Statistical Methods for Dissecting Triploid Endosperm Traits Using Molecular Markers: An Autogamous Model

Rongling Wu,*[,1] Chang-Xing Ma,* Maria Gallo-Meagher,[†] Ramon C. Littell* and George Casella*

**Department of Statistics and [†]Agronomy Department, University of Florida, Gainesville, Florida 32611*

Manuscript received April 12, 2002
Accepted for publication June 27, 2002

ABSTRACT

The endosperm, a result of double fertilization in flowering plants, is a triploid tissue whose genetic composition is more complex than diploid tissue. We present a new maximum-likelihood-based statistical method for mapping quantitative trait loci (QTL) underlying endosperm traits in an autogamous plant. Genetic mapping of quantitative endosperm traits is qualitatively different from traits for other plant organs because the endosperm displays complicated trisomic inheritance and represents a younger generation than its mother plant. Our endosperm mapping method is based on two different experimental designs: (1) a one-stage design in which marker information is derived from the maternal genome and (2) a two-stage hierarchical design in which marker information is derived from both the maternal and offspring genomes (embryos). Under the one-stage design, the position and additive effect of a putative QTL can be well estimated, but the estimates of the dominant and epistatic effects are upward biased and imprecise. The two-stage hierarchical design, which extracts more genetic information from the material, typically improves the accuracy and precision of the dominant and epistatic effects for an endosperm trait. We discuss the effects on the estimation of QTL parameters of different sampling strategies under the two-stage hierarchical design. Our method will be broadly useful in mapping endosperm traits for many agriculturally important crop plants and also make it possible to study the genetic significance of double fertilization in the evolution of higher plants.

ONE of the most important events in the evolution of higher plants is the occurrence of double fertilization, a phenomenon independently discovered by Navashin of Russia in 1898 and Guignard of France in 1899 (Friedman 1990, 1998; Jensen 1998). During the process of double fertilization, one of the two sperm cells from a pollen tube fertilizes the haploid egg cell to form a diploid zygote (the new sporophytic generation) and the other sperm cell fertilizes the diploid central cell and fuses with the central cell (polar) nuclei, thus giving rise to the *triploid endosperm*. Different targets for fertilization by the sperm cells are regarded as evolutionarily significant, because such reproductive behavior provides an opportunity for differential segregation of those traits associated with the production of nutritive tissue (endosperm) for the seed from those traits associated with successful embryo development. With the initiation of zygote and endosperm from separate fertilizations, the opportunity exists for optimal, independent specialization of each tissue (Friedman 1990, 1998).

The endosperm is also tremendously important to human nutrition. Grain quality depends critically upon amylose content in rice, protein content and percentage of amino acid in wheat, gum content in barley, and sugar content in sweet corn. Genetic improvement of such endosperm traits that affect food quality has received considerable attention in plant breeding (Benner *et al.* 1989; Sadimantara *et al.* 1997; Mazur *et al.* 1999; van der Meer *et al.* 2001). Quantitative genetic models for analyzing the trisomic inheritance of endosperm traits have been developed and applied to practical data analyses in a variety of grain crops (Gale 1976; Mo 1987; Bogyo *et al.* 1988; Pooni *et al.* 1992; Zhu and Weir 1994). However, traditional quantitative methods may not efficiently resolve the detection of genetic factors underlying endosperm traits because they fail to estimate the map positions of these genetic factors on the chromosomes. DNA-based molecular markers for mapping genetic factors conditioning a quantitative trait, or quantitative trait loci (QTL), are being widely used to map QTL for endosperm traits (Tan *et al.* 1999; Wang and Larkins 2001; Wang *et al.* 2001). For example, using 83 simple sequence repeat loci, Wang *et al.* (2001) detected two QTL, one on the short arm of chromosome 4 and the other on the long arm of chromosome 7, together accounting for 25% of the variance for elongation factor 1α content in maize endosperm.

Current statistical analyses of the association between markers and endosperm traits are based on the assumption that endosperm traits are controlled under the same genetic composition as other tissues developing

[1]*Corresponding author:* Department of Statistics, 533 McCarty Hall C, University of Florida, Gainesville, FL 32611.
E-mail: rwu@stat.ufl.edu

from the diploid embryo. This assumption is obviously violated because the endosperm has three unique properties. First, the endosperm is triploid and has a more complex genetic composition than the embryo. For a locus with two alleles *A* and *a*, four genotypic combinations *AAA*, *AAa*, *Aaa*, and *aaa* are possible, *vs.* three genotypes *AA*, *Aa*, and *aa* for the diploid embryo. Second, the endosperm of a cross between two different genotypes will differ from that of the reciprocal cross. Third, the occurrence of the fertilized egg is the beginning of a new generation, so the embryo and endosperm on a plant represent the next generation. A precise resolution into mapping endosperm traits using marker information from diploid organs needs the development of a bridge between these two tissues with different levels of ploidy. Although this is statistically challenging, currently developed computational algorithms, such as the EM algorithm, provide a powerful means for mapping the QTL contained in the triploid endosperm on the basis of marker information from diploids.

In this article, we have developed a maximum-likelihood-based method, implemented with the EM algorithm, to map QTL responsible for endosperm traits using a genetic linkage map of polymorphic markers. For a particular plant, the formation of endosperm QTL genotypes depends on how its polar nuclei, which are the duplication of the female gametes (eggs), are fertilized by the male gametes (pollen). Our models are different for *autogamous* and *allogamous* plants. For autogamous plants, such as rice, the sperm cells fertilizing the two central cells, which generate the endosperm, are derived from the same plant. For allogamous plants the sperm cells are derived from a pollen pool of the mapping population. For this reason, the endosperm QTL genotypes will have different segregation patterns between autogamous and allogamous plants. Here we report the development of a statistical method for endosperm mapping in autogamous plants. This method is studied under different statistical strategies by extensive simulation experiments.

EXPERIMENTAL DESIGNS

Consider an endosperm trait measured in a backcross or F_2 population. Traditional diploid trait mapping, as proposed by Lander and Botstein (1989), uses marker information and phenotypic measurements derived from the same generation and the same ploidy level. However, because the endosperm is triploid, its precise molecular characterization can be difficult. For example, it is impossible to distinguish directly between two endosperm genotypes *AAa* and *Aaa* from commonly used dominant (randomly amplified polymorphic DNA or amplified fragment length polymorphism) or codominant marker systems (restriction fragment length polymorphism or microsatellite). For dominant markers, three genotypes *AAA*, *AAa*, and *Aaa* cannot be distinguished from one another. For this reason, marker information in endosperm mapping should not be derived from the triploid endosperm rather than from other diploid tissues. Assuming that an endosperm-specific trait is controlled only by the endosperm QTL genotype and that no gene interactions from maternal effects exert to affect the trait expression, we need to predict endosperm QTL behavior (generation $t + 1$) using molecular markers from the maternal genome (generation t) or offspring genome (generation $t + 1$). It should be pointed out that, although generation $t + 1$ is used here for the endosperm, it does not mean that the endosperm can reproduce to generate the progeny. Marker information can be sampled from a diploid tissue of a maternal plant and/or its progeny's diploid tissue (*e.g.*, embryo), which represents two different experimental designs for mapping endosperm QTL contained in seeds. Below, these two designs are described.

Consider a backcross plant of an autogamous species. The diploid marker genotype of this plant is determined only by the gamete genotypes of the heterozygous F_1, whereas its endosperm QTL genotypes are determined by the combination of the polar nucleus of two central cells and a sperm nuclei, which this plant produces for the next generation. Within individual plants, the frequencies of polar nuclei genotypes are identical to those of the female gamete (egg) genotypes because the two central cells are formed from the egg cell through mitosis and, thereby, the polar nuclei genotypes can be regarded as the homogeneous duplication of the egg genotypes (Friedman 1990, 1998). It is clear that, for autogamous plants as considered in this study, the frequencies of the sperm genotypes are identical to those of the egg genotypes and of the polar nuclei genotypes. On the basis of these properties, we can calculate the frequencies of the endosperm QTL genotypes that a backcross plant produces, given the diploid marker genotypes of this plant and its embryos.

Two different experimental designs can be used to predict triploid endosperm QTL genotypes on the basis of diploid marker genotypes. The first design is one in which the endosperm is predicted from a diploid tissue of backcross plants (in generation t). We call this design a *one-stage design*. The second design uses marker genotypes from both backcross plants (maternal genome) and their embryos (offspring genome) to predict the endosperm, which is called a *two-stage hierarchical design* because two successive generations (t and $t + 1$) are genotyped. The one-stage design is simpler in terms of the material genotyped, whereas the two-stage hierarchical design is more precise because within-family variation of a backcross plant is considered.

Suppose there are two flanking markers, M_η and $M_{\eta+1}$, derived from the diploid tissue and embryos of a backcross plant. The recombination fraction between the two markers is denoted by r. A putative QTL, located between the two markers (measured by the recombina-

TABLE 1

Conditional probabilities of endosperm QTL genotypes, conditional upon diploid marker genotypes for $\mathcal{M}_\eta$ and $\mathcal{M}_{\eta+1}$ in a backcross design

Backcross (generation t)	Marker genotype:	Endosperm (generation $t+1$) *QQQ*	*QQq*	*Qqq*	*qqq*
$Z_{11}^{(t)}$		$\frac{(1-r_1)(1-r_2)}{4(1-r)}$	$\frac{(1-r_1)(1-r_2)}{4(1-r)}$	$\frac{(1-r_1)(1-r_2)}{4(1-r)}$	$\frac{1-r_1-r_2+5r_1r_2}{4(1-r)}$
$Z_{10}^{(t)}$		$\frac{(1-r_1)r_2}{4r}$	$\frac{(1-r_1)r_2}{4r}$	$\frac{(1-r_1)r_2}{4r}$	$\frac{4r_1+r_2-5r_1r_2}{4r}$
$Z_{01}^{(t)}$		$\frac{r_1(1-r_2)}{4r}$	$\frac{r_1(1-r_2)}{4r}$	$\frac{r_1(1-r_2)}{4r}$	$\frac{r_1+4r_2-5r_1r_2}{4r}$
$Z_{00}^{(t)}$		$\frac{r_1r_2}{4(1-r)}$	$\frac{r_1r_2}{4(1-r)}$	$\frac{r_1r_2}{4(1-r)}$	$\frac{4-4r_1-4r_2+5r_1r_2}{4(1-r)}$

Marker genotypes of the backcross $Z_{11}^{(t)} = M_\eta m_\eta M_{\eta+1} m_{\eta+1}$, $Z_{10}^{(t)} = M_\eta m_\eta m_{\eta+1} m_{\eta+1}$, $Z_{01}^{(t)} = m_\eta m_\eta M_{\eta+1} m_{\eta+1}$, $Z_{00}^{(t)} = m_\eta m_\eta m_{\eta+1} m_{\eta+1}$. r_1, r_2 and r are the recombination fractions between marker $\mathcal{M}_\eta$ and the QTL, between the QTL and marker $\mathcal{M}_{\eta+1}$, and between the two flanking markers.

tion fraction r_1 with $\mathcal{M}_\eta$ and r_2 with $\mathcal{M}_{\eta+1}$), is segregating in a trisomic manner and exerts an effect on an endosperm trait. Let $G_i^{(t)}$ denote the diploid marker genotype of backcross plant i (generation t) at the two flanking markers, $G_{ij}^{(t+1)}$ denote the embryo marker genotype of the jth seed (generation $t+1$) that this plant produces, and g_k denote the kth ($k = 0, 1, 2, 3$ denotes different numbers of the increasing allele Q) QTL genotype of the triploid endosperm. A general expression for the conditional probability of an endosperm QTL genotype under the one-stage and two-stage hierarchical design can be written, respectively, as

$$p_{ik} = \text{Prob}(g_k | G_i^{(t)}, \mathcal{M}_\eta - \mathcal{M}_{\eta+1}, r_1, r_2)$$
$$p_{ijk} = \text{Prob}(g_k | G_i^{(t)}, G_{ij}^{(t+1)}, \mathcal{M}_\eta - \mathcal{M}_{\eta+1}, r_1, r_2),$$
$$i = 1, \ldots, M; j = 1, \ldots, N_i; k = 0, \ldots, 3;$$

where M is the number of the backcross plants and N_i is the number of seeds derived from backcross plant i, which provide both embryo genotypic and endosperm phenotypic information. Expressions of p_{ik} or p_{ijk} are derived for all possible marker genotypes of the backcross plant (one-stage design) or for all possible marker genotypes of the backcross plant and all its possible embryo marker genotypes (two-stage hierarchical design). These expressions are given in Tables 1 and 2, respectively, with detailed derivations described in APPENDIX A.

STATISTICAL MODEL

We have formulated a statistical mixture model, in which different QTL genotypes in the endosperm are viewed as components of a normal mixture. This mixture model is defined by the frequency of each of the endosperm QTL genotypes and the density corresponding to each genotype.

Additive-dominant model: The phenotypic value (y) of an endosperm trait due to a single QTL for the ith backcross plant under the one-stage design or the jth autogamous seed of the ith backcross plant under the two-stage hierarchical design can be statistically modeled by

$$y_i = \mu + x_i a + z_i^1 d_1 + z_i^2 d_2 + \varepsilon_i, \quad \text{for the one-stage design}$$
$$y_{ij} = \mu + x_{ij} a + z_{ij}^1 d_1 + z_{ij}^2 d_2 + \varepsilon_{ij}, \quad \text{for the two-stage hierarchical design,} \qquad (1)$$

where μ is the overall mean; a is the additive effect of the increasing allele Q at the QTL; d_1 is the first dominant effect, *i.e.*, the dominance effect of *QQ* to *q* when Q is dominant or that of *q* to *QQ* when *q* is dominant; d_2 is the second dominance effect, which reflects the dominance effect of Q to *qq* when Q is dominant or *qq* to Q when *q* is dominant; and x's and z's are the indicator variables describing an endosperm QTL genotype, defined as

Endosperm genotype	x_i, x_{ij}	z_i^1, z_{ij}^1	z_i^2, z_{ij}^2
QQQ	$\frac{3}{2}$	0	0
QQq	$\frac{1}{2}$	1	0
Qqq	$-\frac{1}{2}$	0	1
qqq	$-\frac{3}{2}$	0	0

(2)

TABLE 2

Conditional probabilities (p_{ijk}) of endosperm QTL genotypes, conditional upon two-generation diploid marker genotypes from backcross plants and their embryos

Marker genotype		QTL genotype			
Backcross	Embryo	*QQQ*	*QQq*	*Qqq*	*qqq*
$Z_{11}^{(t)}$	$Z_{22}^{(t+1)}$	$\frac{(1-r_1)^3(1-r_2)^3}{(1-r)^3}$	$\frac{r_1r_2(1-r_1)^2(1-r_2)^2}{(1-r)^3}$	$\frac{r_1r_2(1-r_1)^2(1-r_2)^2}{(1-r)^3}$	$\frac{r_1r_2[r_1r_2(1-r_1)(1-r_2)+(1-r)^2]}{(1-r)^3}$
	$Z_{21}^{(t+1)}$	$\frac{r_2(1-r_1)^3(1-r_2)^2}{r(1-r)^2}$	$\frac{r_1(1-r_1)^2(1-r_2)(1-2r_2+2r_2^2)}{2r(1-r)^2}$	$\frac{r_1(1-r_1)^2(1-r_2)(1-2r_2+2r_2^2)}{2r(1-r)^2}$	$\frac{r_1r_2[r_1(1-r_1)(1-r_2)^2+r(1-r)]}{r(1-r)^2}$
	$Z_{20}^{(t+1)}$	$\frac{r_2^2(1-r_1)^3(1-r_2)}{r^2(1-r)}$	$\frac{r_1r_2(1-r_1)^2(1-r_2)^2}{r^2(1-r)}$	$\frac{r_1r_2(1-r_1)^2(1-r_2)^2}{r^2(1-r)}$	$\frac{r_1[r_1(1-r_1)(1-r_2)^3+r_2r^2]}{r^2(1-r)}$
	$Z_{12}^{(t+1)}$	$\frac{r_1(1-r_1)^2(1-r_2)^3}{r(1-r)^2}$	$\frac{r_2(1-r_1)(1-r_2)^2(1-2r_1+2r_1^2)}{2r(1-r)^2}$	$\frac{r_2(1-r_1)(1-r_2)^2(1-2r_1+2r_1^2)}{2r(1-r)^2}$	$\frac{r_1r_2[r_2(1-r_1)^2(1-r_2)+r(1-r)]}{r(1-r)^2}$
	$Z_{11}^{(t+1)}$	$\frac{2r_1r_2(1-r_1)^2(1-r_2)^2}{(1-r)(1-2r+2r^2)}$	$\frac{(1-r_1)(1-r_2)(1-2r_1+2r_1^2)(1-2r_2+2r_2^2)}{2(1-r)(1-2r+2r^2)}$	$\frac{(1-r_1)(1-r_2)(1-2r_1+2r_1^2)(1-2r_2+2r_2^2)}{2(1-r)(1-2r+2r^2)}$	$\frac{r_1r_2[2(1-r_1)^2(1-r_2)^2+(1-2r+2r^2)]}{(1-r)(1-2r+2r^2)}$
	$Z_{10}^{(t+1)}$	$\frac{r_1r_2^2(1-r_1)^2(1-r_2)}{r(1-r)^2}$	$\frac{r_2(1-r_1)(1-r_2)^2(1-2r_1+2r_1^2)}{2r(1-r)^2}$	$\frac{r_2(1-r_1)(1-r_2)^2(1-2r_1+2r_1^2)}{2r(1-r)^2}$	$\frac{r_1[(1-r_1)^2(1-r_2)^3+r_2r(1-r)]}{r(1-r)^2}$
	$Z_{02}^{(t+1)}$	$\frac{r_1^2(1-r_1)(1-r_2)^3}{r^2(1-r)}$	$\frac{r_1(1-r_1)^2r_2(1-r_2)^2}{r^2(1-r)}$	$\frac{r_1(1-r_1)^2r_2(1-r_2)^2}{r^2(1-r)}$	$\frac{r_2[r_2(1-r_1)^3(1-r_2)+r_1r^2]}{r^2(1-r)}$
	$Z_{01}^{(t+1)}$	$\frac{r_1^2r_2(1-r_1)(1-r_2)^2}{r(1-r)^2}$	$\frac{r_1(1-r_1)^2(1-r_2)(1-2r_2+2r_2^2)}{2r(1-r)^2}$	$\frac{r_1(1-r_1)^2(1-r_2)(1-2r_2+2r_2^2)}{2r(1-r)^2}$	$\frac{r_2[(1-r_1)^3(1-r_2)^2+r_1r(1-r)]}{r(1-r)^2}$
	$Z_{00}^{(t+1)}$	$\frac{r_1^2r_2^2(1-r_1)(1-r_2)}{(1-r)^3}$	$\frac{r_1r_2(1-r_1)^2(1-r_2)^2}{(1-r)^3}$	$\frac{r_1r_2(1-r_1)^2(1-r_2)^2}{(1-r)^3}$	$\frac{(1-r_1)^3(1-r_2)^3+r_1r_2(1-r)^2}{(1-r)^3}$
$Z_{10}^{(t)}$	$Z_{20}^{(t+1)}$	$\frac{r_2(1-r_1)^3}{r}$	$\frac{r_1r_2(1-r_1)^2}{r}$	$\frac{r_1r_2(1-r_1)^2}{r}$	$\frac{r_1[r_1r_2(1-r_1)+(1-r_2)]}{r}$
	$Z_{10}^{(t+1)}$	$\frac{r_1r_2(1-r_1)^2}{r}$	$\frac{r_2(1-r_1)(1-2r_1+2r_1^2)}{2r}$	$\frac{r_2(1-r_1)(1-2r_1+2r_1^2)}{2r}$	$\frac{r_1[r_2(1-r_1)^2+(1-r_2)]}{r}$
	$Z_{00}^{(t+1)}$	$\frac{r_1^2r_2(1-r_1)}{r}$	$\frac{r_1r_2(1-r_1)^2}{r}$	$\frac{r_1r_2(1-r_1)^2}{r}$	$\frac{r_2(1-r_1)^3+r_1(1-r_2)}{r}$
$Z_{01}^{(t)}$	$Z_{02}^{(t+1)}$	$\frac{r_1(1-r_2)^3}{r}$	$\frac{r_1r_2(1-r_2)^2}{r}$	$\frac{r_1r_2(1-r_2)^2}{r}$	$\frac{r_2[r_1r_2(1-r_2)+(1-r_1)]}{r}$
	$Z_{01}^{(t+1)}$	$\frac{r_1r_2(1-r_2)^2}{r}$	$\frac{r_1(1-r_2)(1-2r_2+2r_2^2)}{2r}$	$\frac{r_1(1-r_2)(1-2r_2+2r_2^2)}{2r}$	$\frac{r_2[r_1(1-r_2)^2+(1-r_1)]}{r}$
	$Z_{00}^{(t+1)}$	$\frac{r_1r_2^2(1-r_2)}{r}$	$\frac{r_1r_2(1-r_2)^2}{r}$	$\frac{r_1r_2(1-r_2)^2}{r}$	$\frac{r_1(1-r_2)^3+r_2(1-r_1)}{r}$
$Z_{00}^{(t)}$	$Z_{00}^{(t+1)}$	$\frac{r_1r_2}{4(1-r)}$	$\frac{r_1r_2}{4(1-r)}$	$\frac{r_1r_2}{4(1-r)}$	$\frac{4-4r_1-4r_2+5r_1r_2}{4(1-r)}$

See Table 1 for explanations of the symbols.

The variables ε_i, ε_{ij} are the residuals including the aggregate effect of both polygenes and error effect and distributed as $N(0, \sigma^2_\varepsilon)$. The probabilities with which x's and z's take an assigned value depend on the genomic positions of the QTL in the interval bracketed by flanking markers given in Tables 1 and 2.

Consider an endosperm mapping population composed of M backcross plants and N_i randomly selected autogamous seeds from the ith backcross plant. These backcross plants and the embryos of their seeds are genotyped simultaneously, whereas phenotypes are measured for the endosperm of their seeds. The likelihood of the marker data and the endosperm trait values controlled by the putative QTL can be represented under the one-stage or two-stage hierarchical design by

$$\ell(\mathbf{\Omega}) = \begin{cases} \prod_{i=1}^{M}[\sum_{k=0}^{3} p_{ik} f_k(y_i)], \\ \quad \text{for the one-stage design} \\ \prod_{i=1}^{M}\prod_{j=1}^{N_i}[\sum_{k=0}^{3} p_{ijk} f_k(y_{ij})], \\ \quad \text{for the two-stage hierarchical design,} \end{cases} \tag{3}$$

where $\mathbf{\Omega} = (\mu, a, d_1, d_2, r_1 \text{ or } r_2, \sigma^2_e)^T$ is the vector for unknown QTL-effect parameters, QTL-position parameters, and residual variance to be estimated; p_{ik}, p_{ijk} are the proportions of each mixture normal (*i.e.*, endosperm QTL genotype); and f_k is the normal density corresponding to the kth genotype, with mean μ_k described in Equations 2 and 3. On the basis of quantitative genetic models of endosperm traits (Gale 1976; Mo 1987; Bogyo *et al.* 1988; Pooni *et al.* 1992), unknown QTL-effect parameters $\mathbf{e} = (\mu, a, d_1, d_2)^T$ can be exactly expressed as a linear combination of QTL genotypic means $\mathbf{m} = (\mu_3, \mu_2, \mu_1, \mu_0)^T$. We first estimate $\mathbf{m}$ and then solve for $\mathbf{e}$ because the former has a simpler structure.

The maximum-likelihood estimates (MLEs) of the unknown parameters $\mathbf{\Omega} = (\mathbf{m}, r_1 \text{ or } r_2, \sigma^2_e)^T$ under the one-stage or two-stage hierarchical design can be computed by implementing an expectation maximization (EM) algorithm (Dempster *et al.* 1977; Meng and Rubin 1993). The log-likelihood of Equation 3 for the two-stage hierarchical design is given by

$$\log \ell(\mathbf{\Omega}) = \sum_{i=1}^{M}\sum_{j=1}^{N_i} \log\left[\sum_{k=0}^{3} p_{ijk} f_k(y_{ij})\right] \tag{4}$$

with derivatives

$$\begin{aligned} \frac{\partial}{\partial \mathbf{\Omega}_\phi}\log \ell(\mathbf{\Omega}) &= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=0}^{3} \frac{p_{ijk}\frac{\partial}{\partial \mathbf{\Omega}_\phi} f_k(y_{ij})}{\sum_{k=0}^{3} p_{ijk} f_k(y_{ij})} \\ &= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=0}^{3} \frac{p_{ijk} f_k(y_{ij})}{\sum_{k=0}^{3} p_{ijk} f_k(y_{ij})} \frac{\partial}{\partial \mathbf{\Omega}_\phi}\log f_k(y_{ij}) \\ &= \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=0}^{3} P_{ijk}\frac{\partial}{\partial \mathbf{\Omega}_\phi}\log f_k(y_{ij}), \end{aligned}$$

where we define

$$P_{ijk} = \frac{p_{ijk} f_k(y_{ij})}{\sum_{k=0}^{3} p_{ijk} f_k(y_{ij})}, \tag{5}$$

which could be thought of as a posterior probability that the endosperm from the jth seed of the ith backcross plant has a QTL genotype k. We then implement the EM algorithm with the expanded parameter set $\{\mathbf{\Omega}, \mathbf{P}\}$, where $\mathbf{P} = \{P_{ijk}\}$. Conditional on $\mathbf{P}$, we solve for the zeros of $(\partial/\partial \mathbf{\Omega}_\phi)\log \ell(\mathbf{\Omega})$ (appendix b) to get our estimates of $\mathbf{\Omega}$ (the M step). The estimates are then used to update $\mathbf{P}$ (the E step), and the process is repeated until convergence. The values at convergence are the MLEs. The formulas for estimating the unknown parameters in the M step under both one-stage and two-stage hierarchical designs are given in appendix b.

Epistatic model: Suppose there are two biallelic QTL $\mathcal{Q}_1$ of alleles Q_1 and q_1 and $\mathcal{Q}_2$ of alleles Q_2 and q_2 to epistatically affect an endosperm-specific trait. At each QTL there are four possible genotypes for triploid endosperm. Thus, two-QTL genotypes in the endosperm can be arrayed by

$$\begin{bmatrix} Q_1Q_1Q_1 \\ Q_1Q_1q_1 \\ Q_1q_1q_1 \\ q_1q_1q_1 \end{bmatrix} \otimes \begin{bmatrix} Q_2Q_2Q_2 \\ Q_2Q_2q_2 \\ Q_2q_2q_2 \\ q_2q_2q_2 \end{bmatrix},$$

where $\otimes$ is the Kronecker product.

If the two putative QTL are tested on different intervals, the conditional probabilities of the two-QTL genotypes can be calculated independently for each QTL, *i.e.*, $p_{ik_1k_2} = p_{ik_1}p_{ik_2}$ ($i = 1, \ldots, M$; $k_1, k_2 = 0, \ldots, 3$) for the one-stage design and $p_{ijk_1k_2} = p_{ijk_1}p_{ijk_2}$ ($j = 1, \ldots, N_i$) for the two-stage hierarchical design, assuming that there is no crossover interference between the two marker intervals. Denote $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ as the matrices for the conditional probability of QTL genotypes for $\mathcal{Q}_1$ conditional upon the marker interval $\mathcal{M}_\eta$–$\mathcal{M}_{\eta+1}$ and for $\mathcal{Q}_2$ conditional upon a different marker interval $\mathcal{M}_{\eta'}$–$\mathcal{M}_{\eta'+1}$, respectively. The conditional probability matrix ($\mathbf{\Theta}$) of joint QTL genotypes at $\mathbf{Q}_1$ and $\mathbf{Q}_2$ conditional upon the two-marker intervals can be expressed as $\mathbf{\Theta} = \mathbf{\Theta}_1 \otimes \mathbf{\Theta}_2$. If two linked QTL are located within the same marker interval, the conditional probabilities of the two-QTL endosperm genotypes conditional upon the diploid marker genotypes of the flanking markers (bracketing two putative QTL) should be derived. These two-QTL conditional probabilities for the backcross population of an autogamous plant are given on our statistical genetics webpage (http://www.ifasstat.ufl.edu/genetics/~endosperm). In a practical data analysis, however, modeling two QTL within the same marker interval should be carefully considered because of a confounding effect, unless an adequately large sample size is used.

The phenotypic value of an endosperm trait due to the two putative QTL can be modeled under the one-stage or two-stage hierarchical design. As our analysis is similar between these two models, only the two-stage hierarchical design is described. The phenotypic value of an endosperm trait from the ith backcross plant and its jth autogamous seed under the two-hierarchical design is written as

$$\begin{aligned}
y_{ij} = \mu &+ x_{ij1}a_1 + x_{ij2}a_2 && \text{Additive effects}\\
&+ z^1_{ij1}d^1_1 + z^2_{ij1}d^2_1 + z^1_{ij2}d^1_2 + z^2_{ij2}d^2_2 && \text{Dominant effects}\\
&+ (x_{ij1}x_{ij2})i_{aa} && \text{Additive} \times \text{additive effect}\\
&+ (x_{ij1}z^1_{ij2})j^1_{ad} + (x_{ij1}z^2_{ij2})j^2_{ad} && \text{Additive} \times \text{dominance effects}\\
&+ (z^1_{ij1}x_{ij2})k^1_{da} + (z^2_{ij1}x_{ij2})k^2_{da} && \text{Dominant} \times \text{additive effects}\\
&+ (z^1_{ij1}z^1_{ij2})l^{11}_{dd} && \text{First dominant} \times \text{first dominant effect}\\
&+ (z^2_{ij1}z^2_{ij2})l^{22}_{dd} && \text{Second dominant} \times \text{second dominant effect}\\
&+ (z^1_{ij1}z^2_{ij2})l^{12}_{dd} + (z^2_{ij1}z^1_{ij2})l^{21}_{dd} && \text{First dominant} \times \text{second dominant or second dominant} \times \text{additive effects}\\
&+ \varepsilon_{ij} && \text{Error effects,}
\end{aligned} \tag{6}$$

where x's and z's are the indicator variables describing an endosperm QTL genotype at each QTL for endosperm ij, as defined in expression (2), and a's, d's, i_{aa}, j_{ad}'s, k_{da}'s, and l_{dd}'s are the corresponding additive, dominant, additive × additive, additive × dominant, dominant × additive, and dominant × dominant epistatic effects between the two QTL (see Equation 6). The EM algorithm is implemented to estimate all unknown parameters including the overall mean, QTL effects and position, and residual variance (APPENDIX B).

HYPOTHESIS TESTING

Additive-dominant model: A number of hypothesis tests can be formulated for endosperm inheritance. The first hypothesis considers the existence of any QTL affecting the expression of an endosperm trait, which is expressed as

$$\begin{aligned}
&H_0\text{: } a = d_1 = d_2 = 0\\
&H_1\text{: at least one of the equalities above does not hold,}
\end{aligned} \tag{7}$$

for the backcross B_1. The test statistic for testing the above hypotheses is calculated as the log-likelihood ratio of the full model (H_1) over the reduced model (H_0),

$$\text{LR} = -2\log\left[\frac{L(\tilde{\boldsymbol{\Omega}})}{L(\hat{\boldsymbol{\Omega}})}\right],$$

where $\tilde{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\Omega}}$ denote the ML estimates of the unknown parameters under H_0 and H_1, respectively. The log-likelihood ratio (LR) is asymptotically χ^2 distributed with 3 d.f. However, the critical threshold value for declaring the existence of an endosperm QTL is generally calculated on the basis of permutation tests (CHURCHILL and DOERGE 1994).

Second, a hypothesis test can be made for the additive or dominant effects of the QTL on the endosperm trait, on the basis of

$$\begin{aligned}
&H_0\text{: } a = 0\\
&H_1\text{: } a \neq 0,
\end{aligned} \tag{8}$$

whose log-likelihood ratio test statistic is asymptotically χ^2 distributed with 1 d.f., and

$$\begin{aligned}
&H_0\text{: } d_1 = d_2 = 0\\
&H_1\text{: at least one of the equalities above does not hold}
\end{aligned} \tag{9}$$

whose log-likelihood ratio test statistic is asymptotically χ^2 distributed with 2 d.f.

Epistatic model: The existence of any QTL affecting the expression of an endosperm trait under the two-QTL model can be tested on the basis of the hypotheses

$$\begin{aligned}
H_0\text{: } &a_1 = a_2 = d^1_1 = d^2_1 = d^1_2 = d^2_2 = i_{aa} = j^1_{ad} = j^2_{ad}\\
&= k^1_{da} = k^2_{da} = l^{11}_{dd} = l^{12}_{dd} = l^{21}_{dd} = l^{22}_{dd} = 0\\
H_1\text{: at least one of the equalities above does not hold.}
\end{aligned} \tag{10}$$

The test statistic for testing the above hypotheses is calculated as the log-likelihood ratio of the full model (H_1) over the reduced model (H_0), which is asymptotically χ^2 distributed with 15 d.f.

Like the hypothesis tests for the additive or dominant effects of individual QTL, the significance of QTL epistasis effects on the expression of an endosperm trait can be tested on the basis of

$$\begin{aligned}
H_0\text{: } &i_{aa} = j^1_{ad} = j^2_{ad} = k^1_{da} = k^2_{da} = l^{11}_{dd} = l^{12}_{dd} = l^{21}_{dd}\\
&= l^{22}_{dd} = 0\\
H_1\text{: at least one of the equalities above does not hold.}
\end{aligned} \tag{11}$$

Under the null hypothesis, the LR of Equation 11 is asymptotically χ^2 distributed with 9 d.f. Specifically, different components of epistatic (additive × additive, additive × dominant, dominant × additive, and dominant × dominant) effects can also be tested.

If two QTL detected are on the same interval, the degree of their linkage can also be tested. Testing the QTL linkage is equivalent to testing $r_2 = 0$, where r_2 is the recombination fraction between the two QTL. This test can be extended to test for a particular value of the recombination fraction.

MONTE CARLO SIMULATION

Simulation scenarios: We performed a series of simulation experiments to examine the statistical properties of our endosperm mapping method. A linkage group length of 100 cM is simulated for a backcross population. Assume that there are six equidistant markers or-

dered as $\mathcal{M}_1$–$\mathcal{M}_6$ on the group. In the backcross, these six markers generate a total of 64 genotypes whose frequencies are simulated on the basis of the recombination fractions between all pairs of two successive markers. We use the Kosambi map function to convert the map distance into the recombination fraction.

In our simulation experiments, we use different sampling strategies for marker information (one-stage *vs.* two-stage hierarchical design), different levels of heritability ($H^2 = 0.2$ *vs.* 0.6) and different sample sizes ($M =$ 200 *vs.* 400). For the two-stage hierarchical design, different sampling strategies are designed on the basis of allocations of a given sample size between the backcross plants and their progeny (seeds). The sampling strategies used here are (1) 400 × 1 (one seed is sampled from each of 400 backcross plants), (2) 40 × 10 (40 seeds are sampled from each of 10 backcross plants), (3) 20 × 20 (20 seeds are sampled from each of 20 backcross plants), and (4) 10 × 40 (10 seeds are sampled from each of 40 backcross plants). In addition to their possible effects on parameter estimation, these different strategies have different utilities in practice. For example, strategy 1 does not require genotyping many embryos, but requires the maintenance of a large backcross population. In strategy 4, only a small backcross population is maintained, but it needs many embryos to be genotyped.

Additive-dominant model: Suppose there is a QTL located on the middle point of the linkage group (*i.e.*, 50 cM away from each end of the group), which affects a quantitative endosperm trait. The additive and dominant effects of the QTL are hypothesized as $a = 1$, $d_1 = 0.8$, and $d_2 = 0.5$. Given the overall mean $\mu = 10$, the genotypic means of the four possible endosperm QTL genotypes are calculated using Equation 1. In the endosperm progeny of the backcross, the frequencies of these triploid QTL genotypes are 1/8 for *QQQ*, 1/8 for *QQq*, 1/8 for *Qqq*, and 5/8 for *qqq*. The genetic variance due to this QTL is 1.2. When the heritability H^2 of an endosperm trait is given, the residual variance can be calculated, from which the residual effects (and therefore phenotypic values) of the individuals are simulated.

The characterization of the threshold for declaring the existence of a QTL is a difficult issue. The permutation test proposed by Churchill and Doerge (1994) is regarded as a useful approach for calculating the threshold because it is not dependent on the distribution of the test statistic. However, permutation tests require expensive computations. We thus use chromosome-wide permutation tests to characterize the thresholds only under a sample size of 400 for the one-stage design and under a 40 × 10 strategy for the two-stage hierarchical design. A set of endosperm phenotypic values for the one-stage design is simulated using the residual variances of $\sigma_\varepsilon^2 = 1.2$ and 6.4, which correspond to the heritability levels of 0.6 and 0.2, respectively, when a QTL is assumed. Similarly, for the two-stage hierarchical design, two simulation scenarios, with $\sigma_\varepsilon^2 = 1.2$ and 6.4, are designed. In each case, our model is used to estimate QTL parameters for endosperm inheritance and calculate the corresponding LR.

The distribution of the LR values over 1000 permutation replicates can be approximated by a χ^2 distribution. The 95th, 99th, and 99.9th percentiles of the distribution of the maximum are used as empirical critical values to declare the existence of a QTL on the linkage groups at the significance levels $\alpha = 0.05$, 0.01, and 0.001. Under the one-stage design, these percentiles are 10.2551, 13.1874, and 15.7615 for $H^2 = 0.6$ and 10.1746, 15.5654, and 23.3611 for $H^2 = 0.2$, respectively. These percentiles under the two-stage hierarchical design (40 × 10) are 12.6697, 15.9342, and 20.3486 for $H^2 =$ 0.6 and 12.9303, 15.8864, and 18.7674 for $H^2 = 0.2$.

Epistatic model: Suppose there are two QTL affecting a quantitative endosperm trait, the first located on the middle point of the marker interval $\mathcal{M}_2$–$\mathcal{M}_3$ and the second on the middle point of the marker interval $\mathcal{M}_4$–$\mathcal{M}_5$. The additive and dominant effects of the first QTL are hypothesized as $a_1 = 1.0$, $d_1^1 = 0.8$, and $d_1^2 = 0.4$ and those of the second QTL as $a_2 = 0.5$, $d_2^1 = 0.5$, and $d_2^2 = 0.5$. The epistatic interaction effects between the two QTL are hypothesized as $i_{aa} = 0.6$ (additive × additive), $j_{ad}^1 = j_{ad}^2 = -0.4$ (additive × dominant), $k_{da}^1 = k_{da}^2 = -0.2$ (dominant × additive), and $l_{dd}^{11} = l_{dd}^{12} = l_{dd}^{21} = l_{dd}^{22} = 0.2$ (dominant × dominant). Given the overall mean $\mu = 10$, the genotypic means of the 16 possible endosperm QTL genotypes at the two QTL are calculated using Equation 6. In the endosperm progeny of the backcross, the frequencies of the triploid QTL genotypes for each QTL are 1/8 for *QQQ*, 1/8 for *QQq*, 1/8 for *Qqq*, and 5/8 for *qqq*. Thus, when these two QTL are on different marker intervals, the frequencies of their 16 genotypes can be arrayed by $(1/8, 1/8, 1/8, 5/8)^T \otimes (1/8, 1/8, 1/8, 5/8)^T$. When these two QTL are on the same marker interval, the frequencies of the QTL genotypes in the endosperm progeny of the backcross plants can be calculated on the basis of the joint probabilities of the QTL-marker genotypes. With the genotypic means and frequencies of the QTL genotypes, the genetic variances due to these two QTL can be calculated for each model.

We perform LR tests across a grid of locations on the chromosome to infer the most likely genome positions of two QTL. The declaration for the existence of QTL is based on a critical threshold for the LR test statistic that controls the chromosome-wise type I error rate. Permutation tests proposed by Churchill and Doerge (1994) are used to calculate the threshold values for each simulation scenario.

RESULTS

Additive-dominant model: *One-stage design:* The position and effects of the hypothesized QTL can be reason-

TABLE 3

The MLEs of QTL parameters and their MSEs (in parentheses) for different heritabilities (H^2) and sample sizes (M) under the one-stage design based on 100 repeated simulations

H^2	Position (50)	$\mu_3 = 11.5$	$\mu_2 = 11.3$	$\mu_1 = 10$	$\mu_0 = 8.5$	σ^2_ε	$\mu = 100$	$a = 1$	$d_1 = 0.8$	$d_2 = 0.5$
The number of backcross plants $M = 200$										
0.2	49.16	11.1816	11.1232	10.4401	8.5241	5.7311	9.8528	0.8858	0.8275	1.0302
	(97.52)	(1.5357)	(0.9000)	(1.3539)	(0.0636)	(0.7825)	(0.3893)	(0.1824)	(1.7827)	(2.0774)
0.6	50.30	11.2298	11.2106	10.2420	8.4921	0.9652	9.8610	0.9126	0.8933	0.8373
	(9.64)	(0.4184)	(0.2988)	(0.6325)	(0.0127)	(0.0174)	(0.1193)	(0.0428)	(0.6100)	(0.8869)
The number of backcross plants $M = 400$										
0.2	49.44	11.2540	11.1057	10.4118	8.5072	5.8561	9.8806	0.9156	0.7673	0.9890
	(26.16)	(1.0335)	(0.7473)	(1.2068)	(0.0361)	(0.3272)	(0.2733)	(0.1162)	(1.4579)	(1.8538)
0.6	49.80	11.0950	11.3546	10.3275	8.4984	0.9696	9.7967	0.8656	1.1252	0.9636
	(4.56)	(0.4659)	(0.2485)	(0.5524)	(0.0036)	(0.0065)	(0.1209)	(0.0506)	(0.6217)	(0.8394)

The position of the QTL is described by the map distance (in centimorgans) from one end of the linkage group (100 cM long). σ^2_ε hypothesized is 6.4 at $H^2 = 0.2$ and 1.2 at $H^2 = 0.6$.

ably well estimated under this design, but, as expected, with better accuracy and precision for a larger than a smaller sample size and for a higher rather than lower heritability level (Table 3). The mean-squared error (MSE) for the position MLEs among 100 replications of simulation is very high (97.52) at $M = 200$ and $H^2 = 0.2$, suggesting that the QTL position cannot be precisely estimated in this case. But such a high sampling error can be reduced by a factor of 3 when M is increased to 400 or by a factor of 8 when H^2 is increased to 0.6. This information also can be seen in Figure 1, in which a more precise localization of the QTL displays a narrower peak for the profile of the log-likelihood ratio test statistics across the length of the linkage group.

The MLEs of the genotypic means (μ's) can be well estimated, although a larger heritability and larger sample size will lead to better estimates (Table 3). These estimates are used to solve for one additive effect (a) and two dominant effects (d_1 and d_2) using the linear equation given in APPENDIX B. The MLEs of the additive and dominant effects display different accuracy and precision. Whereas the MLE of the additive effect has acceptable precision at $M = 200$ and $H^2 = 0.2$, those of the two dominant effects are highly *upward* biased and imprecise. Moreover, the dominant effect (d_2) due to the intralocus interaction of Q and qq appears to be not only overestimated more seriously, but also has lower estimation precision than the dominant effect (d_1) due to the intralocus interaction of QQ and q. For example, at $M = 200$ and $H^2 = 0.2$, d_2 is overestimated by onefold, whereas d_1 is overestimated by ~30%. When the heritability of an endosperm trait and/or the sample size used is increased, the precision of the dominant effect estimation improves, but to a lesser extent than does the precision of the additive effect estimation.

In general, the one-stage design that ignores marker segregation within backcross families can be used to estimate the position and effect of a QTL on the endosperm. But acceptable estimation precision for dominant effects under the one-stage design requires a high heritability level of an endosperm trait and a large sample size.

Two-stage hierarchical design: The precision of the MLEs for the QTL position and effects improves significantly under the two-stage hierarchical design (Table 4) compared to the one-stage design (Table 3). Under the two-

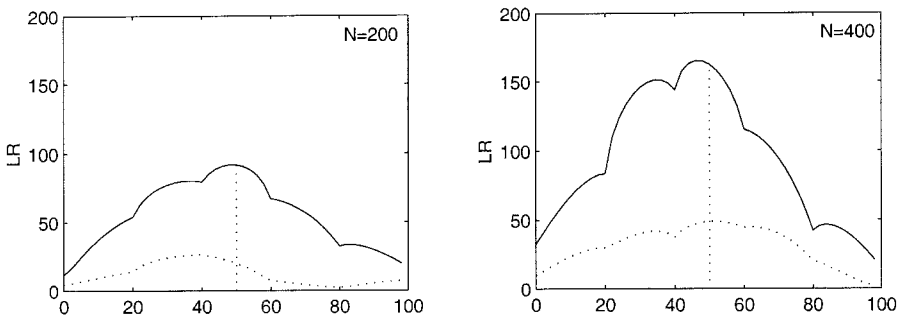

FIGURE 1.—The profiles of the log-likelihood-ratio test statistics calculated as a function of genome position on the simulated linkage group for different sample sizes and heritabilities under the one-stage design. The dotted and solid curves refer to heritabilities of 0.2 and 0.6, respectively. The vertical dotted line refers to the true position of a hypothesized QTL on the linkage group.

TABLE 4

The MLEs of QTL parameters and their MSEs (in parentheses) for different heritabilities (H^2) and different sampling designs under the two-stage hierarchical design based on 100 repeated simulations

H^2	Position (50)	$\mu_3 = 11.5$	$\mu_2 = 11.3$	$\mu_1 = 10$	$\mu_0 = 8.5$	σ^2_ε	$\mu = 100$	$a = 1$	$d_1 = 0.8$	$d_2 = 0.5$
					400×1					
0.2	49.72	11.5102	11.1195	10.1503	8.4921	6.2559	10.0012	1.0060	0.6153	0.6522
	(14.48)	(0.6387)	(0.5824)	(0.4539)	(0.0295)	(0.2818)	(0.1672)	(0.0742)	(0.8813)	(0.6242)
0.6	50.16	11.4318	11.2290	10.1034	8.4945	1.2007	9.9631	0.9791	0.7763	0.6298
	(1.84)	(0.1452)	(0.0953)	(0.0952)	(0.0049)	(0.0153)	(0.0396)	(0.0158)	(0.1736)	(0.1305)
					40×10					
0.2	50.12	11.4795	11.2140	10.1272	8.5144	6.2162	9.9970	0.9884	0.7228	0.6244
	(19.60)	(0.5637)	(0.6333)	(0.5100)	(0.0325)	(0.3135)	(0.1484)	(0.0665)	(0.9750)	(0.6412)
0.6	50.10	11.5062	11.2689	10.0314	8.5118	1.1801	10.0090	0.9981	0.7608	0.5214
	(1.80)	(0.1055)	(0.0753)	(0.0860)	(0.0078)	(0.0123)	(0.0283)	(0.0126)	(0.1600)	(0.0929)
					20×20					
0.2	49.96	11.5187	11.1467	10.1642	8.4966	6.1711	10.0076	1.0074	0.6354	0.6602
	(17.92)	(0.6004)	(0.6230)	(0.5761)	(0.0304)	(0.3425)	(0.1535)	(0.0720)	(0.9028)	(0.8259)
0.6	50.10	11.4748	11.2085	10.0890	8.5063	1.2009	9.9905	0.9895	0.7232	0.5932
	(2.36)	(0.1404)	(0.0922)	(0.1125)	(0.0047)	(0.0103)	(0.0355)	(0.0165)	(0.2254)	(0.1224)
					10×40					
0.2	49.90	11.6398	11.1121	10.1218	8.5009	6.2893	10.0704	1.0463	0.5186	0.5745
	(16.60)	(0.6398)	(0.4459)	(0.5555)	(0.0424)	(0.2999)	(0.1622)	(0.0795)	(0.9146)	(0.6733)
0.6	50.08	11.5145	11.2876	10.0449	8.5013	1.1935	10.0079	1.0044	0.7775	0.5392
	(1.92)	(0.1119)	(0.0688)	(0.0957)	(0.0055)	(0.0128)	(0.0288)	(0.0133)	(0.1531)	(0.1034)

See Table 3 for explanations of the symbols.

stage hierarchical design, the QTL can be localized with high mapping resolution when H^2 is high (0.6; Figure 2). For the two-stage hierarchical design, different sampling strategies have different precision of parameter estimation. The estimation precision is best under strategy 40×10 when H^2 is higher, but the precision is best under strategy 400×1 when H^2 is lower. For the same design, the precision of parameter estimation differs

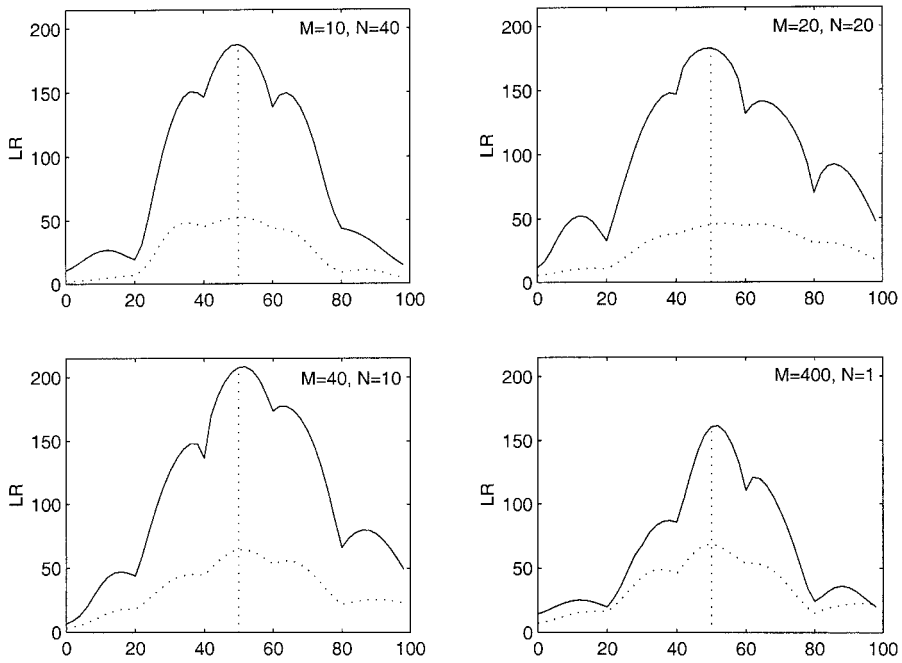

FIGURE 2.—The profiles of the log-likelihood-ratio test statistics calculated as a function of genome position on the simulated linkage group for different sampling designs and heritabilities under the two-stage hierarchical design. The dotted and solid curves refer to heritabilities of 0.2 and 0.6, respectively. The vertical dotted line refers to the true position of a hypothesized QTL on the linkage group.

TABLE 5

The MLEs of QTL parameters and mean square errors (MSE) of the estimates among 100 replicated simulations at different heritability (H^2) levels and sample sizes (M) under the one-stage design

	$H^2 = 0.2$ ($\sigma_\varepsilon^2 = 3.6$)				$H^2 = 0.6$ ($\sigma_\varepsilon^2 = 0.6$)			
	$M = 200$		$M = 400$		$M = 200$		$M = 400$	
True parameter	MLE	MSE	MLE	MSE	MLE	MSE	MLE	MSE
$\tau_1 = 30$	26.00	17.00	30.09	33.00	29.60	33.00	30.89	4.56
$\tau_2 = 70$	61.00	81.00	68.27	46.82	66.20	46.60	71.44	42.33
$\mu = 10.0$	9.79	0.17	9.70	0.25	10.00	0.05	9.94	0.05
$a_1 = 1.0$	0.82	0.03	0.87	0.08	0.89	0.03	0.93	0.03
$a_2 = 0.5$	1.02	0.37	0.42	0.30	0.56	0.05	0.36	0.05
$d_1^1 = 0.8$	1.41	1.12	1.08	1.16	0.66	0.23	0.74	0.45
$d_1^2 = 0.4$	1.46	1.42	1.04	0.87	0.94	0.52	0.65	0.39
$d_2^1 = 0.5$	−0.06	0.32	0.55	0.75	0.30	0.49	0.66	0.15
$d_2^2 = 0.5$	1.49	2.54	0.81	0.65	0.24	0.37	0.71	0.57
$i_{aa} = 0.6$	0.71	0.04	0.63	0.02	0.56	0.02	0.55	0.01
$j_{ad} = -0.4$	−0.01	0.18	−0.27	0.14	−0.33	0.11	−0.29	0.11
$k_{da} = -0.2$	−0.03	0.04	−0.12	0.42	−0.23	0.28	−0.13	0.10
$l_{dd} = 0.2$	−0.48	5.08	−0.39	5.87	−0.21	1.32	−0.31	3.24
$\hat{\sigma}_\varepsilon^2$	2.20	0.63	2.76	0.22	0.57	0.01	0.61	0.02
Percentage[a]	10		55		50		90	

[a] The percentage of the simulations in which a significant QTL at the significance level $\alpha = 0.05$ is detected.

between different types of parameters. A general trend is that the MLEs of the dominant effects have lower precision than those of the additive effect and the QTL position. But compared to the one-stage design, the estimates of the dominant effects are much less biased for the two-stage hierarchical design (Table 4).

In summary, the two-stage hierarchical design is better than the one-stage design in terms of the accuracy and precision of QTL parameters. The two-stage hierarchical design is particularly more advantageous than the one-stage design in estimating the dominant effects of an endosperm trait.

Epistatic model: *One-stage design:* The accuracy and precision of the estimates of the hypothesized QTL locations and effects, as well as the power to detect a significant QTL, depend on the magnitude of QTL effect, the nature of the effect, the level of heritability, and sample size (Table 5). The QTL of a larger effect can be better mapped to a genomic location than that of a smaller effect. Figure 3, A and B, illustrates the landscapes of the log-likelihood ratio test statistics as a function of the locations of the two QTL for the one-stage design. The maxima of the landscape is closer to the coordinate of the true positions of the two QTL when an endosperm trait has a larger (Figure 3A) than lower heritability (Figure 3B).

For a small sample size (200) and a low heritability trait (0.20), the additive effect of a large QTL and its additive × additive effect with other QTL can be well estimated. The estimation of the additive effect of a small QTL displays significantly increased accuracy and precision when sample size and/or the heritability is increased. The dominant effects among different alleles at the same QTL cannot be well estimated for a small sample size and low-heritability endosperm trait (Table 5). Both the accuracy and precision of the estimates of dominant effects can be increased with increased sample size and heritability. It seems that the estimation precision for dominant effects can be better improved with the increased level of heritability than increased sample size. It is difficult to precisely estimate the additive × dominant, dominant × additive, and dominant × dominant epistatic effects of the two QTL for an endosperm trait of low heritability. Even the estimates of these parameters cannot be well improved when the sample size used is increased from 200 to 400.

In general, the one-stage design that ignores marker segregation within backcross families can be used to estimate the position and additive and additive × additive effect of a large QTL on the endosperm. But it is difficult to achieve acceptable estimation precision for dominant and epistatic additive × dominant, dominant × dominant, and dominant × dominant effects under the one-stage design. It appears that a high heritability level is more important in improving the estimates of these parameters than the increase of sample size.

Two-stage hierarchical design: When an endosperm trait is under low genetic control ($H^2 = 0.2$), the two-stage hierarchical design, which captures both between- and within-family variation, is not advantageous in the esti-

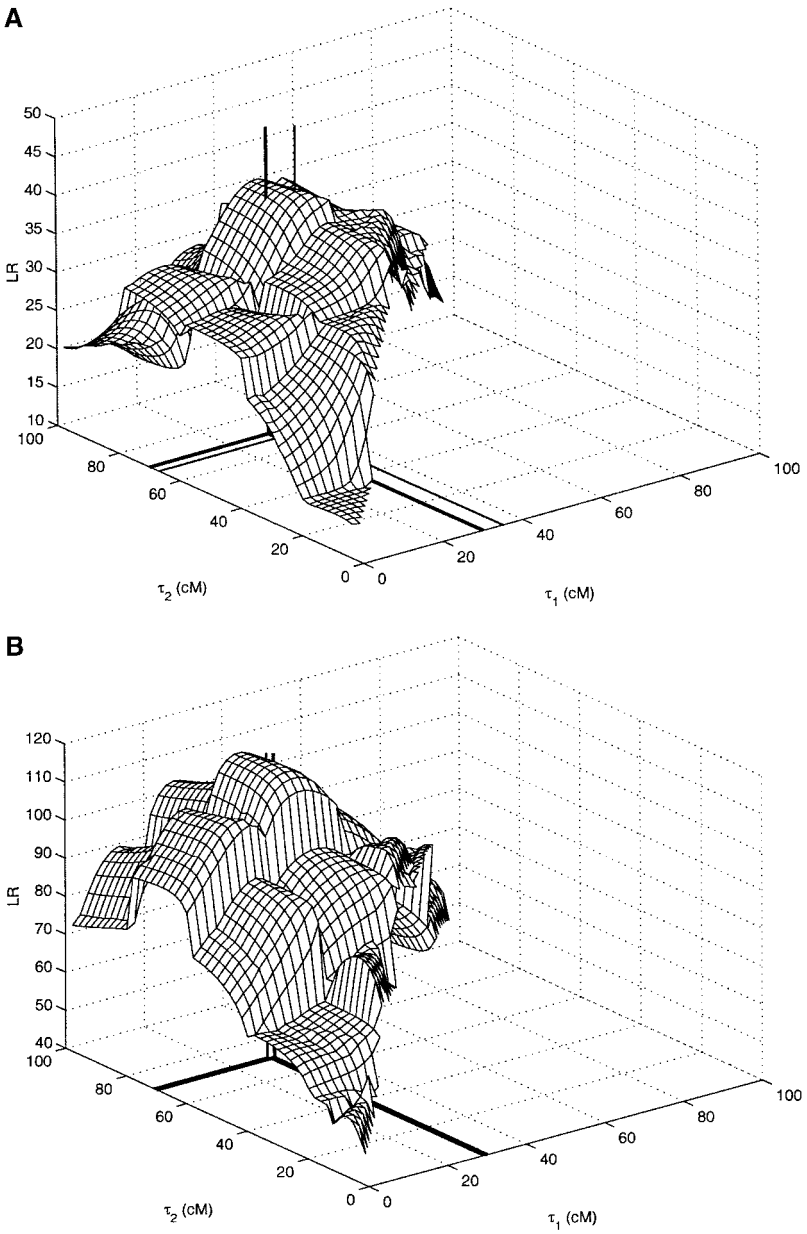

FIGURE 3.—The landscapes of log-likelihood-ratio test statistics as a function of genome positions of two QTL affecting an endosperm trait of heritabilities 0.2 (A) and 0.6 (B) under the one-stage design. τ_1 and τ_2 denote the genome positions of the two epistatic QTL, respectively. The thick bottom lines indicate true QTL positions.

mation precision of two epistatic QTL over the one-stage design, which captures only between-family variation (Table 6). This is in sharp contrast to the result of the one-QTL-fitting model in which the two-stage hierarchical design can provide more precise estimation of QTL parameters than the one-stage design (Tables 3 and 4).

The two-stage hierarchical design displays striking advantages in estimating the epistasis of QTL for an endosperm trait of high heritability ($H^2 = 0.6$; Table 5). In this case, the position of the two epistatic QTL, including one with a smaller effect, can be precisely mapped, as seen from reduced MSEs. For example, the MSE of the position of the smaller QTL is 42.33 under the one-stage design (Table 5), whereas it is reduced to 9.44–22.05 for the same sample size under the two-stage hierarchical design (Table 6). Figure 4, A–C, illustrates the landscapes of the log-likelihood-ratio test statistics of the two putative QTL with epistasis calculated from one simulation run for the different sampling strategies. The peaks of the LR landscapes are close to the true positions of the two QTL, suggesting that their positions can be well estimated.

Aside from precise estimation of additive and additive × additive (i_{aa}) effects, the precision of the estimates of different dominant effects and additive × dominant (j_{ad}), dominant × additive (k_{da}), and dominant × dominant effects (l_{dd}) can be improved for a high heritability endosperm trait from the two-stage hierarchical design (Table 6). It appears that the MLEs of the additive effects, i_{aa}, j_{ad}, and k_{da}, are more precise than those of the

TABLE 6

The MLEs of QTL parameters and mean square errors (MSE) of the estimates among 100 replicated simulations at different heritability levels (H^2) and sampling designs under the two-stage hierarchical design

	$H^2 = 0.2$ ($\sigma_\varepsilon^2 = 3$)						$H^2 = 0.6$ ($\sigma_\varepsilon^2 = 0.5$)							
	10 × 40		20 × 20		40 × 10		10 × 40		20 × 20		40 × 10		400 × 1	
True parameter	MLE	MSE	MLE	MSE	MLE	MSE	MLE	MSE	MLE	MSE	MLE	MSE	MLE	MSE
$\tau_1 = 30$	30.33	13.67	30.00	20.00	33.00	35.67	29.56	5.44	29.74	6.89	30.60	5.40	29.60	4.60
$\tau_2 = 70$	68.33	30.33	68.00	34.00	65.44	53.44	70.56	9.44	71.42	22.05	71.00	20.60	68.40	13.80
$\mu = 10.0$	9.92	0.22	9.90	0.40	9.82	0.24	9.98	0.05	10.00	0.04	10.00	0.04	9.95	0.03
$a_1 = 1.0$	0.90	0.15	0.64	0.23	0.78	0.34	0.92	0.06	0.96	0.04	0.97	0.02	0.93	0.05
$a_2 = 0.5$	0.16	0.50	0.33	0.27	0.27	0.58	0.29	0.15	0.46	0.07	0.46	0.03	0.44	0.10
$d_1^1 = 0.8$	0.68	0.69	0.65	0.68	0.81	1.03	0.72	0.30	0.85	0.18	0.75	0.13	0.74	0.11
$d_1^2 = 0.4$	0.65	0.79	0.72	1.67	1.22	1.78	0.49	0.09	0.49	0.26	0.64	0.30	0.65	0.28
$d_2^1 = 0.5$	0.32	1.59	0.41	1.12	0.57	1.03	0.39	0.20	0.58	0.23	0.39	0.23	0.55	0.07
$d_2^2 = 0.5$	0.55	0.86	1.13	0.85	0.49	0.46	0.48	0.10	0.74	0.21	0.52	0.10	0.59	0.23
$i_{aa} = 0.6$	0.48	0.10	0.33	0.16	0.53	0.18	0.51	0.02	0.56	0.02	0.58	0.01	0.56	0.03
$j_{ad} = -0.4$	−0.37	0.35	0.13	0.40	−0.20	0.46	−0.36	0.07	−0.33	0.09	−0.42	0.04	−0.31	0.05
$k_{da} = -0.2$	0.08	0.59	−0.15	0.40	0.13	0.94	0.03	0.18	−0.13	0.10	−0.11	0.08	−0.13	0.15
$l_{dd} = 0.2$	0.07	3.71	−0.54	9.19	−0.33	6.63	0.41	0.74	−0.38	1.68	0.18	0.92	−0.03	0.84
σ_ε^2	2.50	0.31	2.73	0.23	2.81	0.12	0.47	0.00	0.47	0.00	0.47	0.00	0.46	0.00
Percentage[a]	60		40		45		90		95		100		100	

[a] The percentage of the simulations in which a significant QTL at the significance level $\alpha = 0.05$ is detected.

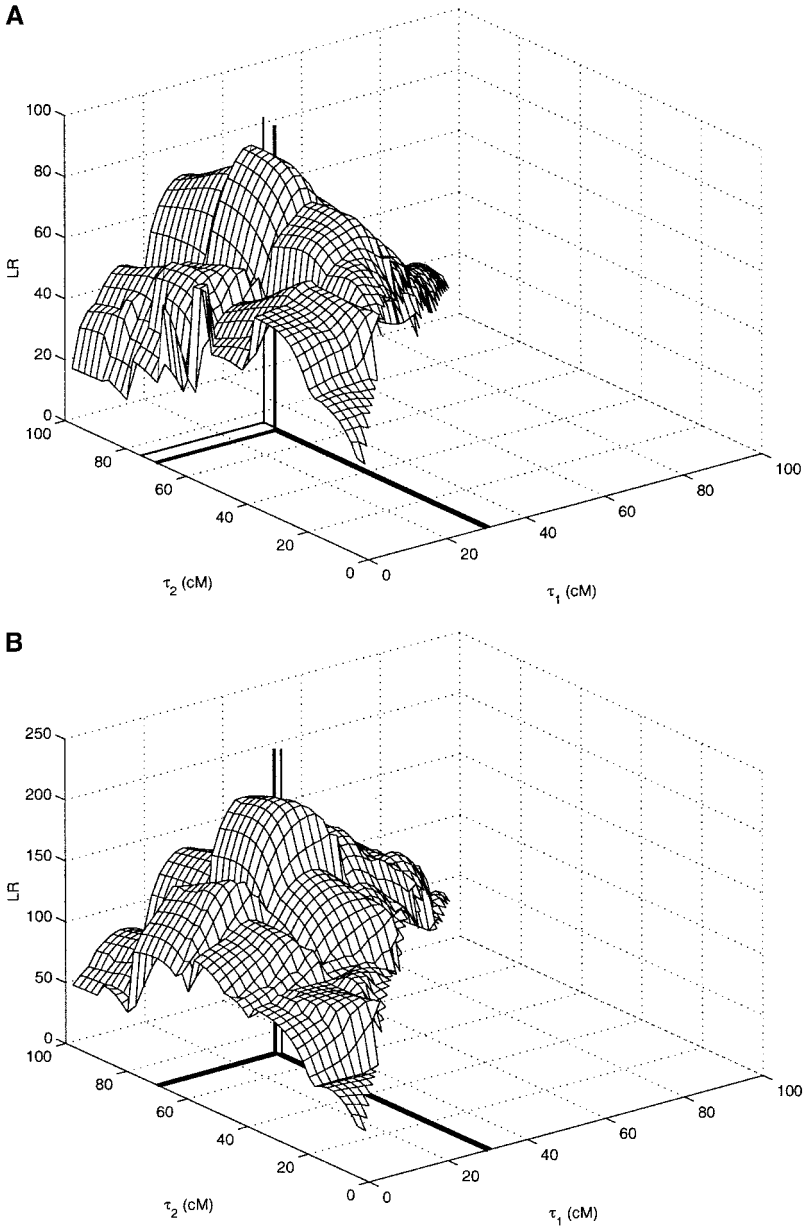

FIGURE 4.—The landscapes of log-likelihood-ratio test statistics as a function of genome positions of two QTL affecting an endosperm trait of heritabilities 0.2 (A) and 0.6 (B) under a 40 × 10 sampling design of the two-stage hierarchical design. τ_1 and τ_2 denote the genome positions of the two epistatic QTL, respectively. The thick bottom lines indicate true QTL positions.

dominant effects and l_{dd} among all different sampling strategies. However, different sampling strategies display different estimation precision of the MLEs of the dominant effects and l_{dd}. For example, strategy 10 × 40 can provide more precise estimates of the dominant effects of one Q over two q's (d_1^2 and d_2^2), whereas more precise estimates of the dominant effects of two Q's over one q (d_1^1 and d_2^1) can be provided by strategy 400 × 1 (Table 6). Relatively speaking, strategy 10 × 40 can provide a better estimate of l_{dd} than other designs.

In summary, the two-stage hierarchical design is better for the estimation of epistatic QTL than the one-stage design when an endosperm trait has a high heritability. Different sampling designs have different effects on the estimation of dominant effects and dominant × dominant epistatic effects. Depending on the nature of experimental material and breeders' purposes, these sampling designs can be selectively used.

DISCUSSION

The genetic improvement of grain quality in crop plants has now become a major focus in many plant breeding programs (SADIMANTARA *et al.* 1997; MAZUR *et al.* 1999; TAN *et al.* 1999; VAN DER MEER *et al.* 2001; WANG *et al.* 2001). Compared to yield improvement, however, quality improvement will be much more challenging because traits affecting grain quality are endosperm specific, and the endosperm being a triploid tissue has complex trisomic inheritance. Also, since the

endosperm is one generation ahead of its mother plant, it is difficult to predict the segregation patterns of the endosperm genes on the basis of marker information from the maternal parent. These have been two major obstructions in improving grain quality through a marker-assisted selection strategy. In this article, we employ traditional quantitative genetic principles and powerful statistical technologies to develop a new theoretical method for mapping QTL underlying endosperm traits in autogamous plants.

Unlike traditional diploid mapping, genetic mapping of the triploid endosperm requires estimation of a large number of genetic parameters because there are more copies of alleles at each locus. Statistically, an increased number of unknown parameters to be estimated can create numerous problems in estimation, such as larger biases, larger sampling errors, and lower power. We compare the differences between two alternative experimental designs in the capacity to minimize the problems due to an increased number of unknowns in our mapping model. The first model, the *one-stage design*, draws marker information only from the current generation of plants, and the second model, the *two-stage hierarchical design*, draws marker information from both the current generation and the self-fertilized progeny that the plants generate. In practice, the one-stage design is less expensive because no marker information for the progeny of the experimental plants is needed. However, theoretical simulations indicate that the estimation of QTL parameters, especially dominant effects, under the one-stage design can be seriously overestimated when the heritability of an endosperm trait and/or sample size is not large. The two-stage hierarchical design, which extracts marker information from two successive generations, can improve the accuracy and precision of QTL parameters including dominant effects. This is especially remarkable when an endosperm has a low heritability and/or the experiment is based on a limited number of backcross progeny. For the two-stage hierarchical design, different allocations of a given number of samples between the backcross plants and their autogamous progeny also affect the parameter estimation.

The endosperm used as a mapping tissue has several theoretical advantages in its own right. For example, since the endosperm is triploid and contains extra gene copies, a number of hypotheses regarding the role of gene interactions within or between loci in plant development, adaptation, and evolution can be addressed using the endosperm as a model study material. Moreover, the endosperm is formed due to the fusing of two polar nuclei cells with a sperm nucleus. The endosperm is an ideal material to study maternal effects and interaction effects with nuclei genes on plant behavior. Our mapping model can be readily extended to include maternal effects. Also, our limited knowledge about the genetic mechanisms underlying endosperm traits makes it practically impossible to understand the evolutionary significance of double fertilization in higher plants (Friedman 1990, 1998), which produces the endosperm. The genetic mapping of endosperm provides a powerful means for addressing this fundamental question in plant evolution.

In this article, we extend a one-QTL analysis to include QTL with epistasis in the control of an endosperm trait. The role of epistasis in trait control, evolution, and breeding has been reconciled recently thanks to the use of powerful molecular markers that can effectively dissect a complex trait into its individual QTL locus components (Doebley *et al.* 1995; Lark *et al.* 1995; Lukens and Doebley 1999; Cheverud 2000; Kim and Rieseberg 2001). However, because epistasis results from the dependence of different genes activated in a physiological process or biochemical pathway (Phillips 1998) so that the precise estimation of epistatic effects on quantitative variation is always difficult, many statistical methods for QTL mapping in the current literature assume no epistasis, or they are simply based on a two-way analysis of variance specifying the interaction effect of a given pair of markers (see the references listed above). A few methodologies with power to detect epistasis include Kao *et al.*'s (1999) multiple-interval mapping, Du and Hoeschele's (2000) finite locus model, and Jannink and Jansen's (2001) one-dimensional genome search. These methods allow for the determination of epistasis between different QTL or between QTL and polygenic background. Our method can be specifically used to map epistatic QTL affecting the expression of complex traits on triploid endosperm derived from double fertilization in flowering plants.

The idea described in our mapping model can be extended to map pleiotropic QTL affecting both grain yield and grain quality, although this will be a challenging statistical issue. Traits for grain yield, *e.g.*, seed number and seed weight, are located on the diploid tissues of backcross plants. As a consequence, the QTL for grain yield should be modeled in the mode of disomic inheritance. However, endosperm traits for grain quality undergo trisomic inheritance and, thus, they should be modeled differently. A statistical framework should be built to model the pleiotropic effect of a QTL on diploid tissues and triploid endosperm. Such a framework will permit plant breeders to design a more efficient breeding program for selecting superior genotypes with high yield and high quality.

In this study, we report our results in a backcross population for an autogamous plant system. The application of this model in an F_2 design is not difficult, except for the segregation of more marker genotypes in both the current and progeny generations. For an autogamous plant, the eggs and two polar nuclei cells are self fertilized so that the frequencies of male gamete genotypes are identical to those of female gamete genotypes. But in an allogamous plant, such as maize, each female gamete from each mother plant will be pollinated

by all possible male gametes from the pollen pool. This difference should be considered when the current model is used to study the genetics of the allogamous endosperm. All of these issues deserve in-depth investigations.

We thank three anonymous reviewers for their helpful comments on earlier versions of this manuscript. This work is partially supported by a grant from National Science Foundation (DMS9971586) to C.G., an Outstanding Young Investigators Award of the National Science Foundation of China (30128017), and a University of Florida Research Opportunity Fund (02050259) to R.W. The publication of this manuscript is approved as Journal Series no. R-08586 by the Florida Agricultural Experiment Station.

LITERATURE CITED

Benner, M. S., R. L. Phillips, J. A. Kirihara and J. W. Messing, 1989 Genetic analysis of methionine-rich storage protein accumulation in maize. Theor. Appl. Genet. **78:** 761–767.

Bogyo, T. P., R. C. M. Lance, P. Chevalier and R. A. Nilan, 1988 Genetic models for quantitatively inherited endosperm characters. Heredity **60:** 61–67.

Cheverud, J. M., 2000 Detecting epistasis among quantitative trait loci, pp. 5881 in *Epistasis and the Evolutionary Process*, edited by J. B. Wolf, E. D. Brodie III and M. J. Wade. Oxford University Press, New York.

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

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum likelihood from incomplete data via EM algorithm. J. R. Stat. Soc. Ser. B **39:** 1–38.

Doebley, J., A. Stec and C. Gustus, 1995 *teosinte branched1* and the origin of maize: evidence for epistasis and the evolution of dominance. Genetics **141:** 333–346.

Du, F. X., and I. Hoeschele, 2000 Estimation of additive, dominance and epistatic variance components using finite locus models implemented with a single-site Gibbs and a descent graph sampler. Genet. Res. **76:** 187–198.

Friedman, W. E., 1990 Double fertilization in *Ephedra*, a nonflowering seed plant: its bearing on the origin of angiosperms. Science **247:** 951–954.

Friedman, W. E., 1998 The evolution of double fertilization and endosperm: an "historical" perspective. Sex Plant Reprod. **11:** 6–16.

Gale, M. D., 1976 High α-amylase breeding and genetical aspects of the problem. Cereal Res. Commun. **4:** 231–243.

Jannink, J. L., and R. Jansen, 2001 Mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics **157:** 445–454.

Jensen, W. A., 1998 Double fertilization: a personal view. Sex Plant Reprod. **11:** 1–5.

Kao, C. H., Z-B. Zeng and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152:** 1203–1216.

Kim, S. C., and L. H. Rieseberg, 2001 The contribution of epistasis to species differences in annual sunflowers. Mol. Ecol. **10:** 683–690.

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

Lark, K. G., K. Chase, F. Adler, L. M. Mansur and J. H. Orf, 1995 Interactions between quantitative trait loci in soybean in which trait variation at one locus is conditional upon a specific allele at another. Proc. Natl. Acad. Sci. USA **92:** 4656–4660.

Lukens, L. N., and J. Doebley, 1999 Epistatic and environmental interactions for quantitative trait loci involved in maize evolution. Genet. Res. **74:** 291–302.

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

Meng, X. L., and D. B. Rubin, 1993 Maximum likelihood estimation via the ECM algorithm—A general framework. Biometrika **80:** 267–278.

Mo, H. D., 1987 Genetic expression for endosperm traits, pp. 478–487 in *Proceedings of the 2nd International Conference on Quantitative Genetics*. Sinauer Associates, Sunderland, MA.

Phillips, P. C., 1998 The language of gene interaction. Genetics **149:** 1167–1171.

Pooni, H., S. I. Kumar and G. S. Khush, 1992 A comprehensive model for disomically inherited metrical traits expressed in triploid tissues. Heredity **69:** 166–174.

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

Tan, Y. F., J. X. Li, S. B. Yu, Y. Z. Xing, C. G. Xu *et al.*, 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.

van der Meer, I. M., A. G. Bovy and D. Bosch, 2001 Plant-based raw material: improved food quality for better nutrition via plant genomics. Curr. Opin. Biotech. **12:** 488–492.

Wang, X. L., and B. A. Larkins, 2001 Genetic analysis of amino acid accumulation in opaque-2 maize endosperm. Plant Physiol. **125:** 1766–1777.

Wang, X. L., Y. M. Woo, C. S. Kim and B. A. Larkins, 2001 Quantitative trait locus mapping of loci influencing elongation factor 1 alpha content in maize endosperm. Plant Physiol. **125:** 1271–1282.

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

Communicating editor: P. D. Keightley

APPENDIX A: DERIVATION OF CONDITIONAL PROBABILITIES

We describe the procedures for deriving the conditional probabilities of endosperm QTL genotypes upon marker genotypes of the current backcross (one-stage design) and marker genotypes of the backcross and its progeny for an autogamous species (two-stage hierarchical design). Suppose two inbred lines P_1 and P_2, with respective genotypes $M_\eta M_\eta QQM_{\eta+1}M_{\eta+1}$ and $m_\eta m_\eta qqm_{\eta+1}$ $m_{\eta+1}$ at two flanking markers $\mathcal{M}_\eta$ and $\mathcal{M}_{\eta+1}$ and the QTL they bracket. Denote the recombination fraction between the two markers by r and those between $\mathcal{M}_\eta$ and the QTL as well as the QTL and $\mathcal{M}_{\eta+1}$ by r_1 and r_2. The F_1 hybrid of the two lines has genotype $M_\eta m_\eta QqM_{\eta+1}$ $m_{\eta+1}$, which produces a total of eight joint marker-QTL gamete genotypes (assuming generation t), denoted by G, with the corresponding frequencies expressed as

Joint gamete genotype	Symbol	Frequency
$M_\eta QM_{\eta+1}$	$G^{(t)}_{111}$	$\frac{1}{2}(1-r_1)(1-r_2)$
$M_\eta qM_{\eta+1}$	$G^{(t)}_{101}$	$\frac{1}{2}r_1r_2$
$M_\eta Qm_{\eta+1}$	$G^{(t)}_{110}$	$\frac{1}{2}(1-r_1)r_2$
$M_\eta qm_{\eta+1}$	$G^{(t)}_{100}$	$\frac{1}{2}r_1(1-r_2)$
$m_\eta QM_{\eta+1}$	$G^{(t)}_{011}$	$\frac{1}{2}r_1(1-r_2)$
$m_\eta qM_{\eta+1}$	$G^{(t)}_{001}$	$\frac{1}{2}(1-r_1)r_2$
$m_\eta Qm_{\eta+1}$	$G^{(t)}_{010}$	$\frac{1}{2}r_1r_2$
$m_\eta qm_{\eta+1}$	$G^{(t)}_{000}$	$\frac{1}{2}(1-r_1)(1-r_2)$

Crossing the F_1 with the homozygous P_2 generates the B_1 backcross generation *t*, in which there are eight joint marker-QTL diploid genotypes, each corresponding to a joint gamete genotype above. These joint diploid genotypes can be categorized into four groups of marker genotypes (denoted by *Z*),

Marker diploid genotype	Symbol	Frequency
$M_\eta m_\eta M_{\eta+1} m_{\eta+1}$	$Z_{11}^{(t)}$	$\frac{1}{2}(1-r)$
$M_\eta m_\eta m_{\eta+1} m_{\eta+1}$	$Z_{10}^{(t)}$	$\frac{1}{2}r$
$m_\eta m_\eta M_{\eta+1} m_{\eta+1}$	$Z_{01}^{(t)}$	$\frac{1}{2}r$
$m_\eta m_\eta m_{\eta+1} m_{\eta+1}$	$Z_{00}^{(t)}$	$\frac{1}{2}(1-r)$

For a traditional backcross design, the recombination fraction between the two markers is estimated on the basis of the numbers of these different marker genotypes in the population. The subsequent QTL analysis is performed by associating the marker genotypes with phenotypic values measured on diploid organs of the backcross progeny. However, when mapping endosperm traits, we must consider how the embryo and endosperm are formed through reproductive behavior. Whereas the embryo is formed due to the fusing of one haploid egg and one sperm cell, the formation of the endosperm results from the fusing of two central cells, whose genotype (denoted by *GG*) is the duplication of the egg, and one sperm cell. Since there are different levels of heterozygosity, each of the eight joint backcross genotypes above produces different compositions of the egg genotypes and polar nuclei genotypes. The first backcross diploid genotype derived from the first gamete genotype $G_{111}^{(t)}$ of the F_1 produces eight egg genotypes (and therefore polar nuclei genotypes), whereas those from the rest of the gamete genotypes produce four (from $G_{101}^{(t)}$, $G_{110}^{(t)}$, and $G_{011}^{(t)}$), two (from $G_{100}^{(t)}$, $G_{001}^{(t)}$, and $G_{010}^{(t)}$), and one egg genotype (from $G_{000}^{(t)}$). For an autogamous plant, the genotypes and frequencies of sperms it produces are identical to those of eggs. Because the one-stage and two-stage hierarchical designs use different marker information, we derive the conditional probabilities of the endosperm QTL genotypes separately for these two models.

One-stage design: In this model, we use only marker genotypes from the backcross plants, without considering the marker information from the progeny of the backcross. Thus, our interest here is how to generate endosperm QTL genotypes given a particular marker genotype of the backcross. The endosperm QTL genotypes (generation $t+1$) produced by the first backcross genotype for an autogamous plant are the product of the array of the polar nuclei genotypes and the array of the sperm genotypes,

$$\begin{bmatrix} QQ & \frac{1}{2} \\ qq & \frac{1}{2} \end{bmatrix} \times \begin{bmatrix} Q & \frac{1}{2} \\ q & \frac{1}{2} \end{bmatrix}^T,$$

which results in four endosperm QTL genotypes *QQQ*, *QQq*, *Qqq*, and *qqq*, each with a frequency of 1/4.

The second backcross genotype $G_{101}^{(t)}$ with the same marker genotype $Z_{11}^{(t)}$ produces one polar nuclei QTL genotype (*qq*) and one sperm QTL genotype (*q*), which thus results in only one single endosperm QTL genotype *qqq*. The conditional probabilities of the endosperm QTL genotypes upon the diploid marker genotypes $Z_{11}^{(t)}$ of the backcross can be calculated as

$$\text{Prob}\left[QQQ|Z_{11}^{(t)}\right] = \frac{(1/2)(1-r_1)(1-r_2)\cdot 1/4}{(1/2)(1-r)} = \frac{(1-r_1)(1-r_2)}{4(1-r)},$$

$$\text{Prob}\left[QQq|Z_{11}^{(t)}\right] = \frac{(1/2)(1-r_1)(1-r_2)\cdot 1/4}{(1/2)(1-r)} = \frac{(1-r_1)(1-r_2)}{4r}$$

$$\text{Prob}\left[Qqq|Z_{11}^{(t)}\right] = \frac{(1/2)(1-r_1)(1-r_2)\cdot 1/4}{(1/2)(1-r)} = \frac{(1-r_1)(1-r_2)}{4r}$$

$$\text{Prob}\left[qqq|Z_{11}^{(t)}\right] = \frac{(1/2)(1-r_1)(1-r_2)\cdot 1/4 + (1/2)r_1r_2}{(1/2)(1-r)}$$

$$= \frac{1-r_1-r_2+5r_1r_2}{4(1-r)},$$

where crossover interference is ignored.

Similarly, we can derive the conditional probabilities of the endosperm QTL genotypes upon the other three diploid marker genotypes in the backcross (see Table 1).

Two-stage hierarchical design: When the marker information from both the backcross and its progeny is considered simultaneously, we should see the segregation of both markers and QTL in the progeny of the backcross. Under the two-stage hierarchical design, the endosperm genotypes (in generation $t+1$) produced by the first backcross genotype for an autogamous plant are the product of the array of eight polar nuclei genotypes (denoted by *GG*) and the array of eight sperm genotypes,

$$\begin{bmatrix} GG_{111}^{(t+1)} & \frac{1}{2}(1-r_1)(1-r_2) \\ GG_{101}^{(t+1)} & \frac{1}{2}r_1r_2 \\ GG_{110}^{(t+1)} & \frac{1}{2}(1-r_1)r_2 \\ GG_{100}^{(t+1)} & \frac{1}{2}r_1(1-r_2) \\ GG_{011}^{(t+1)} & \frac{1}{2}r_1(1-r_2) \\ GG_{001}^{(t+1)} & \frac{1}{2}(1-r_1)r_2 \\ GG_{010}^{(t+1)} & \frac{1}{2}r_1r_2 \\ GG_{000}^{(t+1)} & \frac{1}{2}(1-r_1)(1-r_2) \end{bmatrix} \times \begin{bmatrix} G_{111}^{(t+1)} & \frac{1}{2}(1-r_1)(1-r_2) \\ G_{101}^{(t+1)} & \frac{1}{2}r_1r_2 \\ G_{110}^{(t+1)} & \frac{1}{2}(1-r_1)r_2 \\ G_{100}^{(t+1)} & \frac{1}{2}r_1(1-r_2) \\ G_{011}^{(t+1)} & \frac{1}{2}r_1(1-r_2) \\ G_{001}^{(t+1)} & \frac{1}{2}(1-r_1)r_2 \\ G_{010}^{(t+1)} & \frac{1}{2}r_1r_2 \\ G_{000}^{(t+1)} & \frac{1}{2}(1-r_1)(1-r_2) \end{bmatrix},$$

which results in nine triploid endosperm marker genotypes, each containing four QTL genotypes *QQQ*(Q_3), *QQq*(Q_2), *Qqq*(Q_1), and *qqq*(Q_0). In the same manner, nine diploid embryo marker genotypes are formed with frequencies

Marker embryo genotype	Symbol	Frequency
$M_\eta M_\eta M_{\eta+1} M_{\eta+1}$	$Z_{22}^{(t+1)}$	$\frac{1}{4}(1-r)^2$
$M_\eta M_\eta M_{\eta+1} m_{\eta+1}$	$Z_{21}^{(t+1)}$	$\frac{1}{2}r(1-r)$
$M_\eta M_\eta m_{\eta+1} m_{\eta+1}$	$Z_{20}^{(t+1)}$	$\frac{1}{4}r^2$
$M_\eta m_\eta M_{\eta+1} M_{\eta+1}$	$Z_{12}^{(t+1)}$	$\frac{1}{2}r(1-r)$
$M_\eta m_\eta M_{\eta+1} m_{\eta+1}$	$Z_{11}^{(t+1)}$	$\frac{1}{2}(1-2r+2r^2)$
$M_\eta m_\eta m_{\eta+1} m_{\eta+1}$	$Z_{10}^{(t+1)}$	$\frac{1}{2}r(1-r)$
$m_\eta m_\eta M_{\eta+1} M_{\eta+1}$	$Z_{02}^{(t+1)}$	$\frac{1}{4}r^2$
$m_\eta m_\eta M_{\eta+1} m_{\eta+1}$	$Z_{01}^{(t+1)}$	$\frac{1}{2}r(1-r)$
$m_\eta m_\eta m_{\eta+1} m_{\eta+1}$	$Z_{00}^{(t+1)}$	$\frac{1}{4}(1-r)^2$

By isolating DNA from embryos, their marker genotypes can be characterized. The second diploid backcross genotype derived from the F_1's second gamete genotype $G_{101}^{(t)}$ has the same marker genotype $Z_{11}^{(t)}$ as the first diploid backcross genotype. The second backcross genotype produces four polar nuclei genotypes and four sperm genotypes, which are arrayed as

$$\begin{bmatrix} GG_{101}^{(t+1)} & \frac{1}{2}(1-r) \\ GG_{100}^{(t+1)} & \frac{1}{2}r \\ GG_{001}^{(t+1)} & \frac{1}{2}r \\ GG_{000}^{(t+1)} & \frac{1}{2}(1-r) \end{bmatrix} \times \begin{bmatrix} G_{101}^{(t+1)} & \frac{1}{2}(1-r) \\ G_{100}^{(t+1)} & \frac{1}{2}r \\ G_{001}^{(t+1)} & \frac{1}{2}r \\ G_{000}^{(t+1)} & \frac{1}{2}(1-r) \end{bmatrix}^T,$$

which also results in nine triploid endosperm marker genotypes, but each containing only a single QTL genotype *qqq* because of no allele *Q* in this backcross genotype.

The same endosperm genotypes derived from the first and second backcross genotype will be summed up because of their same marker genotype $Z_{11}^{(t)}$. With similar analyses we can sum up the same endosperm genotypes for the third and fourth, fifth and sixth, and seventh and eighth backcross genotypes.

The conditional probabilities of the endosperm QTL genotypes, conditional upon diploid marker genotypes of the backcross (t) and its autogamous progeny ($t+1$) under the two-stage hierarchical design can be derived according to Bayes' theorem. For example, the conditional probability of endosperm QTL genotypes *QQQ*, *QQq*, *Qqq*, and *qqq*, conditional upon a diploid marker genotype Z_{111} in the two successive generations can be calculated as

$$\text{Prob}(QQQ|Z_{111}^{(t)}, Z_{111}^{(t+1)}) = \frac{\text{Prob}(Z_{111}^{(t)}, Z_{111}^{(t+1)}, QQQ)}{\text{Prob}(Z_{111}^{(t)}, Z_{111}^{(t+1)})} = \frac{(1-r_1)^3(1-r_2)^3}{(1-r)^3},$$

$$\text{Prob}(QQq|Z_{111}^{(t)}, Z_{111}^{(t+1)}) = \frac{\text{Prob}(Z_{111}^{(t)}, Z_{111}^{(t+1)}, QQq)}{\text{Prob}(Z_{111}^{(t)}, Z_{111}^{(t+1)})} = \frac{r_1 r_2(1-r_1)^2(1-r_2)^2}{(1-r)^3},$$

$$\text{Prob}(Qqq|Z_{111}^{(t)}, Z_{111}^{(t+1)}) = \frac{\text{Prob}(Z_{111}^{(t)}, Z_{111}^{(t+1)}, Qqq)}{\text{Prob}(Z_{111}^{(t)}, Z_{111}^{(t+1)})} = \frac{r_1 r_2(1-r_1)^2(1-r_2)^2}{(1-r)^3},$$

$$\text{Prob}(qqq|Z_{111}^{(t)}, Z_{111}^{(t+1)}) = \frac{\text{Prob}(Z_{111}^{(t)}, Z_{111}^{(t+1)}, qqq)}{\text{Prob}(Z_{111}^{(t)}, Z_{111}^{(t+1)})} = \frac{r_1^2 r_2^2(1-r_1)(1-r_2) + r_1 r_2(1-r)^2}{(1-r)^3}.$$

Similarly, we can derive the conditional probabilities of the endosperm QTL genotypes, conditional upon other diploid marker genotypes at the backcross and its autogamous progeny (see Table 2).

APPENDIX B: ESTIMATORS OF MODEL PARAMETERS IN THE M STEP

Additive-dominant model: The MLEs of the unknown QTL parameters are obtained by differentiating the likelihood with respect to each unknown, setting the derivatives equal to zero, and solving the log-likelihood equations. The EM algorithm is used to obtain the MLE of the genetic mean of each endosperm genotype at a putative QTL bracketed by a marker interval.

The expressions of the log-likelihood equations for estimating genotypic means and residual variance in the M step are given as

$$\hat{\mu}_k = \frac{\sum_{i=1}^{M} P_{ik} y_i}{\sum_{i=1}^{M} P_{ik}} \quad \text{or} \quad \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i} P_{ijk} y_{ij}}{\sum_{i=1}^{M}\sum_{j=1}^{N_i} P_{ijk}},$$

$$\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{3} P_{ik}(y_i - \hat{\mu}_k)^2 \quad \text{or} \quad \frac{1}{\sum_{i=1}^{M} N_i}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k=1}^{3} P_{ijk}(y_{ij} - \hat{\mu}_k)^2,$$

for the one-stage design and the two-stage hierarchical design, respectively, where P_{ik} and P_{ijk} are the posterior probabilities of the QTL genotypes under the two different designs, respectively (see Equation 5). The MLE of the QTL position is obtained by treating the recombination fraction between the QTL and one marker (r_1 or r_2) as a fixed parameter.

After the genotypic means (**m**) are estimated, a linear transformation is used to estimate the QTL effect parameters (**e**). We have

$$\mathbf{m} = \mathbf{De},$$

$$\hat{\mathbf{e}} = \mathbf{D}^{-1}\hat{\mathbf{m}},$$

where

$$\mathbf{D} = \begin{bmatrix} 1 & \frac{3}{2} & 0 & 0 \\ 1 & \frac{1}{2} & 1 & 0 \\ 1 & -\frac{1}{2} & 0 & 1 \\ 1 & -\frac{3}{2} & 0 & 0 \end{bmatrix},$$

and

$$\mathbf{D}^{-1} = \begin{bmatrix} \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & -\frac{1}{3} \\ -\frac{2}{3} & 1 & 0 & -\frac{1}{3} \\ -\frac{1}{3} & 0 & 1 & -\frac{2}{3} \end{bmatrix}.$$

On the basis of invariance property of maximum-likelihood estimates, the estimate of **e** is the MLE because the estimate of **m** is the MLE.

Epistatic model: The two-QTL quantitative genetic model for a triploid endosperm trait can be expressed in matrix form,

$$\mathbf{m} = \mathbf{Ae},$$

where $\mathbf{m} = (\mu_{k_1k_2})_{16\times 1}$ is the vector for the genotypic means of 16 QTL genotypes ($k_1, k_2 = 0, \ldots, 3$), which can be estimated in the M step (a similar M step was described above for the additive-dominant model); $\mathbf{e} = (\mu, a_1, a_2, d_1^1, d_1^2, d_2^1, d_2^2, i_{aa}, j_{ad}^1, j_{ad}^2, k_{da}^1, k_{da}^2, l_{dd}^{11}, l_{dd}^{12}, l_{dd}^{21}, l_{dd}^{22})^{\mathrm{T}}$ is the vector for the unknown QTL effects specified in the quantitative inheritance of the endosperm, and **A** is the design matrix relating the genotypic means to the QTL effects. Based on quantitative genetic models of the triploid endosperm (Gale 1976; Mo 1987; Bogyo *et al.* 1988; Pooni *et al.* 1992), we have

$$\mathbf{A} = \begin{bmatrix}
1 & 3/2 & 3/2 & 0 & 0 & 0 & 0 & 9/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 3/2 & 1/2 & 0 & 0 & 1 & 0 & 3/4 & 3/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 3/2 & -1/2 & 0 & 0 & 0 & 1 & -3/4 & 3/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 3/2 & -3/2 & 0 & 0 & 0 & 0 & -9/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1/2 & 3/2 & 1 & 0 & 0 & 0 & 3/4 & 0 & 3/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1/2 & 1/2 & 1 & 0 & 1 & 0 & 1/4 & 1/2 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\
1 & 1/2 & -1/2 & 1 & 0 & 0 & 1 & -1/4 & 3/2 & -1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\
1 & 1/2 & -3/2 & 1 & 0 & 0 & 0 & -3/4 & 0 & -3/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -1/2 & 3/2 & 0 & 1 & 0 & 0 & -3/4 & 0 & 3/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -1/2 & 1/2 & 0 & 1 & 1 & 0 & -1/4 & -1/2 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\
1 & -1/2 & -1/2 & 0 & 1 & 0 & 1 & 1/4 & -3/2 & -1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\
1 & -1/2 & -3/2 & 0 & 1 & 0 & 0 & 3/4 & 0 & -3/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -3/2 & 3/2 & 0 & 0 & 0 & 0 & -9/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -3/2 & 1/2 & 0 & 0 & 1 & 0 & -3/4 & -3/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -3/2 & -1/2 & 0 & 0 & 0 & 1 & 3/4 & -3/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -3/2 & -3/2 & 0 & 0 & 0 & 0 & 9/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}.$$

Then, the MLE of **e** is obtained as $\hat{\mathbf{e}} = \mathbf{A}^{-1}\hat{\mathbf{m}}$. The sampling variance of $\hat{\mathbf{e}}$ is calculated by $\mathbf{A}^{-1}(\mathbf{A}^{-1})^{\mathrm{T}}\sigma^2$, whose elements on the diagonal are ($5/16$, $1/9$, $1/4$, $121/144$, $121/144$, $3/4$, 1, $4/81$, $28/81$, $28/81$, $157/324$, $157/324$, $775/81$, $40/3$, $775/81$, $40/3$). Using the design matrix **A**, the additive and additive × additive effects can be estimated precisely, with the sampling variances of the MLEs being a small proportion ($1/4 \sim 4/81$) of the estimated residual variance. Compared to additive and additive × additive effects, the precision of the MLEs of different dominant effects and additive × dominant and dominant × additive effects is reduced. The MLEs of different dominant × dominant effects have the lowest precision, whose sampling variances are enlarged relative to the residual variance ($775/81 \sim 40/3$).

To reduce the sampling variances of the MLEs of the QTL effect parameters, especially the dominant × dominant effects, we use single j_{ad}, k_{da}, and l_{dd} to capture information of different additive × dominant, dominant × additive, and dominant × dominant effects, respectively. In this case, the number of the unknown parameters in **e** is reduced to 11. We thus have a new design matrix

$$\mathbf{A} = (a_{kl})_{16\times 11} = \begin{bmatrix}
1 & 3/2 & 3/2 & 0 & 0 & 0 & 0 & 9/4 & 0 & 0 & 0 \\
1 & 3/2 & 1/2 & 0 & 0 & 1 & 0 & 3/4 & 3/2 & 0 & 0 \\
1 & 3/2 & -1/2 & 0 & 0 & 0 & 1 & -3/4 & 3/2 & 0 & 0 \\
1 & 3/2 & -3/2 & 0 & 0 & 0 & 0 & -9/4 & 0 & 0 & 0 \\
1 & 1/2 & 3/2 & 1 & 0 & 0 & 0 & 3/4 & 0 & 3/2 & 0 \\
1 & 1/2 & 1/2 & 1 & 0 & 1 & 0 & 1/4 & 1/2 & 1/2 & 1/2 \\
1 & 1/2 & -1/2 & 1 & 0 & 0 & 1 & -1/4 & 3/2 & -1/2 & 1/2 \\
1 & 1/2 & -3/2 & 1 & 0 & 0 & 0 & -3/4 & 0 & -3/2 & 0 \\
1 & -1/2 & 3/2 & 0 & 1 & 0 & 0 & -3/4 & 0 & 3/2 & 0 \\
1 & -1/2 & 1/2 & 0 & 1 & 1 & 0 & -1/4 & -1/2 & 1/2 & 1/2 \\
1 & -1/2 & -1/2 & 0 & 1 & 0 & 1 & 1/4 & -3/2 & -1/2 & 1/2 \\
1 & -1/2 & -3/2 & 0 & 1 & 0 & 0 & 3/4 & 0 & -3/2 & 0 \\
1 & -3/2 & 3/2 & 0 & 0 & 0 & 0 & -9/4 & 0 & 0 & 0 \\
1 & -3/2 & 1/2 & 0 & 0 & 1 & 0 & -3/4 & -3/2 & 0 & 0 \\
1 & -3/2 & -1/2 & 0 & 0 & 0 & 1 & 3/4 & -3/2 & 0 & 0 \\
1 & -3/2 & -3/2 & 0 & 0 & 0 & 0 & 9/4 & 0 & 0 & 0
\end{bmatrix}.$$

Below, we describe a procedure for estimating the 11 QTL parameters. In the E step, we have derived the posterior probability of a given endosperm QTL genotype k_1k_2 for the two epistatic QTL under the two-stage hierarchical design,

$$P_{ijk_1k_2} = \frac{p_{ijk_1k_2} f_{k_1k_2}(y_{ij})}{\sum_{k_1=0}^{3}\sum_{k_2=0}^{3} p_{ijk_1k_2} f_{k_1k_2}(y_{ij})},$$

$$i = 1, \ldots, M,\ j = 1, \ldots, N_i,\ k_1, k_2 = 0, \ldots, 3,$$

where $f_{k_1k_2}(y_{ij})$ is the normal distribution of endosperm phenotypes at the two QTL. We have

$$c_l = \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k_1=0}^{3}\sum_{k_2=0}^{3} P_{ijk_1k_2} y_{ij} a_{k_1k_2l}, \quad l = 1, \ldots, 11,$$

$$b_{lm}^{(t)} = \sum_{k_1=0}^{3}\sum_{k_2=0}^{3} a_{k_1k_2l} a_{k_1k_2m} \sum_{i=1}^{M}\sum_{j=1}^{N_i} P_{ijk_1k_2} f_{k_1k_2}(y_{ij}), \quad l, m = 1, \ldots, 11,$$

and

$$\mathbf{B} = (b_{lm})_{11\times 11},$$
$$\mathbf{C} = (c_1 \ldots c_{11})^T.$$

Then, the MLEs of the unknown QTL effects can be obtained by

$$\hat{\mathbf{e}} = B^{-1}\mathbf{C}.$$

The estimation of residual variance is given by

$$\hat{\sigma}^2 = \frac{1}{\sum_{i=1}^{M} N_i} \sum_{i=1}^{M}\sum_{j=1}^{N_i}\sum_{k_1=0}^{3}\sum_{k_2=0}^{3} P_{ijk_1k_2}(y_{ij} - \hat{\mu}_{k_1k_2})^2.$$