A general mixture model approach for mapping quantitative trait loci from diverse cross designs involving multiple inbred lines

YUEFU LIU* AND ZHAO-BANG ZENG
Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA

(*Received 19 March 1999 and in revised form 27 September 1999 and 8 November 1999*)

Summary

Most current statistical methods developed for mapping quantitative trait loci (QTL) based on inbred line designs apply to crosses from two inbred lines. Analysis of QTL in these crosses is restricted by the parental genetic differences between lines. Crosses from multiple inbred lines or multiple families are common in plant and animal breeding programmes, and can be used to increase the efficiency of a QTL mapping study. A general statistical method using mixture model procedures and the EM algorithm is developed for mapping QTL from various cross designs of multiple inbred lines. The general procedure features three cross design matrices, **W**, that define the contribution of parental lines to a particular cross and a genetic design matrix, **D**, that specifies the genetic model used in multiple line crosses. By appropriately specifying **W** matrices, the statistical method can be applied to various cross designs, such as diallel, factorial, cyclic, parallel or arbitrary-pattern cross designs with two or multiple parental lines. Also, with appropriate specification for the **D** matrix, the method can be used to analyse different kinds of cross populations, such as F_2 backcross, four-way cross and mixed crosses (e.g. combining backcross and F_2). Simulation studies were conducted to explore the properties of the method, and confirmed its applicability to diverse experimental designs.

1. Introduction

Most current statistical methods for mapping quantitative trait loci (QTL) based on inbred line designs apply to crosses of two inbred lines (Lander & Botstein, 1989; Haley & Knott, 1992; Zeng, 1994; Jansen & Stam, 1994; Kao & Zeng, 1997). However, many practical breeding programmes utilize a number of crosses that may originate from multiple inbred lines, such as diallel crosses. As molecular marker technology becomes more efficient and is applied to these breeding populations, the need for general statistical methods for QTL mapping analyses that are applicable to these populations becomes apparent. The advantage of this analysis is that it integrates and utilizes QTL mapping information for practical breeding purposes and improves the accuracy and efficiency of practical breeding programmes. From a genetic point of view, QTL analysis of these crosses could be more efficient. In two-line crosses, such as backcrosses and F_2, statistical analyses deal with the segregation of only two QTL alleles at a locus. If the lines share a common QTL allele, the QTL will not be detectable. In multiple-line crosses, multiple QTL alleles may segregate, and the probability of detecting segregating QTL alleles increases. Of course, the statistical analyses with multiple-line crosses are more complex, particularly when different crosses have different and irregular crossing patterns.

There have been several studies on statistical methods of QTL mapping in multiple-line crosses. Rebai *et al.* (1994) extended the regression method of Haley & Knott (1992) to several F_2 populations from a diallel design of multiple inbred lines. They assumed that different lines are fixed with different QTL alleles. Rebai & Goffinet (1993) also examined statistical power for QTL detection in F_2 populations of a diallel cross, assuming that a QTL is exactly on a marker. In a recent paper, Xu (1998) proposed fixed and random model procedures for a specific cross design of multiple inbred lines, where the F_2 populations are independent from each other and t F_1 populations stem from $2t$

* Corresponding author. Tel: +1 (919) 515 2586. Fax: +1 (919) 515 7315. e-mail: yliu@statgen.ncsu.edu

parental inbred lines. Thus, he studied multiple independent two-line crosses for QTL mapping analysis.

The cross patterns in practical breeding populations can be very complex. Different crosses can have different cross patterns and originate from different overlapping or non-overlapping inbred lines. They can be related to each other in a variety of ways. In this paper, we outline a general approach of mixture model analyses that are applicable to various crosses from two or multiple inbred lines. We first focus on a group of F_2 populations from a diallel design of L lines as an example. In simulation studies, we show how the methods can be used for different kinds of combinations of crosses. The methods are sufficiently flexible to be used for a variety of cross populations, such as complete and partial diallel cross designs, factorial cross designs, cyclic cross designs and multiple two-line cross designs, and a mixture of different cross populations.

2. Method

(i) *Data structure*

We first consider a data set of F_2 populations from a diallel cross (without selfing) between inbred lines P_i ($i = 1, 2, \ldots, L$). For L inbred lines, there could be at most $L(L-1)/2$ different F_2 populations, ignoring the differences between reciprocal crosses. The data structures considered in this paper are not restricted to these F_2 populations. To facilitate description of data structure, let $F_1(i,j)$ and $F_2(i,j)$ be the F_1 and F_2 populations originating from P_i and P_j, for $i,j = 1, 2, \ldots, L$ and $j \neq i$; that is $F_1(i,j) = P_i \times P_j$ and $F_2(i,j) = F_1(i,j) \times F_1(i,j)$. Furthermore, denote by $B[(i,j),j]$ the backcross of an $F_1(i,j)$ to parental line P_j and $B[(i,j), k]$ the three-way cross of $F_1(i,j)$ with P_k. Likewise, a four-way cross between $F_1(i,j)$ and $F_1(k,l)$ is denoted by $F_2[(i,j),(k,l)]$. Other crosses may be defined in a similar way.

In the present study, we assume that the inbred line origin of each marker allele segregating in the cross population is known or can be inferred in probability. Further, we also assume that the linkage map of markers that is applicable to all crosses is known. Of course, not all markers are necessarily segregating with different alleles in all crosses, because some inbred lines may have the same marker alleles, or some markers may be uninformative in some crosses. These markers will be recorded as missing or partially missing data in some crosses. In QTL mapping analysis, we can use a Markov chain process (e.g. Jiang & Zeng, 1977) to infer QTL genotype probabilities conditional on observed marker information.

In the data, each record for an individual consists of the type of cross, parental inbred lines, genotypes or phenotype observations of different marker loci, and trait values. If reciprocal crosses are also involved, we need to identify them in the data in order to fit the recoprocal effect in the model.

(ii) *Genetic model*

In dealing with crosses from multiple inbred lines, we need a multiple-allele genetic model to specify the relationship between QTL genotype and trait phenotype. Traditionally, a genetic model is constructed based on the least square principle with genetic effects defined as deviations from the population mean and lower-order terms (Cockerham, 1954; Kempthorne, 1957). In such a model, the additive, dominance and epistatic effects are a function of an array of genotypic values and are frequency dependent.

In this application, however, our population is not a single homogeneous population, but a group of related and heterogeneous cross populations. We are seeking a genetic model that can link the genetic effect parameters of different populations together, and also can be used for different populations more or less autonomously. This would enable us to utilize the relatedness of different cross populations at QTL level for the statistical inference of the genetic architecture of the traits in the whole population, and also make the statistical analysis flexible enough to be applicable to a variety of combinations of cross populations.

For these reasons, we choose to define the genetic effects as deviations from the mean of the inbred lines, rather than the mean of the whole population. At one locus, let

$$\mu = \frac{\sum_{i=1}^{L} g_{ii}}{L},$$

where g_{ii} is the homozygote genotypic value in line i and L is the number of inbred lines. The the additive effect at the locus for line i can be defined as a deviation of g_{ii} from μ:

$$a_i = g_{ii} - \mu.$$

Clearly, the a_i values ($i = 1, 2, \ldots, L$) sum to zero. The dominance effect between Q_i and Q_j is defined as a deviation of heterozygote genotypic value g_{ij} from the averaged additive effect and the mean:

$$d_{ij} = g_{ij} - \frac{a_i + a_j}{2} - \mu.$$

Thus, the model of a multiallelic system for a $F_2(i,j)$ population can be written as

$$\begin{pmatrix} g_{ii} \\ g_{ij} \\ g_{jj} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} a_i \\ a_j \\ d_{ij} \end{pmatrix} + \begin{pmatrix} \mu \\ \mu \\ \mu \end{pmatrix} \quad (1)$$

https://doi.org/10.1017/S0016672300004493 Published online by Cambridge University Press

or in matrix notation

$$\mathbf{G}_{ij} = \mathbf{D}\alpha_{ij} + \mathbf{1}\mu. \quad (2)$$

The **D** matrix is designed to summarize the information of the genetic model and can be different for different kinds of cross populations. For other cross populations, other than F_2, the **D** matrix can be specified correspondingly. With mixed cross populations in a data set, a subscript for **D** is needed to indicate the type of cross involved.

For a population of $F_2(1,2)$ and $F_2(2,3)$, for example, a_2 is estimated from both crosses, and d_{12} and d_{23} are estimated from $F_2(1,2)$ and $F_2(2,3)$ separately. The constraint $a_1 + a_2 + a_3 = 0$ will ensure the model is not over-parameterized.

The model can also be used for a single F_2 (say $F_2(1,2)$) population. With the constraint $a_1 + a_2 = 0$ in this case, (1) is reduced to

$$\begin{pmatrix} g_{11} \\ g_{12} \\ g_{22} \end{pmatrix} = \begin{pmatrix} \mu \\ \mu \\ \mu \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} a \\ d \end{pmatrix}.$$

This model also applies to multiple loci, ignoring epistasis. The issue of epistasis is, however, complicated and ignored here. Epistatic parameters need to be defined with reference to a specific cross. In order to take epistasis into account properly, some reparameterization of the additive and dominance effects is required.

(iii) *Statistical model and likelihood analysis*

Based on the above genetic model, we define the statistical model of composite interval mapping for multiple F_2 data from crosses of multiple inbred lines as

$$y_{ijk} = (z^{[1]}_{ijk}\mathbf{w}^{[1]}_{ijk} + z^{[2]}_{ijk}\mathbf{w}^{[2]}_{ijk})\,\mathbf{a} + z^{[3]}_{ijk}\mathbf{w}^{[3]}_{ijk}\,\mathbf{d} + \mathbf{x}_{ijk}\beta + e_{ijk}, \quad (3)$$

where

$$z^{[1]}_{ijk} = \begin{cases} 1 & \text{if the QTL genotype is } Q_iQ_i \\ \frac{1}{2} & \text{if the QTL genotype is } Q_iQ_j \\ 0 & \text{if the QTL genotype is } Q_jQ_j, \end{cases}$$

$$z^{[2]}_{ijk} = \begin{cases} 0 & \text{if the QTL genotype is } Q_iQ_i \\ \frac{1}{2} & \text{if the QTL genotype is } Q_iQ_j \\ 1 & \text{if the QTL genotype is } Q_jQ_j \end{cases}$$

and

$$z^{[3]}_{ijk} = \begin{cases} 0 & \text{if the QTL genotype is } Q_iQ_i \\ 1 & \text{if the QTL genotype is } Q_iQ_j \\ 0 & \text{if the QTL genotype is } Q_jQ_j. \end{cases}$$

y_{ijk} is the trait value of the kth individual in the F_2 population originating from inbred lines i and j; **a** is a column vector of additive effects, $\mathbf{a} = (a_1, a_2, \ldots, a_L)'$; **d** is a column vector of dominance effects between different alleles, $\mathbf{d} = (d_{12}, d_{13}, \ldots, d_{ij}, \ldots, d_{(L-1)L})'$ for $i < j$; $\mathbf{w}^{[1]}_{ijk}$ is a row vector of length L with the ith element one and others zero, indicating the allelic transmission from parental inbred line i to the cross progeny in F_2; $\mathbf{w}^{[2]}_{ijk}$ is a row vector with the jth element one and others zero, indicating the allelic transmission from parental inbred line j to the cross progeny; $\mathbf{w}^{[3]}_{ijk}$ is a row vector of length $L(L-1)/2$ with the ijth element one and others zero, indicating the coincidence of the allele Q_i from line i with allele Q_j from line j in the cross progeny; β is the column vector of fixed effects that may include the overall mean μ_g, cross means μ_{ij}, some systematic environmental effects, reciprocal effects (if appropriate), and selected marker effects to control the genetic background; $\mathbf{x}_{ijk}$ is the ijkth row vector of the design matrix **X** that relates the fixed effects to observations; e_{ijk} is a residual of the model and assumed to be identically, independently and normally distributed with mean zero and variance σ^2. For the case where $e_{ijk} \sim N(0, \sigma^2_{ij})$ or where repeated measurements are available, likelihood analysis and estimation formulae are presented in the Appendix.

In this model, trait values (**Y**) are modelled as a function of parameters including QTL effects (**a** and **d**) and other fixed effects (β) given the indicator variables (z) for QTL genotypes, line cross information (**W**), and design matrix (**X**) of other fixed effects. Individual QTL genotypes are generally not observed. However, because genetic marker genotypes are observed, the probability distribution of QTL genotypes can be inferred by a Markov chain analysis conditioned on marker genotypes, the genomic position of the putative QTL and the experimental design (Jiang & Zeng, 1997). Let p_{ijkl} ($l = 1, 2, 3$) be the probabilities of the three QTL genotypes for individual k in $F_2(i,j)$ conditioned on marker genotypes and the genomic position of the putative QTL. The likelihood function of the model for $\theta = (\mathbf{a}', \mathbf{d}', \sigma^2, \beta')$ is

$$L(\theta \mid \mathbf{Y}) = \prod_{i=1}^{L-1}\prod_{j=i+1}^{L} (2\pi\sigma^2)^{-\frac{n_{ij}}{2}} \prod_{k=1}^{n_{ij}} \sum_{l=1}^{3} p_{ijkl} \exp\left[-\frac{1}{2\sigma^2}(y_{ijk} - \mu_{ijkl})^2\right] \quad (4)$$

with

$$\mu_{ijk1} = \mathbf{w}^{[1]}_{ijk}\,\mathbf{a} + \mathbf{x}_{ijk}\beta,$$

$$\mu_{ijk2} = (\tfrac{1}{2}\mathbf{w}^{[1]}_{ijk} + \tfrac{1}{2}\mathbf{w}^{[2]}_{ijk})\,\mathbf{a} + \mathbf{w}^{[3]}_{ijk}\,\mathbf{d} + \mathbf{x}_{ijk}\beta,$$

$$\mu_{\mathrm{ijk3}} = \mathbf{w}^{[2]}_{ijk}\,\mathbf{a} + \mathbf{x}_{ijk}\beta.$$

Generally, the analysis for QTL is performed through a search process by analysing the likelihood for different genomic positions. For a given position (and thus p_{ijkl}), the likelihood analysis can be

https://doi.org/10.1017/S0016672300004493 Published online by Cambridge University Press

performed by the EM algorithm (Dempster *et al.*, 1977; Little & Rubin, 1987; Kao & Zeng, 1997). We treat QTL genotypes as missing data. Instead of maximizing the above likelihood function directly, we can maximize the conditional expectation of complete data log-likelihood with respect to QTL genotype given observation $\mathbf{Y}$ and current estimates of parameters $\theta^{[t]}$:

$$Q(\theta\,|\,\theta^{[t]}) = \sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\sum_{k=1}^{n_{ij}}\sum_{l=1}^{3}\ln[\phi(y_{ijk}\,|\,\mu_{ijkl},\sigma^2)p_{ijkl}]\pi_{ijkl},$$

where ϕ is a normal density function and

$$\pi_{ijkl} = \frac{p_{ijkl}\,\phi(y_{ijk}\,|\,\mu_{ijkl},\sigma^2)}{\sum_{l=1}^{3} p_{ijkl}\,\phi(y_{ijk}\,|\,\mu_{ijkl},\sigma^2)} \tag{5}$$

is the posterior probability of the QTL genotype. Calculating π_{ijkl} is the E-step of the EM algorithm. To make the estimation formulae easier to fit different genetic models, we can express μ_{ijkl} in formula (5) as

$$\mu_{ijkl} = \mathbf{D}^l\alpha_{ij} + \mathbf{x}_{ijk}\beta,$$

where $\mathbf{D}^l$ is the lth row vector of the $\mathbf{D}$ matrix in (2), $\alpha_{ij} = (a_i\ a_j\ d_{ij})'$. The M-step is to maximize $Q(\theta\,|\,\theta^{[t]})$ with respect to θ. The estimators can be expressed as

$$\hat{\mathbf{a}} = \mathbf{T}_{aa}^{-1}[\mathbf{S}_a'(\mathbf{Y}-\mathbf{X}\beta) - \mathbf{T}_{ad}\,\mathbf{d}], \tag{6}$$

$$\hat{\mathbf{d}} = \mathbf{T}_{dd}^{-1}[\mathbf{S}_d'(\mathbf{Y}-\mathbf{X}\beta) - \mathbf{T}_{da}\,\mathbf{a}], \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{N}[(\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta) + \alpha'\mathbf{T}\alpha - 2(\mathbf{Y}-\mathbf{X}\beta)'\mathbf{S}\alpha], \tag{8}$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y}-\mathbf{S}\alpha), \tag{9}$$

with

$$\begin{aligned}\mathbf{T}_{aa} &= \mathbf{W}_1'\{\mathbf{W}_1 * [\mathbf{\Pi}(\mathbf{D}_1 \circ \mathbf{D}_1)]\} + \mathbf{W}_1'\{\mathbf{W}_2 * [\mathbf{\Pi}(\mathbf{D}_1 \circ \mathbf{D}_2)]\}\\ &\quad + \mathbf{W}_2'\{\mathbf{W}_1 * [\mathbf{\Pi}(\mathbf{D}_2 \circ \mathbf{D}_1)]\}\\ &\quad + \mathbf{W}_2'\{\mathbf{W}_2 * [\mathbf{\Pi}(\mathbf{D}_2 \circ \mathbf{D}_2)]\},\\ \mathbf{T}_{ad} &= \mathbf{W}_1'\{\mathbf{W}_3 * [\mathbf{\Pi}(\mathbf{D}_1 \circ \mathbf{D}_3)]\} + \mathbf{W}_2'\{\mathbf{W}_3 * [\mathbf{\Pi}(\mathbf{D}_2 \circ \mathbf{D}_3)]\},\\ \mathbf{T}_{da} &= \mathbf{W}_3'\{\mathbf{W}_1 * [\mathbf{\Pi}(\mathbf{D}_3 \circ \mathbf{D}_1)]\} + \mathbf{W}_3'\{\mathbf{W}_2 * [\mathbf{\Pi}(\mathbf{D}_3 \circ \mathbf{D}_2)]\},\\ \mathbf{T}_{dd} &= \mathbf{W}_3'\{\mathbf{W}_3 * [\mathbf{\Pi}(\mathbf{D}_3 \circ \mathbf{D}_3)]\},\\ \mathbf{T} &= \begin{pmatrix}\mathbf{T}_{aa} & \mathbf{T}_{ad}\\ \mathbf{T}_{da} & \mathbf{T}_{dd}\end{pmatrix},\\ \alpha &= (\mathbf{a}'\mathbf{d}')',\\ \mathbf{S} &= (\mathbf{S}_a\,\mathbf{S}_d),\\ \mathbf{S}_a &= (\mathbf{\Pi}\mathbf{D}_1) * \mathbf{W}_1 + (\mathbf{\Pi}\mathbf{D}_2) * \mathbf{W}_2,\\ \mathbf{S}_d &= (\mathbf{\Pi}\mathbf{D}_3) * \mathbf{W}_3,\end{aligned}$$

where the $\mathbf{T}$ and $\mathbf{S}$ matrices are just intermediate variables defined to simplify formulae (6)–(9). In these equations, $\mathbf{\Pi} = \{\pi_{ijkl}\}_{N\times 3}$ with $N = \sum_{i=1}^{L-1}\sum_{j=i+1}^{L} n_{ij}$; $\mathbf{W}_1$ is a $N \times L$ matrix with $\mathbf{w}_{ijk}^{[1]}$ being the ijkth row vector; $\mathbf{W}_2$ has the same size as $\mathbf{W}_1$ with $\mathbf{w}_{ijk}^{[2]}$ being the ijkth row vector; $\mathbf{W}_3$ is a $N \times L(L-1)/2$ matrix in the diallel design with $\mathbf{w}_{ijk}^{[3]}$ being the ijkth row vector. $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}_3$ are the first, second and third column vectors of the $\mathbf{D}$ matrix in (2), corresponding to the indicator variables $z_{ijk}^{[1]}$, $z_{ijk}^{[2]}$ and $z_{ijk}^{[3]}$ respectively. The symbol $\circ$ stands for the Hardamard product, which is an element-by-element product of two same-order matrices. The symbol $*$ is used here to denote the element-by-element product of each column in a matrix by a column vector, i.e. for $\mathbf{A} = \{a_{ij}\}_{n\times m}$ and $\mathbf{b} = \{b_i\}_{n\times 1}$, $\mathbf{A} * \mathbf{b} = \{a_{ij}\,b_i\}_{n\times m}$. The EM is performed in iterations between the E-step (5) and the M-step (6), (7), (8) and (9), starting with an initial guess of parameter values. The converged values of parameter estimates are the maximum likelihood estimates (MLE).

Though the above formulae were derived based on the connected F_2 populations, it is also directly applicable to a single F_2 populations, two backcross populations from which dominance effects are estimable, a single or multiple four-way crosses, etc. When there is only one QTL effect in the model for each cross of a data set, such as backcrosses, the formulae can still be used directly by setting $\mathbf{D}_2$ and $\mathbf{D}_3$ to zero and not estimating other parameters.

The above formulae have a very neat structure and can be used for diverse data structures and experimental designs. Their general nature will be more apparent if we rewrite the formulae for location parameters in terms of normal equations:

$$\begin{pmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{S}_a & \mathbf{X}'\mathbf{S}_d\\ \mathbf{S}_a'\mathbf{X} & \mathbf{T}_{aa} & \mathbf{T}_{ad}\\ \mathbf{S}_d'\mathbf{X} & \mathbf{T}_{da} & \mathbf{T}_{dd}\end{pmatrix}^{[t]}\begin{pmatrix}\beta^{[t+1]}\\ \mathbf{a}^{[t+1]}\\ \mathbf{d}^{[t+1]}\end{pmatrix} = \begin{pmatrix}\mathbf{X}'\mathbf{Y}\\ \mathbf{S}_a'\mathbf{Y}\\ \mathbf{S}_d'\mathbf{Y}\end{pmatrix}^{[t]}.$$

These are the mixture model normal equations for ML estimation and can accommodate different numbers of parameters just as normal equations for least square estimation. Note that the row dimension of matrix $\mathbf{D}$ is equal to the number of mixture components and the column dimension is the number of genetic parameters for each cross, which depends on genetic models and can be different for different kinds of cross populations. Thus, in analysing mapping data with mixed kinds of cross populations, it is necessary to add a code in the data to distinguish different types of crosses in order to let the computer program associate the appropriate $\mathbf{D}$ matrix to each cross.

(iv) *Hypothesis testing*

A test for QTL is performed through a likelihood ratio test under the hypotheses:
H_0: $a_i = 0$ and $d_{ij} = 0$ for $i, j = 1, 2, \ldots, L$ and $i \neq j$, i.e. no QTL at the testing position;

https://doi.org/10.1017/S0016672300004493 Published online by Cambridge University Press

H_1: At least one of the genetic effects a_i or d_{ij} is not zero, implying that the testing position may contain a QTL.

For crosses of multiple inbred lines, the likelihood ratio test statistic takes the form

$$LR = -2\log\frac{L\left(\bigcap_{i=1}^{L} a_i = 0 \text{ and } \bigcap_{i<j}^{L} d_{ij} = 0, \hat{\beta}_0, \hat{\sigma}_0^2\right)}{L(\hat{a}_i, \hat{d}_{ij}, \hat{\beta}, \hat{\sigma}^2)}, \quad (10)$$

where $\hat{a}_i$, $\hat{d}_{ij}$, $\hat{\beta}$, $\hat{\sigma}^2$ are MLE of parameters a_i, d_{ij}, β, σ^2 under H_1, and $\hat{\beta}_0$ and $\hat{\sigma}_0^2$ are MLE under H_0.

Determining the threshold of the test statistic is a complicated issue. Many factors, such as the sample size, genome size, genetic map density and proportion of missing data, could affect the distribution of the test statistic under the null hypothesis (Lander & Botstein, 1989; Zeng, 1994; Churchill & Doerge, 1994; Kao, 1995). With appropriate adjustment to the critical level based on an estimate of the effective number of independent tests involved (Lander & Botstein, 1989; Zeng, 1994), an approximate threshold can be calculated. However, when multiple parameters are tested simultaneously, the issue can be further complicated. Therefore, we recommend using the union intersection method (Casella & Berger, 1990) and split the above hypothesis into many subsets of hypotheses. Taking the diallel design of three lines as an example, the hypothesis can be split into the following six subsets:

$H_0 1: a_1 = 0, H_1 1: a_1 \neq 0,$

$H_0 2: a_2 = 0, H_1 2: a_2 \neq 0,$

$H_0 3: a_3 = 0, H_1 3: a_3 \neq 0,$

$H_0 4: d_{12} = 0, H_1 4: d_{12} \neq 0,$

$H_0 5: d_{13} = 0, H_1 5: d_{13} \neq 0,$

$H_0 6: d_{23} = 0, H_1 6: d_{23} \neq 0.$

Each subset of hypotheses tests one parameter. If all the subsets of the null hypothesis are not rejected based on the separate likelihood ratio tests, the null hypothesis will not be rejected. The rejection of any subset of the null hypothesis will lead to rejection of the null hypothesis. The subsets of the null hypothesis can be tested for every position along the genome or only at the positions where the total LRT statistic reaches its local maximum to save computing time. The position of a QTL is estimated at a position where the null hypothesis is rejected and the maximum likelihood is achieved. Correspondingly, the effects of the putative QTL are estimated by MLE at the estimated QTL position.

The maximum likelihood estimates of parameters β and σ^2 under the null hypothesis are the estimates based on the reduced model

$$y_{ijk} = \mathbf{x}_{ijk}\beta + e_{ijk}$$

with all a_i's and d_{ij}'s constrained to zero. Since it is a multiple regression model, the MLE of β and σ^2 are simply

$$\hat{\beta}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \quad (11)$$

$$\hat{\sigma}_0^2 = \frac{1}{N}(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta). \quad (12)$$

In testing a subset of the hypoothesis, the maximum likelihood under a specific null hypothesis is obtained with appropriate parameters in vector **a** and/or in vector **d** constrained to zero.

3. Simulation study

(i) *Simulation methods*

For simplicity, we simulate only one QTL in a linkage group of 100 cM with five different combinations of cross populations to demonstrate the general applicability of the procedure. The markers are evenly distributed in the linkage group with interval size (*int*) 5 or 10 cM. The position of the QTL is set to 34 cM. We simulate several crosses from three inbred lines, each fixed with different alleles in the first four cases. The genotype values of the three homozygotes, Q_1Q_1, Q_2Q_2 and Q_3Q_3, and the three heterozygotes, Q_1Q_2, Q_1Q_3 and Q_2Q_3, in the populations are assumed to be $-0{\cdot}3$, $-0{\cdot}4$, $0{\cdot}6$, $-0{\cdot}45$, $0{\cdot}25$ and $-0{\cdot}1$ respectively. In the fifth case, we assume $Q_1 \equiv Q_2$, i.e. Q_1 and Q_2 are the same allele. No systematic environmental effects are simulated for simplicity. The mean values of crossbred populations $F_2(1,2)$, $F_2(1,3)$, $F_2(2,3)$, $B[(1,2),1]$, $B[(1,2),2]$ and $B[(2,3),3]$ are assumed to be $1{\cdot}50$, $2{\cdot}00$, $2{\cdot}50$, $1{\cdot}25$, $1{\cdot}75$ and $2{\cdot}75$, respectively. The residual variance is scaled to give the variance explained by the QTL, $R^2 = 0{\cdot}1$, $0{\cdot}3$ or $0{\cdot}6$. One hundred replicates are simulated for each parameter combination, and the search analysis for the QTL is performed at 1 cM intervals for each replicate. The five cases of population structures and genetic models are:

Case 1: Three F_2 populations, $F_2(1,2)$, $F_2(1,3)$ and $F_2(2,3)$, from a diallel cross design of the three inbred lines. The sample size of each cross is assumed to be 50, 100 or 1000.

Case 2: Three F_2 populations, $F_2(1,2)$, $F_2(1,3)$ and $F_2(2,3)$, plus three backcross populations, $B[(1,2),1]$, $B[(1,2),2]$ and $B[(2,3),3]$, from the three inbred lines. The sample size for each F_2 population is 50, and for each backcross population 50, 100 or 500. This case is used to show the advantage of combining different types of crosses in the analysis.

Case 3: A single F_2 population, $F_2(1,2)$, with sample size 150, 300 or 1500. This case is used for comparison with other cases.

Case 4: Two backcross populations, $B[(1,2),1]$ and $B([1,2),2]$, each with sample size 75, 150 or 750. This

https://doi.org/10.1017/S0016672300004493 Published online by Cambridge University Press

Table 1. *The expected additive and dominance effects as well as means ($\times 10^{-3}$) in the five cases of simulation*

	μ_g	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	a_1	a_2	a_3	d_{12}	d_{13}	d_{23}
Case 1	1967	−500	0	500	—	—	—	−267	−367	633	−100	100	−200
Case 2	1925	−458	42	542	−708	−208	792	−267	−367	633	−100	100	−200
Case 3	1150	—	—	—	—	—	—	50	−50	—	−100	—	—
Case 4	1150	−250	250	—	—	—	—	50	−50	—	−100	—	—
Case 5	1933	−500	0	500	—	—	—	−333	−333	667	0	−200	−200

μ_1, μ_2, μ_3, μ_4, μ_5 and μ_6 stand for cross means of $F_2(1,2)$, $F_2(1,3)$, $F_2(2,3)$, $B[(1,2),1]$, $B[(1,2),2]$ and $B[(2,3),3]$, respectively. μ_g stands for the overall mean of the whole population.

Table 2. *Means and standard deviations of QTL position estimates (cM) over 100 replicates*

N	R^2	Case 1	Case 2	Case 3	Case 4	Case 5
int = 5						
150	0·1	35·7 (10·1)[a]	35·4 (7·31)	33·9 (6·41)	35·5 (9·31)	33·5 (9·77)[d]
	0·3	34·5 (4·11)	33·8 (2·43)	34·5 (1·92)	33·9 (4·00)	34·7 (4·78)
300	0·1	33·8 (6·61)	34·8 (4·61)	34·1 (2·51)	33·0 (5·90)	33·8 (5·95)
	0·3	34·3 (2·35)	33·8 (2·01)	34·1 (1·35)	34·2 (1·46)	34·3 (1·99)
1500	0·1	34·5 (1·46)	34·4 (1·73)	34·1 (1·45)	34·2 (1·41)	34·4 (1·53)
	0·3	34·2 (0·84)	34·0 (0·90)	34·0 (0·50)	34·0 (0·83)	34·1 (0·73)
int = 10						
150	0·1	33·7 (15·7)[b]	35·4 (12·0)[e]	36·4 (9·76)	35·1 (13·3)[e]	35·2 (15·9)[e]
	0·3	33·9 (7·17)	33·4 (4·75)	35·1 (2·31)	34·2 (5·56)	34·4 (7·69)
300	0·1	34·8 (11·4)	34·4 (9·44)	35·4 (2·98)	33·2 (9·75)	34·5 (11·0)
	0·3	33·3 (3·01)	33·6 (2·97)	35·1 (1·44)	33·5 (2·21)	33·3 (2·77)
1500	0·1	33·4 (1·93)	33·7 (2·81)	34·8 (1·26)	33·8 (2·01)	33·7 (2·06)
	0·3	33·8 (0·91)	33·9 (1·05)	34·1 (0·72)	34·2 (0·96)	33·7 (0·92)

[a,b,c,d] and [e] denote means based on the 96, 94, 98, 97 and 92 significant replicates respectively. N is the total sample size. The sample size is $N+150$ in Case 2. *int* stands for the size of marker intervals. R^2 is the proportion of the variance explained by the QTL.

case is used to show that with two backcrosses, both additive and dominance effects of the QTL can be estimated.

Case 5: The same as case 1 except that line 1 is fixed for the same allele as that of line 2, i.e. $Q_1 \equiv Q_2$. Thus the genotypic values of both Q_1Q_1 and Q_2Q_2 are −0·4. The genotypic values of heterozygotes, Q_1Q_3 and Q_2Q_3, are all −0·1 and Q_1Q_2 is the same as Q_1Q_1 and Q_2Q_2.

The expected additive and dominance effects are calculated from the genotypic values, based on the genetic model, and listed in Table 1. The expected values of cross means in five simulated cases are also listed in Table 1 and calculated under the constraint that the sum of cross mean parameters in each simulated case is zero.

(ii) *Results*

The estimates of QTL position and sampling variances of estimates are shown in Table 2 for the five cases. The estimates of QTL effects and sampling variances of estimates are shown in Table 3 for Case 2 and Table 4 for Case 5. For all the cases, the simulation results clearly show that consistent estimates for QTL position and effects are obtained. As expected, the sampling variances of estimates of QTL position and effects decrease as the sample size increases.

Besides the sample size, the marker interval size and the proportion of genetic variance explained by the QTL are two other important factors affecting estimates of QTL position and effects, as observed for the case analysis of two inbred line (Kao & Zeng, 1997). As the marker interval size decreases, the sampling variance of estimates of QTL position and effects decreases. The proportion of genetic variance explained by the QTL significantly affects estimates of QTL position, as expected. The sampling variances of estimates of QTL position and effects for $R^2 = 0{\cdot}6$ (not shown in the tables) are generally about half those for $R^2 = 0{\cdot}3$. The difference in the sampling variances between $R^2 = 0{\cdot}3$ and $R^2 = 0{\cdot}1$ are, however, much larger. It is interesting to note that the sampling variances of QTL position estimates with the highest R^2 (= 0·6) but the lowest sample size ($N = 150$) and lower marker density (*int* = 10) (not shown in Table 2) are very similar to those with the lowest R^2 (= 0·1) but the highest sample size ($N = 1500$) and higher

https://doi.org/10.1017/S0016672300004493 Published online by Cambridge University Press

Table 3. *Means and standard deviations* ($\times 10^{-3}$) *of QTL effect estimates from 100 replicates of the mixed data simulated in case 2*

N	R^2	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{d}_{12}$	$\hat{d}_{13}$	$\hat{d}_{23}$	$\hat{\sigma}^2$
int = 5								
150	0·1	−270 (144)	−369 (137)	640 (129)	−101 (190)	144 (272)	−213 (202)	1046 (103)
	0·3	−265 (69)	−367 (69)	634 (64)	−102 (93)	124 (131)	−199 (99)	272 (28)
300	0·1	−288 (94)	−345 (97)	635 (91)	−95 (127)	159 (265)	−228 (144)	1033 (98)
	0·3	−276 (47)	−354 (43)	632 (45)	−100 (56)	127 (137)	−206 (74)	268 (27)
1500	0·1	−267 (55)	−366 (54)	635 (46)	−100 (59)	87 (200)	−205 (64)	1006 (48)
	0·3	−266 (28)	−366 (27)	633 (24)	−100 (30)	91 (101)	−201 (32)	261 (12)
int = 10								
150	0·1	−274 (161)[c]	−354 (142)[c]	630 (135)[c]	−81 (195)[c]	59 (318)[c]	−204 (234)[c]	1034 (113)[c]
	0·3	−271 (75)	−358 (72)	631 (57)	−93 (101)	76 (141)	−204 (106)	270 (30)
300	0·1	−268 (97)	−364 (107)	634 (106)	−99 (126)	113 (247)	−207 (161)	1019 (94)
	0·3	−265 (49)	−363 (54)	630 (55)	−96 (72)	106 (123)	−199 (81)	265 (29)
1500	0·1	−253 (51)	−374 (53)	629 (46)	−98 (64)	98 (228)	−182 (80)	1011 (53)
	0·3	−259 (27)	−370 (26)	631 (23)	−98 (32)	99 (114)	−190 (39)	262 (15)

[c] Based on the 98 significant replicates.

Table 4. *Means and standard deviations* ($\times 10^{-3}$) *of QTL effect estimates from 100 replicates of three* F_2 *populations with* $Q1 \equiv Q2$ *simulated in Case 5*

N	R^2	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{d}_{12}$	$\hat{d}_{13}$	$\hat{d}_{23}$	$\hat{\sigma}^2$
int = 5								
150	0·1	−386 (210)[d]	−325 (205)[d]	714 (198)[d]	9 (333)[d]	−162 (295)[d]	−247 (332)[d]	1065 (193)[d]
	0·3	−356 (100)	−327 (98)	685 (97)	4 (146)	−180 (139)	−222 (134)	279 (58)
300	0·1	−350 (129)	−353 (127)	705 (129)	34 (199)	−235 (216)	−199 (200)	1098 (136)
	0·3	−339 (62)	−343 (61)	683 (64)	18 (101)	−220 (106)	−196 (93)	285 (38)
1500	0·1	−340 (55)	−327 (55)	669 (54)	−24 (92)	−214 (91)	−224 (85)	1126 (53)
	0·3	−335 (27)	−329 (27)	666 (27)	−12 (45)	−205 (40)	−210 (43)	292 (17)
int = 10								
150	0·1	−328 (234)[e]	−317 (223)[e]	647 (237)[e]	−29 (375)[e]	−176 (361)[e]	−202 (404)[e]	1032 (197)[e]
	0·3	−333 (102)	−328 (101)	663 (102)	−7 (164)	−190 (149)	−208 (174)	273 (56)
300	0·1	−315 (159)	−344 (143)	661 (156)	16 (231)	−183 (216)	−160 (221)	1107 (139)
	0·3	−325 (74)	−340 (67)	666 (67)	15 (102)	−190 (101)	−176 (99)	289 (46)
1500	0·1	−328 (57)	−328 (57)	658 (61)	0 (90)	−178 (82)	−195 (97)	1116 (59)
	0·3	−329 (29)	−329 (29)	660 (31)	0 (45)	−187 (42)	−195 (49)	290 (18)

[d] Based on the 97 significant replicates.
[e] Based on the 92 significant replicates.

marker density *int* = 5. By increasing sample size and marker density, the ability to detect QTL with smaller effects can be significantly increased. In Cases 3 and 4, the QTL additive effects were estimated both as allelic effects (as defined in the model) and as the difference of the two allelic effects. Both parameterizations lead to the same estimated QTL position, as expected.

In comparing results of different cases, several interesting conclusions can be made. First, compared with Case 1, Case 3 has smaller sampling variances of estimates of QTL positions and effects for the same sample size. This reflects the fact that the number of QTL effects to be estimated is reduced from six in Case 1 to two in Case 3. Secondly, compared with Case 4, Case 3 has smaller sampling variances of estimates of QTL position and effects (results not shown), indicating that an F_2 population is better than two backcross populations with the same total sample size. Third, as expected, the sampling variances of additive effects are generally smaller than those of dominance effects in all simulated cases.

4. Discussion

Many current statistical methods for QTL mapping are developed for crosses from two inbred lines. However, in many practical breeding programmes there could be multiple crosses originating from multiple inbred lines and these crosses can be used for QTL mapping (Rebai *et al*., 1994). One advantage of using multiple crosses is to increase the chance of identifying different QTL alleles that are segregating

https://doi.org/10.1017/S0016672300004493 Published online by Cambridge University Press

in different crosses (Xu, 1996, 1998). Another advantage is to increase the statistical power to identify more QTL if data from different crosses are analysed together (Rebai & Goffinet, 1993).

In order to combine different crosses in a single analysis, it is necessary to take the genetic structure of different crosses into account in the statistical model. The genetic structure of a cross population in a multiple-cross data set is the same as that of the corresponding single cross. The genetic relationship between different crosses, however, needs to be connected through a design matrix indicating allelic transition. In this paper we outline a statistical method for a joint QTL mapping analysis in multiple crosses from multiple inbred lines. The method is an extension of composite interval mapping from crosses of two inbred lines (Zeng, 1994; Jiang & Zeng, 1997) to those of multiple inbred lines. The generalization of the method is realized through the introduction of an experimental design matrix **W** and a genetic model matrix **D**. The experimental design matrix **W** contains information on allelic transition from parental inbred lines to cross progeny. The genetic model matrix **D** specifies the relationship between genotypes and genetic parameters in a two- or multiple-allelic system.

The introduction of the **W** matrix makes the method applicable to a variety of cross designs from two or multiple inbred lines, such as complete or partial diallel, factorial and cyclic cross designs, or designs with more complex structures. The use of the **D** matrix also makes the method applicable to different cross populations, such as backcross, double haploid, recombinant inbred lines, F_2, F_3, three-way cross, four-way cross. With appropriate specification of the **W** and **D** matrices together, the method can be applied to a variety of cross designs, data structures and genetic models. In analysing the mixed cross data we need to put a code for each individual in the data to identify the type of cross and apply the appropriate **D** matrix configuration accordingly.

In the simulation, we have demonstrated the application of the method in a partial diallel cross design with three inbred lines in case 1 and a mixed group of crosses (F_2 and backcrosses) in Cases 2 and 4. To test the general applicability of the methods for large data set we have applied the methods and procedures to a simulated data set of many irregular crosses from 10 inbred lines and also to another simulated data set of partial diallel crosses of 30 inbred lines. Consistent results (not shown here) are obtained in both cases. However, as the numbers of inbred lines and crosses increase, it would be more appropriate to regard the genetic effects as random effects for hypotheses testing and parameter estimation. This would not only make the statistical inference applicable to other related populations but also make the statistical analysis more efficient computationally. Sample size is not a major limiting factor to computation, but the number of parameters is. With hundreds of cross populations from many inbred lines in some large-scale breeding programmes it is appropriate to use a random genetic effect model for QTL mapping analysis. We have extended the current study to include random genetic effects, and the results will be published elsewhere.

Of course, the method can be used to analyze a single F_2 or backcross population, as shown in Case 3. In the Appendix we show how to specify the **W** and **D** matrices to reduce the formula for a cross of two inbred lines and to produce results corresponding to those of Kao & Zeng (1997).

In specifying the **D** matrix, some reparameterization may be required for some experimental designs. This is the case, for example, when the number of crosses is smaller than the number of inbred lines, such as a backcross or a four-way cross. For four-way cross data, the number of parameters can be quite large compared with the number of cross populations. A solution to reparameterization for this is to define differences of genetic effects as parameters. For instance, the genetic model can be specified with three genetic parameters as in two-way ANOVA design (Seber, 1977): one for the difference between allelic effects a_1 and a_2, one between a_3 and a_4, and the third for the interaction between the two differences. The **D** matrix under this definition will be

$$\mathbf{D} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 1 \\ \frac{1}{2} & -\frac{1}{2} & -1 \\ -\frac{1}{2} & \frac{1}{2} & -1 \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix}.$$

The three columns specify the configuration for the three corresponding genetic parameters. For more complex cross designs, sometimes reparameterization or constraints need to be made on certain subsets of parameters to ensure the estimability of parameters in the model. If each cross involves different inbred lines, mapping analysis on different crosses can proceed separately, but some joint analyses may still be needed to test certain hypotheses concerning QTL position and some QTL effects, such as QTL by environment or genotype interaction, a situation similar to design II or Jiang & Zeng (1995).

When mapping data are composed of crosses from multiple inbred lines, it might be tempting to extend a two-line cross mapping analysis directly to the cross populations from multiple lines by fitting the data into a model like

$$y_{ijk} = a_{ij} x^*_{ijk} + d_{ij} z^*_{ijk} + x_{ijk}\beta + e_{ijk}.$$

The likelihood of the whole data set under this model is equal to the summation of likelihoods of all individual crosses in the data set. However, it is easy to show that this direct extension is appropriate only

https://doi.org/10.1017/S0016672300004493 Published online by Cambridge University Press

for crosses involving different and non-overlapping lines, such as the case studied by Xu (1998), and is not appropriate for crosses involving some common lines. This is because different crosses that descended from some common lines share some common alleles and are genetically correlated. This genetic structure is not utilized in the above model. Also, the number of the genetic parameters in the above model can be substantially higher than necessary, reducing the power and resolution of the analysis. For example, for a diallel cross design of L parental lines (without reciprocal crosses) there are $L(L-1)/2$ crosses, and the above model would fit $L(L-1)/2$ additive parameters, whereas there are only $L-1$ additive parameters necessary in our model. Thus whenever there are common alleles in different crosses, the information on common alleles should be utilized in the construction of the genetic model.

In practice, one of the purposes of using multiple crosses for mapping QTL is, of course, to use the mapping results to improve the efficiency of a breeding programme. QTL mapping analysis provides estimates of genetic parameters, and these estimates can be used to predict individual genotypic values in a practical breeding programme. There are two points that need to be emphasized in relating QTL mapping to marker-assisted selection. First, a genetic model defined for QTL mapping analysis is population (inbred lines) dependent. The allelic effects in a two-allele model are defined as a difference between two alleles. In a multiple-allele model, allelic effects are defined with reference to a different base. Thus, estimates of genetic parameters of a two-allele model from a cross between two inbred lines can be applied to crosses of the same inbred lines, not crosses from other inbred lines. If a breeding programme has crosses from multiple inbred lines and selection needs to be practised in these and other related populations from the same genetic materials, a multiple-allele model that depicts the genetic structure of multiple lines and their crosses has to be used for mapping QTL in order to obtain mapping results applicable to those populations. This point emphasizes the importance of the current study in the application of genetic marker technology in practical breeding programmes. Secondly, when mapping results are applied to a population that is selected, individual genotypes at QTL loci are not necessarily observed, but their distribution can be inferred from genetic marker information given the estimated genomic positions of QTL. Thus, in predicting individual genotypic values, this distribution of different genotypes can be used to weight different genotypic values for the prediction, a situation similar to the mixture analysis in QTL mapping. Essentially, QTL mapping and marker-assisted selection are better evaluated in the same framework.

Appendix

A. *The formulae for heteroscedastic models*

When different crosses have different residual variances, i.e. $e_{ijk} \sim N(0, \sigma^2_{ij})$, the likelihood function of the parameters is

$$L(\theta \mid \mathbf{Y}) = \prod_{i=1}^{L-1} \prod_{j=i+1}^{L} (2\pi\sigma^2_{ij})^{-\frac{n_{ij}}{2}} \prod_{k=1}^{n_{ij}} \sum_{l=1}^{3} p_{ijkl} \times \exp\left[-\frac{1}{2\sigma^2_{ij}}(y_{ijk}-\mu_{ijkl})^2\right]$$

and the conditional expectation of the complete data log-likelihood with respect to QTL genotype given observation $\mathbf{Y}$ and current guess of parameters $\theta^{[t]}$ is

$$Q(\theta \mid \theta^{[t]}) = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \sum_{k=1}^{n_{ij}} \sum_{l=1}^{3} \ln[\phi(y_{ijk} \mid \mu_{ijkl}, \sigma^2_{ij}) p_{ijkl}] \pi_{ijkl},$$

where

$$\pi_{ijkl} = \frac{p_{ijkl}\phi(y_{ijk} \mid \mu_{ijkl}, \sigma^2_{ij})}{\sum_{l=1}^{3} p_{ijkl}\phi(y_{ijk} \mid \mu_{ijkl}, \sigma^2_{ij})}.$$

The formulae for parameter estimation are listed as follows:

$$\hat{\mathbf{a}} = \mathbf{T}^{-1}_{R,aa}[\mathbf{S}'_a \mathbf{R}^{-1}(\mathbf{Y}-\mathbf{X}\beta) - \mathbf{T}_{R,ad}\mathbf{d}],$$
$$\hat{\mathbf{d}} = \mathbf{T}^{-1}_{R,dd}[\mathbf{S}'_d \mathbf{R}^{-1}(\mathbf{Y}-\mathbf{X}\beta) - \mathbf{T}_{R,da}\mathbf{a}], \quad (13)$$
$$\hat{\beta} = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}(\mathbf{Y}-\mathbf{S}\alpha), \quad (14)$$

where

$$\begin{aligned}
\mathbf{T}_{R,aa} &= \mathbf{W}'_1\mathbf{R}^{-1}\{\mathbf{W}_1 * [\mathbf{\Pi}(\mathbf{D}_1 \circ \mathbf{D}_1)]\} \\
&\quad + \mathbf{W}'_1\mathbf{R}^{-1}\{\mathbf{W}_2 * [\mathbf{\Pi}(\mathbf{D}_1 \circ \mathbf{D}_2]\} \\
&\quad + \mathbf{W}'_2\mathbf{R}^{-1}\{\mathbf{W}_1 * [\mathbf{\Pi}(\mathbf{D}_2 \circ \mathbf{D}_1)]\} \\
&\quad + \mathbf{W}'_2\mathbf{R}^{-1}\{\mathbf{W}_2 * [\mathbf{\Pi}(\mathbf{D}_2 \circ \mathbf{D}_2)]\}, \\
\mathbf{T}_{R,ad} &= \mathbf{W}'_1\mathbf{R}^{-1}\{\mathbf{W}_3 * [\mathbf{\Pi}(\mathbf{D}_1 \circ \mathbf{D}_3)]\} \\
&\quad + \mathbf{W}'_2\mathbf{R}^{-1}\{\mathbf{W}_3 * [\mathbf{\Pi}(\mathbf{D}_2 \circ \mathbf{D}_3)]\}, \\
\mathbf{T}_{R,da} &= \mathbf{W}'_3\mathbf{R}^{-1}\{\mathbf{W}_1 * [\mathbf{\Pi}(\mathbf{D}_3 \circ \mathbf{D}_1)]\} \\
&\quad + \mathbf{W}'_3\mathbf{R}^{-1}\{\mathbf{W}_2 * [\mathbf{\Pi}(\mathbf{D}_3 \circ \mathbf{D}_2)]\}, \\
\mathbf{T}_{R,dd} &= \mathbf{W}'_3\mathbf{R}^{-1}\{\mathbf{W}_3 * [\mathbf{\Pi}(\mathbf{D}_3 \circ \mathbf{D}_3)]\}.
\end{aligned}$$

Here, $\mathbf{R}^{-1}$ is the inverse of the variance matrix of residual vector $\mathbf{e} = \{e_{ijk}\}$. $\mathbf{T}_{R,aa}$, $\mathbf{T}_{R,ad}$, $\mathbf{T}_{R,da}$ and $\mathbf{T}_{R,dd}$ are different from $\mathbf{T}_{aa}$, $\mathbf{T}_{ad}$, $\mathbf{T}_{da}$ and $\mathbf{T}_{dd}$ by including matrix $\mathbf{R}$. $\mathbf{S}$, $\mathbf{S}$, $\mathbf{S}_a$, $\mathbf{S}_d$ and α are defined in the main text.

$$\hat{\sigma}^2_{ij} = \frac{1}{n_{ij}}[(\mathbf{Y}_{ij}-\mathbf{X}_{ij}\beta)'(\mathbf{Y}_{ij}-\mathbf{X}_{ij}\beta) + \alpha'\mathbf{T}_{ij}\alpha - 2(\mathbf{Y}_{ij}-\mathbf{X}_{ij}\beta)'\mathbf{S}_{ij}\alpha]. \quad (15)$$

Here, $\mathbf{Y}_{ij}$, $\mathbf{X}_{ij}$, $\mathbf{S}_{ij}$ are the subset of $\mathbf{Y}, \mathbf{X}, \mathbf{S}$, respectively, that corresponds to cross *ij*. The matrix $\mathbf{T}_{ij}$ and its elements $\mathbf{T}_{aa,ij}$, $\mathbf{T}_{ad,ij}$, $\mathbf{T}_{da,ij}$ and $\mathbf{T}_{dd,ij}$ are calculated from the subsets of $\mathbf{W}_i$ $(i = 1, 2, 3)$ and $\mathbf{\Pi}$ corresponding to cross *ij*, using the formulae in the main text.

https://doi.org/10.1017/S0016672300004493 Published online by Cambridge University Press

The formulae are applicable to the data of multiple measurements on each experimental unit. In that case the elements in matrix **R** are variances of average residual error over the repeated measurements,

$$\sigma_{\bar{e}}^2 = \frac{1+(k-1)\,t}{k}\sigma_e^2,$$

where t is the repeatability, and k is the number of repeated measurements, which can be different for different experimental units.

B. *The reduced form of the general formulae for single-cross data*

In a two-line cross there are only two elements in each row of matrices $\mathbf{W}_1$ and $\mathbf{W}_2$. All elements in the first column of $\mathbf{W}_1$ and the second column of $\mathbf{W}_2$ are one, and the others are zero. $\mathbf{W}_3$ will become a column vector with all elements being one. Consequently, formulae (6)–(9) can be simplified as

$$\begin{pmatrix} \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_1\circ\mathbf{D}_1) & \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_1\circ\mathbf{D}_2) \\ \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_2\circ\mathbf{D}_1) & \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_2\circ\mathbf{D}_2) \end{pmatrix}\begin{pmatrix}\hat{a}_1\\ \hat{a}_2\end{pmatrix} = \begin{pmatrix}(\mathbf{Y}-\mathbf{X}\beta)'\,\mathbf{\Pi}\mathbf{D}_1\\ (\mathbf{Y}-\mathbf{X}\beta)'\,\mathbf{\Pi}\mathbf{D}_2\end{pmatrix} - \begin{pmatrix}\mathbf{1}'\mathbf{\Pi}(\mathbf{D}_1\circ\mathbf{D}_3)\\ \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_2\circ\mathbf{D}_3)\end{pmatrix} d,$$

$$\hat{d} = \frac{(\mathbf{Y}-\mathbf{X}\beta)'\,\mathbf{\Pi}\mathbf{D}_3 - (\mathbf{1}'\mathbf{\Pi}(\mathbf{D}_3\circ\mathbf{D}_1)\,a_1 + \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_3\circ\mathbf{D}_2)\,a_2)}{\mathbf{1}'\mathbf{\Pi}(\mathbf{D}_3\circ\mathbf{D}_3)},$$

$$\hat{\sigma}^2 = \frac{1}{n}[(\mathbf{Y}-\mathbf{X}\beta)'\,(\mathbf{Y}-\mathbf{X}\beta) - \alpha'\mathbf{T}\alpha - 2(\mathbf{Y}-\mathbf{X}\beta)'\,\mathbf{\Pi}\mathbf{D}\alpha],$$

with

$$\mathbf{T} = \begin{pmatrix} \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_1\circ\mathbf{D}_1) & \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_1\circ\mathbf{D}_2) & \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_1\circ\mathbf{D}_3) \\ \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_2\circ\mathbf{D}_1) & \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_2\circ\mathbf{D}_2) & \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_2\circ\mathbf{D}_3) \\ \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_3\circ\mathbf{D}_1) & \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_3\circ\mathbf{D}_2) & \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_3\circ\mathbf{D}_3) \end{pmatrix},$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y}-\mathbf{\Pi}\mathbf{D}\alpha).$$

In a two-line cross design, we usually constrain $a_1+a_2=0$. Let $a_1=a$, $a_2=-a$. Then the genetic design matrix **D** for an F_2 becomes

$$\mathbf{D}^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{pmatrix}$$

and the formulae can be simplified further as

$$\hat{a} = \frac{(\mathbf{Y}-\mathbf{X}\beta)'\,\mathbf{\Pi}\mathbf{D}_1^* - \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_1^*\circ\mathbf{D}_2^*)\,d}{\mathbf{1}'\mathbf{\Pi}(\mathbf{D}_1^*\circ\mathbf{D}_1^*)},$$

$$\hat{d} = \frac{(\mathbf{Y}-\mathbf{X}\beta)'\,\mathbf{\Pi}\mathbf{D}_2^* - \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_1^*\circ\mathbf{D}_2^*)\,a}{\mathbf{1}'\mathbf{\Pi}(\mathbf{D}_2^*\circ\mathbf{D}_2^*)},$$

$$\hat{\sigma}^2 = \frac{1}{n}[(\mathbf{Y}-\mathbf{X}\beta)'\,(\mathbf{Y}-\mathbf{X}\beta) + \gamma'\mathbf{T}\gamma - 2(\mathbf{Y}-\mathbf{X}\beta)'\mathbf{\Pi}\mathbf{D}^*\gamma].$$

where

$$\mathbf{T} = \begin{pmatrix} \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_1^*\circ\mathbf{D}_1^*) & \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_1^*\circ\mathbf{D}_2^*) \\ \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_2\circ\mathbf{D}_1^*) & \mathbf{1}'\mathbf{\Pi}(\mathbf{D}_2^*\circ\mathbf{D}_2^*) \end{pmatrix},$$

$$\gamma = (a\;d)',$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'(\mathbf{Y}-\mathbf{\Pi}\mathbf{D}^*\gamma).$$

These formulae are reduced to those for a two-line cross presented by Kao & Zeng (1997). By further specifying the coefficients in the genetic design matrix **D**, we can obtain the specific formula of the composite interval mapping for a backcross population (Zeng, 1994) or for an F_2 population (Kao, 1995). When further taking out the fixed effects β, we can get the formulae for interval mapping (Lander & Botstein, 1989).

We sincerely thank Bill Hill and two anonymous reviewers for constructive comments. This study was supported in part by grants GM 45344 from National Institutes of Health and no. 9600645 from USDA Plant Genome Program.

References

Casella, G. & Berger, R. L. (1990). *Statistical Inference*. Pacific Grove, California: Brooks/Cole.

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

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* **39**, 859–882.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39**, 1–38.

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

Jansen, R. C. & Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**, 1447–1455.

Jiang, C. & Zeng, Z.-B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**, 1111–1127.

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

Kao, C.-H. (1995). Statistical methods for locating positions and analyzing epistasis of multiple quantitative trait genes using molecular marker information. Dissertation of Department of Statistics at North Carolina State University.

Kao, C.-H. & Zeng, Z.-B. (1997). General formulae for obtaining the MLEs and the asymptotic variance–covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* **53**, 653–665.

Kempthorne, O. (1957). *An Introduction to Genetic Statistics*. Ames, Iowa: The Iowa State University Press.

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

Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

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

Rebai, A., Goffinet, B., Mangin, B. & Perret, D. (1994). Detecting QTLs with diallel schemes. In *Biometrics in Plant Breeding: Applications of Molecular Markers*, *9th meeting of the EUCARPIA*, (ed. J. W. van Ooijen & J. Jansen), pp. 170–177. Wageningen, the Netherlands: CPRO-DLO.

Seber, G. A. F. (1977). *Linear Regression Analysis*. New York: Wiley.

Xu, S. (1996). Mapping quantitative loci using four-way crosses. *Genetical Research* **68**, 175–181.

Xu, S. (1998). Mapping quantitative loci using families of line crosses. *Genetics* **148**, 517–524.

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