Heredity (2003) 90, 228–235

www.nature.com/hdy

Mapping quantitative trait loci underlying triploid endosperm traits

C Xu[1,2], X He[2] and S Xu[1]
[1]*Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA;* [2]*Department of Agronomy, Yangzhou University, Yangzhou 225009, China*

Endosperm, which is derived from two polar nuclei fusing with one sperm, is a triploid tissue in cereals. Endosperm tissue determines the grain quality of cereals. Improving grain quality is one of the important breeding objectives in cereals. However, current statistical methods for mapping quantitative trait loci (QTL) under diploid genetic control have not been effective for dealing with endosperm traits because of the complexity of their triploid inheritance. In this paper, we derive for the first time the conditional probabilities of F_3 endosperm QTL genotypes given different flanking marker genotypes in F_2 plants. Using these probabilities, we develop a multiple linear regression method implemented via the iteratively reweighted least-squares (IRWLS) algorithm and a maximum likelihood method (ML) implemented via the expectation-maximization (EM) algorithm to map QTL underlying endosperm traits. We use the mean value of endosperm traits of F_3 seeds as the dependent variable and the expectations of genotypic indicators for additive and dominance effect of a putative QTL flanked by a pair of markers as independent variables for IRWLS mapping. However, if an endosperm trait is measured quantitatively using a single endosperm sample, the ML mapping method can be used to separate the two dominance effects. Efficiency of the methods is verified through extensive Monte Carlo simulation studies. Results of simulation show that the proposed methods provide accurate estimates of both the QTL effects and locations with very high statistical power. With these methods, we are now ready to map endosperm traits, as we can for regular quantitative trait under diploid control.
Heredity (2003) **90,** 228–235. doi:10.1038/sj.hdy.6800217

Keywords: endosperm traits; iteratively reweighted least-squares method; maximum likelihood estimation; quantitative trait locus; triploid

Introduction

Improving grain quality is one of the important objectives in cereal breeding (Sadimantara *et al*, 1997; Mazur *et al*, 1999; Tan *et al*, 1999; Wang *et al*, 2001). Many grain quality traits, such as amylose content and gel consistency in rice, protein and amino-acid content in wheat, starch and gum content in barley, and sugar content in sweet corn, are endosperm traits. The phenotypes of most endosperm traits are actually distributed in a continuous fashion and their expression is modified by environmental variables. It is not efficient to study their genetic architectures based on a single gene model. Therefore, we need quantitative genetic models to describe the expression of most endosperm traits. Quantitative genetic models have been developed for endosperm traits (Gale, 1976; Mo, 1987; Bogyo *et al*, 1988; Foolad and Jones, 1992; Pooni *et al*, 1992; Zhu and Weir, 1994; Wu *et al*, 1998). However, they are not designed for QTL mapping using molecular markers.

Recent advances in molecular biology provide tools that can generate saturated molecular markers along the genome. These markers segregate and are inherited based on simple Mendelian laws. They can be used to infer the segregation of quantitative trait loci (QTL) located in the neighborhood of the markers, a technique called QTL mapping. Numerous statistical methods have been developed for mapping QTL using simple crosses derived from two inbred lines (Lander and Botstein, 1989; Haley and Knott, 1992; Martinez and Curnow, 1992; Jansen, 1993, 1994; Zeng, 1994; Xu, 1998a, b; Kao and Zeng, 1997). However, these QTL mapping statistics are almost exclusively designed for traits under diploid control. They may not be appropriate for mapping QTL for triploid endosperm traits. Genetically, endosperm traits have several unique properties different from those of diploid traits. First, the endosperm is triploid and has a more complicated genetic constitution than the diploid plant. For a locus with two alleles, *Q* and *q*, four genotypes, *QQQ*, *QQq*, *Qqq* and *qqq*, are possible (Mo, 1987; Bogyo *et al*, 1988), whereas a diploid plant has only three possible genotypes. Second, the occurrence of the fertilized egg is the beginning of a new generation, so that the embryo and endosperm of a plant represent the next generation. Third, the endosperm genotype of a hybrid coming from one mating will differ from that of the reciprocal hybrid. Finally, each single endosperm has an independent genotype, which may be different from each other, and thus endosperm traits are separated based on seeds, whereas diploid traits in plants are separated based on plants. These unique differences associated with endosperm traits should be given sufficient consideration in mapping QTL.

Correspondence: S Xu, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA.
E-mail: xu@genetics.ucr.edu
Received 4 June 2002; accepted 15 October 2002

In this study, we develop a statistical method for mapping QTL with attention particularly paid to these unique properties of endosperm traits. We assume that the genetic variance of an endosperm trait is controlled by the segregation of the triploid genome of the endosperm rather than the diploid genome of the maternal plant. The genotypes of the maternal plant, however, are used only for inferring the genotypes of the endosperm.

Theory and methods

Genetic models for endosperm traits

Based on the genetic characteristics of endosperm traits, various genetic models have been proposed to partition the genetic effects of endosperm traits (Gale, 1976; Mo, 1987; Bogyo *et al*, 1988; Foolad and Jones, 1992; Pooni *et al*, 1992; Zhu and Weir, 1994; Wu *et al*, 1998). In order for the paper to be self-contained, these models are summarized and described here. Consider the simplest case of one quantitative trait locus and two alternative alleles, Q and q, with increasing and decreasing effects, respectively. There are four possible genotypes and three genetic effects, a, d_1 and d_2, where a is the mean substitution effect of Q to q (called the additive effect), d_1 is the interaction effect of QQ and q (called the first dominant effect), d_2 is the interaction effect of Q and qq (called the second dominant effect). The four possible endosperm genotypes and their genotypic values are defined as follows: $G_{QQQ} = \mu + \frac{3}{2}a$, $G_{QQq} = \mu + \frac{1}{2}a + d_1$, $G_{Qqq} = \mu - \frac{1}{2}a + d_2$ and $G_{qqq} = \mu - \frac{3}{2}a$, where μ is the mid-point or mean of the two homozygotes QQQ and qqq.

Linear model

Let y_{ij} be the phenotypic value of the jth endosperm on the ith F_2 plant, which can be described by the following linear model:

$$y_{ij} = x_{oij}\mu + x_{1ij}a + x_{2ij}d_1 + x_{3ij}d_2 + \varepsilon_{ij} \quad (1)$$

where ε_{ij} is the residual error distributed as $N(0, \sigma_\varepsilon^2)$. Note that if the trait is indeed controlled by a single QTL and the genotype of the QTL is observed for every individual, the residual error purely reflects the random environmental noise. If the trait is controlled by multiple QTL, the residual error will contain the effects of other QTL not included in the model in addition to the environmental error. The independent variables, x_{0ij}, x_{1ij}, x_{2ij} and x_{3ij}, are defined as follows. For any genotype, $x_{0ij} = 1$. For genotype QQQ, $x_{1ij} = \frac{3}{2}$ and $x_{2ij} = x_{3ij} = 0$; for genotype QQq, $x_{1ij} = \frac{1}{2}, x_{2ij} = 1$ and $x_{3ij} = 0$, for genotype Qqq, $x_{1ij} = -\frac{1}{2}, x_{2ij} = 0$ and $x_{3ij} = 1$ and for genotype qqq, $x_{1ij} = -\frac{3}{2}, x_{2ij} = 0$ and $x_{3ij} = 0$. The above model cannot be taken as a working model because (i) x_{1ij}, x_{2ij} and x_{3ij} are missing because of the inability to observe the QTL genotype and (ii) most endosperm traits cannot be measured quantitatively using a single endosperm sample, rather a mixed endosperm sample of many seeds is collected and measured together from a single plant. As a result, we need a working model to perform the estimation and test. The working model is

$$\begin{aligned} y_i =& E(x_{0i}|I_M)\mu + E(x_{1i}|I_M)a + E(x_{2i}|I_M)d_1 \\ &+ E(x_{3i}|I_M)d_2 + e_i \end{aligned} \quad (2)$$

where

$$y_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij},\ x_{0i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{0ij},\ x_{1i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{1ij},$$

$$x_{2i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{2ij} \quad \text{and} \quad x_{3i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{3ij}$$

are the mean value of the y_{ij}, x_{0ij}, x_{1ij}, x_{2ij} and x_{3ij}, respectively, for n_i seeds sampled from the ith maternal plant (a single F_2 individual), I_M stands for the F_2 plant marker information. $E(x_{0i}|I_M), E(x_{1i}|I_M), E(x_{2i}|I_M)$ and $E(x_{3i}|I_M)$ are the expectations of x_{0i}, x_{1i}, x_{2i} and x_{3i} conditional on marker information, respectively, and e_i is the residual error, different from ε_{ij}. Note that $E(x_{0i}\,|\,I_M) = 1$ for all i.

Let $\mathbf{X}_i = (x_{0i} \quad x_{1i} \quad x_{2i} \quad x_{3i})$ and $\mathbf{b} = (\mu\ a\ d_1\ d_2)^T$. The model can be expressed in matrix notation as

$$y_i = E(\mathbf{X}_i|I_M)\mathbf{b} + e_i \quad (3)$$

The expectation and variance of model (3) are

$$\begin{aligned} E(y_i|I_M) &= E(\mathbf{X}_i|I_M)\mathbf{b} \quad \text{and} \quad V(y_i|I_M) = V(e_i) \\ &= \mathbf{b}^T V(\mathbf{X}_i|I_M)\mathbf{b} + \sigma_\varepsilon^2/n_i = R_{ii}\sigma_\varepsilon^2 \end{aligned}$$

where

$$R_{ii} = \frac{1}{\sigma_\varepsilon^2}\mathbf{b}^T V(\mathbf{X}_i|I_M)\mathbf{b} + \frac{1}{n_i}$$

To derive the conditional expectations and variances, we need the conditional probabilities of the four possible genotypes of an endosperm given marker information of the maternal plant, denoted by $\mathbf{P}_i = (p_{i(111)}\ p_{i(110)}\ p_{i(100)}\ p_{i(000)})^T$ for the four endosperm genotypes in the order of QQQ, QQq, Qqq and qqq for plant i. Let $M_1M_1QQM_2M_2$ and $m_1m_1qqm_2m_2$ be the joint genotypes of two inbred lines, P_1 and P_2, for three loci (two flanking markers and one QTL). Let r be the recombination fraction between the two markers, and r_1 and r_2 be the recombination fractions of the QTL with the two markers. The F_1 hybrid plant has a genotype of $M_1m_1QqM_2m_2$, which produces eight possible gametes if no interference is assumed. From these gametes, we can easily derive the QTL genotypes and their frequencies of the F_2 plant conditional on flanking markers (Haley and Knott, 1992). Note that we are now to map QTL underlying endosperm trait instead of common diploid plant trait. Therefore, we need to derive the corresponding conditional probabilities of QTL genotypes of the F_3 endosperms from an F_2 maternal plant conditional on marker genotypes of the F_2 plant. These conditional probabilities are given in Table 1. If the QTL genotype of the F_2 plant is QQ, all the F_3 endosperms are QQQ; if the F_2 QTL genotype is qq, all the F_3 endosperm are qqq; if the QTL genotype of the F_2 plant is Qq, there are four possible endosperm QTL genotypes, QQQ, QQq, Qqq and qqq, with an equal probability (1/4). These probabilities are combined with the conditional probabilities of conventional diploid QTL mapping to form the corresponding conditional probabilities of F_3 endosperm QTL (Table 1). From this table we can find $\mathbf{P}_i$ for individual i and are now ready to derive the variances and covariances of the x variables.

Table 1 Conditional probabilities of the triploid endosperm QTL genotypes of F_3 seeds, given the diploid marker genotypes on the maternal plants

Marker genotype for F_2 plant	*Conditional probabilities of endosperm QTL genotypes for F_3 seed*			
	$p_{(111)} = Pr(QQQ \mid I_M)$	$p_{(110)} = Pr(QQq \mid I_M)$	$p_{(100)} = Pr(Qqq \mid I_M)$	$p_{(000)} = Pr(qqq \mid I_M)$
$M_1M_1M_2M_2$	$\frac{2(1-r_1)^2(1-r_2)^2 + r_1(1-r_1)r_2(1-r_2)}{2(1-r)^2}$	$\frac{r_1(1-r_1)r_2(1-r_2)}{2(1-r)^2}$	$\frac{r_1(1-r_1)r_2(1-r_2)}{2(1-r)^2}$	$\frac{2r_1^2r_2^2 + r_1(1-r_1)r_2(1-r_2)}{2(1-r)^2}$
$M_1M_1M_2m_2$	$\frac{4(1-r_1)^2r_2(1-r_2) + r_1(1-r_1)[r_2^2 + (1-r_2)^2]}{4r(1-r)}$	$\frac{r_1(1-r_1)[r_2^2 + (1-r_2)^2]}{4r(1-r)}$	$\frac{r_1(1-r_1)[r_2^2 + (1-r_2)^2]}{4r(1-r)}$	$\frac{4r_1^2r_2(1-r_2) + r_1(1-r_1)[r_2^2 + (1-r_2)^2]}{4r(1-r)}$
$M_1M_1m_2m_2$	$\frac{2(1-r_1)^2r_2^2 + r_1(1-r_1)r_2(1-r_2)}{2r^2}$	$\frac{r_1(1-r_1)r_2(1-r_2)}{2r^2}$	$\frac{r_1(1-r_1)r_2(1-r_2)}{2r^2}$	$\frac{2r_1^2(1-r_2)^2 + r_1(1-r_1)r_2(1-r_2)}{2r^2}$
$M_1m_1M_2M_2$	$\frac{4r_1(1-r_1)(1-r_2)^2 + [r_1^2 + (1-r_1)^2]r_2(1-r_2)}{4r(1-r)}$	$\frac{[r_1^2 + (1-r_1)^2]r_2(1-r_2)}{4r(1-r)}$	$\frac{[r_1^2 + (1-r_1)^2]r_2(1-r_2)}{4r(1-r)}$	$\frac{4r_1(1-r_1)r_2^2 + [r_1^2 + (1-r_1)^2]r_2(1-r_2)}{4r(1-r)}$
$M_1m_1M_2m_2$	$\frac{8r_1(1-r_1)r_2(1-r_2) + [r_1^2 + (1-r_1)^2][r_2^2 + (1-r_2)^2]}{4[r^2 + (1-r)^2]}$	$\frac{[r_1^2 + (1-r_1)^2][r_2^2 + (1-r_2)^2]}{4[r^2 + (1-r)^2]}$	$\frac{[r_1^2 + (1-r_1)^2][r_2^2 + (1-r_2)^2]}{4[r^2 + (1-r)^2]}$	$\frac{8r_1(1-r_1)r_2(1-r_2) + [r_1^2 + (1-r_1)^2][r_2^2 + (1-r_2)^2]}{4[r^2 + (1-r)^2]}$
$M_1m_1m_2m_2$	$\frac{4r_1(1-r_1)r_2^2 + [r_1^2 + (1-r_1)^2]r_2(1-r_2)}{4r(1-r)}$	$\frac{[r_1^2 + (1-r_1)^2]r_2(1-r_2)}{4r(1-r)}$	$\frac{[r_1^2 + (1-r_1)^2]r_2(1-r_2)}{4r(1-r)}$	$\frac{4r_1(1-r_1)(1-r_2)^2 + [r_1^2 + (1-r_1)^2]r_2(1-r_2)}{4r(1-r)}$
$m_1m_1M_2M_2$	$\frac{2r_1^2(1-r_2)^2 + r_1(1-r_1)r_2(1-r_2)}{2r^2}$	$\frac{r_1(1-r_1)r_2(1-r_2)}{2r^2}$	$\frac{r_1(1-r_1)r_2(1-r_2)}{2r^2}$	$\frac{2(1-r_1)^2r_2^2 + [r_1^2 + (1-r_1)^2]r_2(1-r_2)}{2r^2}$
$m_1m_1M_2m_2$	$\frac{4r_1^2r_2(1-r_2) + r_1(1-r_1)[r_2^2 + (1-r_2)^2]}{4r(1-r)}$	$\frac{r_1(1-r_1)[r_2^2 + (1-r_2)^2]}{4r(1-r)}$	$\frac{r_1(1-r_1)[r_2^2 + (1-r_2)^2]}{4r(1-r)}$	$\frac{4(1-r_1)^2r_2(1-r_2) + r_1(1-r_1)[r_2^2 + (1-r_2)^2]}{4r(1-r)}$
$m_1m_1m_2m_2$	$\frac{2r_1^2r_2^2 + r_1(1-r_1)r_2(1-r_2)}{2(1-r)^2}$	$\frac{r_1(1-r_1)r_2(1-r_2)}{2(1-r)^2}$	$\frac{r_1(1-r_1)r_2(1-r_2)}{2(1-r)^2}$	$\frac{2(1-r_1)^2(1-r_2)^2 + r_1(1-r_1)r_2(1-r_2)}{2(1-r)^2}$

Define a 4×4 matrix $\mathbf{H}$ as

$$\mathbf{H} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 3/2 & 1/2 & -1/2 & -3/2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}^{\mathrm{T}}$$

and let $\mathbf{H}_l$ be the lth row of matrix $\mathbf{H}$. We can now connect $\mathbf{X}_i$ with $\mathbf{H}$ by $\mathbf{X}_i = \mathbf{H}_l$ for $l = 1, \ldots 4$, if an endosperm from plant i takes the lth ordered triploid genotype. With the above definitions, we have

$$E(\mathbf{X}_i|I_{\mathrm{M}}) = \mathbf{H}^{\mathrm{T}}\mathbf{P}_i \quad \text{and} \quad V(\mathbf{X}_i|I_{\mathrm{M}}) = \frac{1}{n_i}\mathbf{H}^{\mathrm{T}}[\mathrm{diag}(\mathbf{P}_i) - \mathbf{P}\mathbf{P}_i^{\mathrm{T}}]\mathbf{H}$$

where $\mathrm{diag}(\mathbf{P}_i)$ denotes a diagonal matrix with the diagonal elements filled with vector $\mathbf{P}_i$.

Weighted least-squares estimation

Define $\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \mathbf{X}_2^{\mathrm{T}}, \ldots, \mathbf{X}_k^{\mathrm{T}})^{\mathrm{T}}$ as a $k \times 4$ design matrix for all the k plants, $\mathbf{U} = E(\mathbf{X}|I_{\mathrm{M}})$, $\mathbf{R} = \mathrm{diag}(R_{11}, R_{22}, \ldots, R_{kk})$ as a diagonal matrix and $\mathbf{y} = (y_1, y_2, \ldots, y_k)^{\mathrm{T}}$. The parameters can be estimated using the iteratively reweighted least-squares (IRWLS) method (Xu, 1998a,b). Given an initial guess of $\mathbf{b}$ and σ_ε^2, matrix $\mathbf{R}$ is treated as known. Conditional on $\mathbf{R}$, the solutions for the parameters are

$$\hat{\mathbf{b}} = (\mathbf{U}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{y}$$

and

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{k-4}(\mathbf{y} - \mathbf{U}\mathbf{b})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{U}\mathbf{b})$$

Since $\mathbf{R}$ depends on unknown parameters, it must be updated by the estimates of the parameters and the estimation is then repeated until a certain criterion of convergence has been reached.

The variance–covariance matrix of the estimate $\mathbf{b}$ is

$$V(\hat{\mathbf{b}}) = (\mathbf{U}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{U})^{-1}\hat{\sigma}_\varepsilon^2$$

The variance–covariance matrix of $\hat{\mathbf{b}}$ is used to construct the test statistic for QTL detection. For example, to test the hypothesis $H_0 : a = d_1 = d_2 = 0$, we define

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

so that the null hypothesis can be redefined as $H_0 : \mathbf{L}^{\mathrm{T}}\hat{\mathbf{b}} = \mathbf{0}$. The test statistic for this hypothesis is $F = L\hat{\mathbf{b}}^{\mathrm{T}}[\mathbf{L}^{\mathrm{T}}V(\hat{\mathbf{b}})\mathbf{L}]^{-1}\hat{\mathbf{b}}L^{\mathrm{T}}$. Under the null hypothesis this test statistic will follow approximately a χ^2 distribution with three degrees of freedom. Other hypotheses can be tested by redefining matrix $\mathbf{L}$. For example, to test $H_0 : a = 0$, one simply defines a new $\mathbf{L}$ matrix as $\mathbf{L} = (0 \;\; 1 \;\; 0 \;\; 0)^{\mathrm{T}}$. Xu (1998a, b) showed that the above test statistic is very close to a likelihood ratio test statistic and thus can be converted into an LOD score as usually done in QTL mapping. The relation between the F-like statistic and the LOD score is $\mathrm{LOD} = F/4.61$, which is used in the following simulation studies.

Note that because $E(x_{2i}|I_{\mathrm{M}}) = E(x_{3i}|I_{\mathrm{M}})$ in the F_3 endosperm generation, the two dominance effects, d_1 and d_2, cannot be estimated separately by IRWLS method; rather, they are combined as a single dominance effect. If we want to estimate these two dominance effects separately, we must measure the endosperm trait for each seed and then use a maximum likelihood (ML) method.

ML estimation

The ML method can be implemented via the EM algorithm. If $\mathbf{X}_j$ were observed for every endosperm individual, the MLE of the parameters could be found explicitly in a single step using the following equations:

$$\hat{\mathbf{b}} = \left[\sum_{j=1}^{N}\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j\right]^{-1}\left[\sum_{j=1}^{N}\mathbf{X}_j^{\mathrm{T}}y_j\right]$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N}\sum_{j=1}^{N}(y_j - \mathbf{X}_j\hat{\mathbf{b}})^2 \qquad (4)$$

where N is the number of endosperms measured for the entire mapping population. For example, if there are k plants each with n seeds, then $N = kn$.

In the case where $\mathbf{X}_j$ is missing but the distribution of $\mathbf{X}_j$ is given, the EM algorithm can be adopted to take advantage of the above equations. The EM equations simply replace all the terms related to $\mathbf{X}_j$ by their expectations, that is,

$$\hat{\mathbf{b}} = \left[\sum_{j=1}^{N}E(\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j)\right]^{-1}\left[\sum_{j=1}^{N}E(\mathbf{X}_j^{\mathrm{T}}y_j)\right]$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N}\sum_{j=1}^{N}E[(y_j - \mathbf{X}_j\mathbf{b})^2] \qquad (5)$$

The expectations are obtained conditional on both marker information and the phenotypic value y_j. The connection between the phenotype and the QTL genotype is through the three genetic parameters, but the parameters are what we are trying to find. Therefore, we need iterations on equation (5) by providing some initial values of the parameters to start the iteration. This is the EM algorithm. The E-step is to find the expectations and the M-step is to invoke equation (5) for iterations.

Denote the probability of $\mathbf{X}_j$ conditional on marker information by $\Pr(\mathbf{X}_j = \mathbf{H}_l|\boldsymbol{I}_{\mathrm{M}})$. This probability is simply $\mathbf{P}_j = (p_{j(111)}\; p_{j(110)}\; p_{j(100)}\; p_{j(000)})^{\mathrm{T}}$ and may be called the prior probability. After incorporating the phenotypic value, we obtain the posterior probability, denoted by

$$\Pr(\mathbf{X}_j = \mathbf{H}_l|\boldsymbol{I}_{\mathrm{M}}, y_j) = \frac{\Pr(\mathbf{X}_j = \mathbf{H}_l|\boldsymbol{I}_{\mathrm{M}})f(y_j - \mathbf{H}_l\mathbf{b})}{\sum_{l=1}^{4}\Pr(\mathbf{X}_j = \mathbf{H}_l|\boldsymbol{I}_{\mathrm{M}})f(y_j - \mathbf{H}_l\mathbf{b})}$$

where

$$f(y_j - \mathbf{H}_l\mathbf{b}) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}}\exp\left[-\frac{1}{2\sigma_\varepsilon^2}(y_j - \mathbf{H}_l\mathbf{b})^2\right]$$

The expectations are actually obtained using the posterior probabilities rather than the prior probabilities.

Table 2 Number of makers per chromosome and genetic distances between consecutive markers simulated for the genome consisting of 12 chromosomes in design II

Chr	*No. of markers*	*Chr. length (cM)*	*Genetic distance between flanking markers (cM)*
1	12	151.9	17.3, 13.3, 17.8, 10.2, 21.4, 7.3, 11.8, 14.1, 12.4, 8.2, 18.1
2	8	105.7	9.8, 19.6, 12.6, 9.2, 18.5, 17.4, 18.6
3	9	132.4	16.8, 9.0, 17.8, 19.8, 17.4, 16.0, 16.9, 18.7
4	13	203.4	17.0, 20.0, 20.4, 15.9, 22.0, 14.9, 14.7, 16.3, 19.6, 19.8, 14.0, 8.8
5	15	206.0	7.9, 15.5, 22.6, 17.8, 19.9, 7.9, 15.1, 18.4, 14.6, 18.7, 7.1, 13.8, 14.0, 12.7
6	8	105.0	8.6, 17.1, 11.2, 13.4, 18.1, 17.7, 18.9
7	11	148.5	15.1, 13.6, 14.1, 11.6, 15.9, 18.2, 17.1, 9.1, 13.8, 20.0
8	7	90.3	20.3, 4.7, 19.0, 15.2, 13.2, 17.9
9	9	138.4	15.2, 19.5, 13.7, 16.8, 17.3, 22.2, 18.9, 14.8
10	11	169.5	10.6, 19.9, 19.0, 17.2, 16.7, 15.0, 18.7, 17.9, 21.0, 13.5
11	12	146.1	11.2, 12.3, 15.8, 12.3, 15.6, 14.4, 9.1, 12.8, 12.8, 9.7, 20.1
12	10	116.5	9.3, 15.1, 15.2, 12.2, 16.2, 15.0, 13.8, 10.0, 9.7

Table 3 Locations and sizes of QTL simulated in design II

QTL	*Chromosome*	*Position (cM)*	*Additive effect*	*Dominance effect*		σ_q^2
			a	d_1	d_2	
*qtl*1	3	60.0	0.76	0	0	1.01
*qtl*2	5	102.8	−1.00	0.95	0.79	1.87
*qtl*3	7	105.8	−0.68	0.72	1.05	0.99
*qtl*4	10	66.0	1.03	0.54	−0.60	2.08
*qtl*5	10	134.3	1.27	1.05	1.20	3.04
*qtl*6	12	78.6	1.41	1.38	0.50	3.85

Therefore,

$$\sum_{j=1}^{N} E(\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j) = \sum_{j=1}^{N}\left(\sum_{l=1}^{4}\Pr(\mathbf{X}_j = \mathbf{H}_l|\boldsymbol{I}_{\mathrm{M}}, y_j)\mathbf{H}_l^{\mathrm{T}}\mathbf{H}_l\right)$$

$$\sum_{j=1}^{N} E(\mathbf{X}_j^{\mathrm{T}}y_j) = \sum_{j=1}^{N}\left(\sum_{l=1}^{4}\Pr(\mathbf{X}_j = \mathbf{H}_l|\boldsymbol{I}_{\mathrm{M}}, y_j)\mathbf{H}_l^{\mathrm{T}}y_j\right)$$

and

$$\sum_{j=1}^{N} E(y_j - \mathbf{X}_j\mathbf{b})^2 = \sum_{j=1}^{N}\left(\sum_{l=1}^{4}\Pr(\mathbf{X}_j = \mathbf{H}_l|\boldsymbol{I}_{\mathrm{M}}, y_j)(y_j - \mathbf{H}_l\mathbf{b})^2\right)$$

Likelihood ratio test

Define the log-likelihood value evaluated at the MLE of parameters as

$$L(\hat{\mathbf{b}}, \hat{\sigma}_\varepsilon^2) = \sum_{j=1}^{N}\log\left[\sum_{l=1}^{4}\Pr(\mathbf{X}_j = \mathbf{H}_l|\boldsymbol{I}_{\mathrm{M}})f(y_j - \mathbf{H}_l\hat{\mathbf{b}})\right]$$

This is also called the likelihood value under the full model. We need the likelihood value under $H_0: \mathbf{L}^{\mathrm{T}}\mathbf{b} = \mathbf{0}$ (the restricted model) to test null hypothesis of no genetic effects. Let the likelihood value under the restricted model be

$$L(\hat{\mu}, \hat{\sigma}_\varepsilon^2) = L(\hat{\mathbf{b}}, \hat{\sigma}_\varepsilon^2|\mathbf{L}\mathbf{b} = \mathbf{0})$$

The likelihood ratio test statistic is

$$\begin{aligned}\Lambda &= -2[L(\hat{\mu}, \hat{\sigma}_\varepsilon^2) - L(\hat{\mathbf{b}}, \hat{\sigma}_\varepsilon^2)] \\ &= -2[L(\hat{\mathbf{b}}, \hat{\sigma}_\varepsilon^2|\mathbf{L}\mathbf{b} = \mathbf{0}) - L(\hat{\mathbf{b}}, \hat{\sigma}_\varepsilon^2)]\end{aligned} \quad (6)$$

Simulation studies

Designs of simulation experiments

We designed two simulation experiments, one for a single chromosome (design I) and the other for an entire genome with 12 chromosomes (design II). In design I, the single chromosome is covered by a given number of evenly spaced codominant markers covering 150 cM. A single QTL is located at position 75 cM on the chromosome with the following effects: $a = 4$, $d_1 = d_2 = 2$ and $\mu = 20$. We used various sizes of residual variance to control the desired levels of the heritabilities. Factors considered include (1) marker density, (2) QTL heritability, (3) population size and (4) number of endosperms per plant. Marker density was simulated at two levels: 16 and six markers, which correspond to distances of 10 and 30 cM, respectively, between consecutive markers. QTL heritability was simulated at two levels: 10 and 30%. Sample size of the F_2 population was simulated at three levels: 50, 100 and 200. Number of endosperms collected per plant was simulated at three levels: 5, 10 and 20. The total number of treatment combinations is $2 \times 2 \times 3 \times 3 = 36$.

In design II, a genome consisting of 12 chromosomes was simulated. The number of markers per chromosome and the marker locations were generated randomly. The simulated linkage map information is listed in Table 2. This setup actually mimics the linkage map of the rice genome. In this particular experiment, we simulated six QTL distributed along five of the 12 chromosomes. Their sizes and locations are given in Table 3. The total phenotypic variance of the endosperm trait explained by the six QTL is 50%. The population size is 200 and the number of seeds per plant is 20.

Each treatment combination of the simulation experiments was repeated 100 times. The standard deviation of an estimated parameter among the 100 replicates

provides a measure of the standard error of parameter estimation. The statistical power is determined by counting the number of replicates that have a test statistic (LOD score) greater than the empirical critical values obtained from analyses of 1000 additional samples simulated under the null model (zero heritability).

Results

Table 4 shows the means and standard deviations of the estimated QTL effects and locations as well as the empirical powers calculated from 100 repeated simulations under different marker density, heritabilities and sampling strategies in design I. The results show the general trends of expectation: denser marker maps, higher QTL heritabilities and larger sample sizes tend to produce more accurate and precise estimates, and lower heritabilities, especially with smaller sample sizes, produces less accurate estimates with large estimation errors. In addition, the estimated positions of QTL tend to be biased toward the middle of the chromosome. Estimates of the additive and dominant effects are reasonably close to the true value. It is surprising to see how powerful the methods are in detecting the QTL. Even though the QTL only explains 10% of the trait variation and only 50 F_2 plants, the power is almost 100% when the number of endosperms is 20. When the sample size of F_2 population is 100 or more, the powers are all close to 100% in different treatments. Only two out of the 36 treatment combinations show a statistical power less than 80%.

The mean estimates and standard deviations of locations and effects of the six QTL and corresponding statistical powers in design II are listed in Table 5. From this table, we see that only two out of the six QTL, *qtl*1 and *qtl*3, have powers less than 100% (72 and 71%, respectively, for the two QTL) and all other QTL have 100% power. This is not surprising because the heritabilities of the two QTL are less than 5%. The effect and position estimates of the four larger QTL are fairly accurate and precise. The estimated location of *qtl*4 shows some deviation from the true value. This may be explained by the fact that there are two QTLs on chromosome 10 and *qtl*4 is smaller than *qtl*5 on the same chromosome. The LOD score profiles of a random sample out of the 100 random samples are shown in

Table 4 Means and standard deviation (in parentheses) of the QTLs parameters for different levels of marker density, heritabilities and sample strategies under design I

Marker density (cM)	h^2(%)	*Population size*	*No. of endosperms*	*Power (%)*	*Position (cM)*	*a*	d_1+d_2	*Critical values*
		True value			75.0	4.0	4.0	
			5	74	72.39(19.06)	4.15(0.89)	4.64(11.05)	2.65
		50	10	97	75.40(9.05)	4.14(0.71)	4.66(8.42)	2.49
			20	100	74.96(4.76)	4.08(0.51)	4.66(5.08)	2.42
	10		5	96	76.30(7.56)	3.96(0.70)	4.35(8.60)	2.61
		100	10	100	75.05(5.89)	3.94(0.58)	4.03(5.40)	2.54
			20	100	75.16(2.33)	4.04(0.35)	4.17(3.25)	2.56
			5	100	75.22(3.30)	4.04(0.52)	4.19(4.49)	2.41
10		200	10	100	74.80(1.92)	4.02(0.39)	4.23(3.11)	2.55
			20	100	75.02(1.50)	4.00(0.27)	4.15(2.49)	2.41
			5	100	74.92(4.62)	4.09(0.57)	5.90(5.15)	2.58
		50	10	100	74.49(4.24)	3.99(0.41)	4.44(3.76)	2.61
			20	100	75.09(2.18)	3.97(0.28)	3.57(2.82)	2.65
	30		5	100	74.98(2.43)	3.97(0.36)	3.83(3.72)	2.63
		100	10	100	74.79(1.59)	3.98(0.29)	4.09(2.42)	2.50
			20	100	74.91(1.84)	4.00(0.21)	3.94(1.91)	2.57
			5	100	74.99(1.63)	3.96(0.26)	4.15(2.26)	2.43
		200	10	100	74.99(1.11)	3.98(0.18)	4.02(1.78)	2.49
			20	100	74.95(1.01)	3.99(0.15)	3.85(1.36)	2.39
			5	55	75.15(17.72)	4.67(1.17)	6.66(14.89)	2.50
		50	10	84	73.65(15.99)	4.02(0.75)	3.89(10.45)	2.40
			20	98	75.48(8.71)	3.93(0.61)	4.01(6.88)	2.32
	10		5	90	74.32(11.28)	4.01(0.82)	4.01(8.99)	2.19
		100	10	99	73.03(8.29)	3.93(0.66)	3.83(6.64)	2.26
			20	100	74.34(4.36)	3.93(0.44)	3.72(5.14)	2.32
			5	100	75.73(6.23)	3.87(0.66)	4.61(5.88)	2.22
30		200	10	100	75.01(4.14)	3.91(0.37)	4.11(3.93)	2.19
			20	100	74.76(2.94)	3.99(0.33)	3.60(3.42)	2.18
			5	98	74.30(10.09)	3.97(0.72)	3.77(6.71)	2.34
		50	10	99	74.37(8.79)	3.99(0.56)	3.46(6.26)	2.25
			20	100	74.93(6.65)	3.91(0.44)	4.75(4.42)	2.37
	30		5	100	75.13(5.52)	3.88(0.49)	3.15(5.19)	2.20
		100	10	100	74.88(4.48)	4.00(0.33)	4.13(3.64)	2.25
			20	100	74.26(2.63)	3.97(0.30)	4.13(3.64)	2.33
			5	100	75.24(2.88)	3.92(0.34)	3.71(3.91)	2.28
		200	10	100	75.01(2.19)	3.94(0.26)	3.99(2.99)	2.40
			20	100	74.86(2.10)	3.95(0.19)	3.91(2.22)	2.40

Table 5 Means and standard deviations of the QTL parameters under design II

QTL	*Chr*	*Power (%)*	*Position (cM)*		*a*		d_1+d_2	
			Mean	*Std*	*Mean*	*Std*	*Mean*	*Std*
*qtl*1	3	72	60.85	8.38	0.87	0.14	0.08	1.89
*qtl*2	5	99	104.16	4.61	−0.99	0.18	1.87	1.50
*qtl*3	7	71	106.01	5.91	−0.79	0.12	2.23	1.58
*qtl*4	10	100	75.18	7.81	1.49	0.16	0.33	1.54
*qtl*5	10	100	131.73	3.81	1.64	0.16	2.46	1.47
*qtl*6	12	100	79.21	3.09	1.40	0.16	1.66	1.42

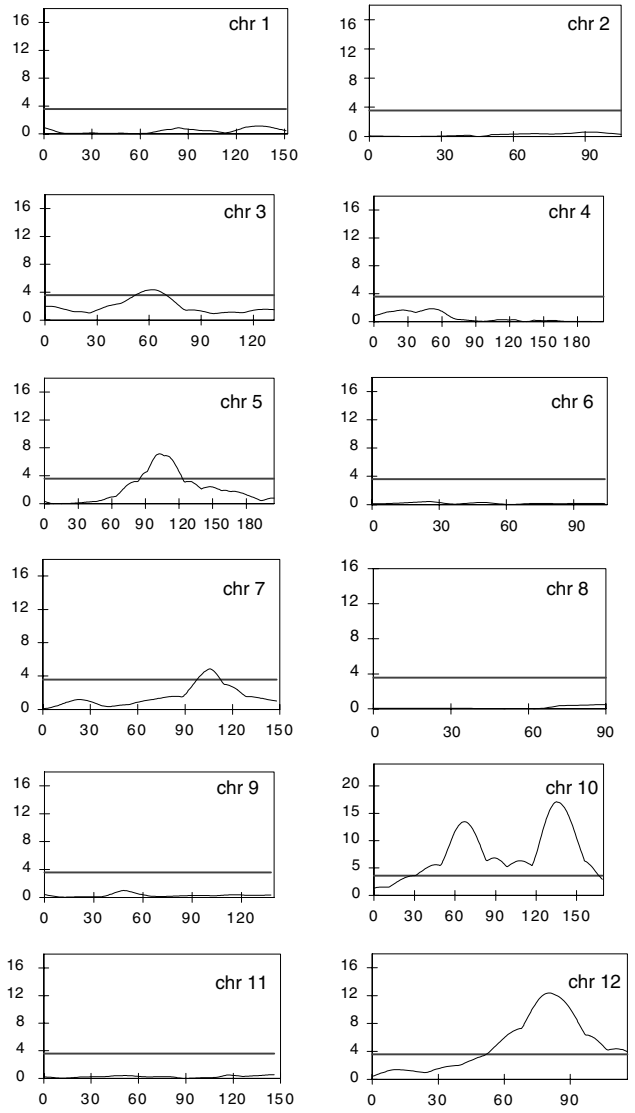

Figure 1 LOD score profiles of QTL detection from a random sample of simulations under design II. The horizontal axis is the map position and vertical axis is the LOD score.

Figure 1. This figure shows an expected output in real data analysis for endosperm mapping.

For the data simulated from design I, we also performed the ML analysis for some treatments, hoping to separate the two dominance effects. The main results are in accordance with those reported above by the IRWLS method. However, the ML method was indeed able to separate the two dominance effects. For example, when the marker density is 10 cM, QTL heritability is 10%, population size is 100 and number of F_3 endosperm per plant is 10, the empirical power is 100%, for the ML method. The MLE of QTL position, additive effect, and first and second dominance effect are 74.37 ± 3.61, 3.99 ± 0.33, 2.26 ± 2.12 and 1.75 ± 2.31, respectively.

Discussion

Endosperm traits belong to a group of characters that determine the grain quality in cereals and they are tremendously important to human nutrition. Genetic improvement of such endosperm traits has received considerable attention in plant breeding (Benner *et al*, 1989; Sadimantara *et al*, 1997; Mazur *et al*, 1999). Quantitative genetics models for analyzing the triploid inheritance of endosperm traits have been developed and applied to practical breeding populations in cereals (Gale, 1976; Mo, 1987; Bogyo *et al*, 1988; Foolad and Jones, 1991; Pooni *et al*, 1992; Zhu and Weir, 1994; Wu *et al*, 1998). However, these traditional methods were not designed for QTL mapping; rather they were developed for analyzing the overall contribution of the genetic variance to the phenotypic variance. Traditional QTL mapping methods for diploid traits have been applied to mapping endosperm traits (Tan *et al*, 1999; Wang and Larkins, 2001; Wang *et al*, 2001). The assumption was that the genetic variance of an endosperm trait is controlled by the segregation of QTL in the diploid maternal plants. Two precautions should be clarified when a diploid mapping procedure is used for mapping a triploid trait. Firstly, the DNA markers detected (eg, using F_2 or BC plants) and traits measured (using F_3 or BC selfing seeds on F_2 or BC plants) are not measured in the same generation. Therefore, the application of a diploid mapping model to endosperm traits is identical to mapping the maternal effects of the trait in question. Secondly, there is no reason to believe that an endosperm trait is only controlled by the genotype of the maternal plant and that there is no contribution from the genotype of the endosperm tissue itself. For the first time, we here consider the triploid control mechanisms of the endosperm traits for QTL mapping and develop the appropriate probability model to infer the triploid genotype from the diploid marker genotypes of the maternal plant.

The proposed IRWLS mapping method is a second-order approximation to ML, was first proposed by Xu (1998a, b), and was demonstrated and compared with the simple linear regression (REG) method proposed by Haley and Knott (1992) and ML method proposed by Lander and Botstein (1989) via Monte Carlo simulation. We chose the IRWLS for two reasons: (1) it is faster than ML and better than the ordinary least-squares method, as shown by Xu (1998a, b); (2) the endosperm trait is measured as the average of several seeds and it is hard to model the average value using the mixed distribution model. However, if an endosperm trait is measured using a single endosperm sample, the ML mapping method can be used to estimate all genetic effects of endosperm QTL, including the two different dominance effects. Since the purpose of this study is not

to compare the efficiencies of different statistical methods, but to apply existing methods to map QTL for endosperm traits, we paid more attention to the genetic model and the implementation of the IRWLS considering that most endosperm traits cannot be measured using a single endosperm sample.

The next step of endosperm mapping is to consider both the maternal diploid genotype and the triploid endosperm genotypes jointly. The model will become more complicated because there will be five genetic effects involved, which are maternal additive effect, maternal dominant effect, endosperm additive, endosperm first dominance and endosperm second dominance effects. This project is currently under investigation and will be reported in a separate paper.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant 39900080) to CX and the National Institutes of Health Grant GM55321, and the USDA National Research Initiative Competitive Grants Program 00-35300-9245 to SX.

References

Benner MS, Phillips RL, Kirhara JA, Messing JW (1989). Genetic analysis of methionine-rich storage protein accumulation in maize. *Theor Appl Genet* **78**: 761–767.

Bogyo TP, M Lance RC, Chevalier P, Nilan RA (1988). Genetic models for quantitatively inherited endosperm characters. *Heredity* **60**: 61–67.

Foolad MR, Jones RA (1992). Models to estimate maternally controlled genetic variation in quantitative seed characters. *Theor Appl Genet* **83**: 360–366.

Gale MD (1976). High α-amylase breeding and genetical aspects of the problem. *Cereal Res Commun* **4**: 231–243.

Haley CS, Knott SA (1992). A simple regression method for maooing 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.

Kao CH, Zeng ZB (1997). General formulas for obtaining the MLEs and the asymptotic variance–covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* **53**: 653–665.

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

Martinez O, Curnow RN (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theor Appl Genet* **85**: 480–488 .

Mazur B, Krebbers E, Tingey S (1999). Gene discovery and product development for grain quality traits. *Science* **285**: 372–375.

Mo HD (1987). Genetic expression for endosperm traits. Proceedings of the Second International Conference on Quantitative Genetics Sinaur Associates, MA, pp 478–487.

Pooni HS, Kumar I, Khush GS (1992). A comprehensive model for disomically inherited metrical traits expressed in triploid tissues. *Heredity* **69**: 166–174.

Sadimantara GR, Abe T, Sasahara T (1997). Genetic analysis of high molecular weight proteins in rice (*Oryza sativa* L.) endosperm. *Crop Sci* **37**: 1177–1180.

Tan YF, Li JX, Yu SB, Xing YZ, Xu CG, Zhang Q (1999). The three important traits for cooking and eating quality of rice grains are controlled by a single locus in an elite rice hybrid, Shanyou 63. *Theor Appl Genet* **99**: 642–648.

Wang XL, Larkins BA (2001). Genetic analysis of amino acid accumulation in opaque-2 maize endosperm. *Plant Physiol* **125**: 1766–1777.

Wang XL, Woo YM, Kim CS, Larkins BA (2001). Quantitative trait locus mapping of loci influencing elongation factor 1 alpha content in maize endosperm. *Plant Physiol* **125**: 1271–1282.

Wu HP, Chen YS, Chao YT (1998). Studies on the genetic model of cytoplasmic and endospermic effect on quantitative characters of plant. *Chin Agron J* **8**: 7–16.

Xu S (1998a). Iteratively reweighted least squares mapping of quantitative trait loci. *Behav Genet* **28**: 341–355.

Xu S (1998b). Further investigation on regression method of mapping quantitative trait loci. *Heredity* **80**: 364–373.

Zeng ZB (1994). Precision mapping of quantitative trait loci. *Genetics* **136**: 1457–1468.

Zhu J, Weir BS (1994). Analysis of cytoplasmic and maternal effects. 2. Genetic models for triploid endosperms. *Theor Appl Genet* **89**: 160–166.