DOI: 10.1534/genetics.103.021642

Multiple-Interval Mapping for Quantitative Trait Loci Controlling Endosperm Traits

Chen-Hung Kao[1]

Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, Republic of China

Manuscript received August 28, 2003
Accepted for publication April 30, 2004

ABSTRACT

Endosperm traits are trisomic inheritant and are of great economic importance because they are usually directly related to grain quality. Mapping for quantitative trait loci (QTL) underlying endosperm traits can provide an efficient way to genetically improve grain quality. As the traditional QTL mapping methods (diploid methods) are usually designed for traits under diploid control, they are not the ideal approaches to map endosperm traits because they ignore the triploid nature of endosperm. In this article, a statistical method considering the triploid nature of endosperm (triploid method) is developed on the basis of multiple-interval mapping (MIM) to map for the underlying QTL. The proposed triploid MIM method is derived to broadly use the marker information either from only the maternal plants or from both the maternal plants and their embryos in the backcross and F_2 populations for mapping endosperm traits. Due to the use of multiple intervals simultaneously to take multiple QTL into account, the triploid MIM method can provide better detection power and estimation precision, and as shown in this article it is capable of analyzing and searching for epistatic QTL directly as compared to the traditional diploid methods and current triploid methods using only one (or two) interval(s). Several important issues in endosperm trait mapping, such as the relation and differences between the diploid and triploid methods, variance components of genetic variation, and the problems if effects are present and ignored, are also addressed. Simulations are performed to further explore these issues, to investigate the relative efficiency of different experimental designs, and to evaluate the performance of the proposed and current methods in mapping endosperm traits. The MIM-based triploid method can provide a powerful tool to estimate the genetic architecture of endosperm traits and to assist the marker-assisted selection for the improvement of grain quality in crop science. The triploid MIM FORTRAN program for mapping endosperm traits is available on the worldwide web (http://www.stat.sinica.edu.tw/chkao/).

CEREAL grains of many crops, such as rice, wheat, barley, and corn, are major food and nutritious resources for human, animal feeds, and industrial products. To enhance the yield and quality of grains, the understanding of the genetic basis underlying the cereal grains becomes increasingly important in crop study. The cereal grains are generally composed of diploid (embryo) and triploid (endosperm) tissues due to double fertilization. During the process of double fertilization, one of the two sperm cells fuses with the egg cell to produce a diploid zygote, which later divides mitotically to form the embryo, and the other sperm cell unites with the central cell (a diploid set of maternal chromosomes) to form a triploid endosperm nucleus, which also undergoes several mitotic divisions to become the endosperm. It is known that the endosperm plays a major role to nourish the embryo in the seed and the young seedling, and the content of endosperms, such as protein, sugar, oil, and carbohydrate concentration, showing quantitative variation is directly related to the quality of cereal grains. The genetic improvement targeting these endosperm traits can provide an efficient way to enhance the grain quality, and it has attracted a lot of attention in plant breeding (Sadimantara *et al.* 1997; Mazur *et al.* 1999; Tan *et al.* 1999; Wang and Larkins 2001; Lou and Zhu 2002). Genetically, the trisomic endosperm represents the next generation and has a more complex genetic mechanism than the diploid tissues. For these reasons, the approach of genetic analysis to endosperm traits is different from that to traits under diploid control, and special treatments are required in the study of endosperm traits.

Most endosperm traits show continuous variations. Quantitative genetic models considering the triploid nature of endosperm traits for studying the underlying genetic basis have been proposed by several researchers (Gale 1976; Mo 1987; Bogyo *et al.* 1988; Zhu and Weir 1994). These models generally focus on partitioning the phenotypic variance of an endosperm trait into various genetic and nongenetic (environmental) components. These variance components do not provide all the detailed information, such as the number, positions, and effects about the underlying quantitative trait loci (QTL).

[1] *Author e-mail:* chkao@stat.sinica.edu.tw

To unlock this QTL information, the ideas of the traditional QTL mapping methods utilizing the well-distributed genetic markers along the genome to infer the QTL parameters can be used. The traditional QTL mapping methods use the information about traits and markers from the same generation, *e.g.*, backcross or F_2 populations, to detect QTL controlling traits in diploid organisms (Lander and Botstein 1989; Haley and Knott 1992; Jansen 1993; Zeng 1994; Kao *et al.* 1999; Kao and Zeng 2002). Although they are designed for traits under diploid control, some researchers have applied them to mapping for QTL controlling endosperm traits (Tan *et al.* 1999; Wang and Larkins 2001; Wang *et al.* 2001). Such application implicitly relies on an invalid assumption that the endosperm traits are directly controlled by the diploid maternal genomes, not by the triploid endosperm genomes. Consequently, the traditional QTL mapping methods have limited power and precision in mapping endosperm traits (Wu *et al.* 2002a).

Wu *et al.* (2002a,b) and Xu *et al.* (2003) pioneered statistical methods to map endosperm traits by taking the triploid nature of endosperms into account using the marker information from the maternal plants (one-stage design) in the backcross or F_2 population. Wu *et al.* (2002a) further proposed a triploid QTL mapping method by using the marker information from both the maternal plants and their embryos (two-stage design), to improve the mapping of endosperm traits in the backcross population. Their methods have been shown to be able to provide improved QTL resolution. As these methods consider only one (or two) QTL at a time in the model, they can bias QTL identification and estimation when multiple QTL are located in the same linkage group (Lander and Botstein 1989; Jansen 1993; Zeng 1994). To deal with these problems and further improve the endosperm trait mapping, a potential way is to extend the current one-QTL model to a multiple-QTL model such that more genetic variation can be controlled in the model, as has been done in mapping traits in diploid tissues (Kao and Zeng 1997; Kao *et al.* 1999; Zeng *et al.* 1999). In this article, a triploid method based on multiple-interval mapping (MIM) using multiple marker intervals simultaneously to fit multiple putative QTL into the model is developed to achieve these purposes. This MIM-based triploid method can broadly take either the one- or two-stage design in either the backcross or F_2 population into account to analyze endosperm traits. As shown in this article, the proposed method can detect QTL responsible for endosperm traits with more power and better precision, and it can readily analyze and search for epistatic QTL due to its multiple-QTL approach. Besides, some related issues in mapping endosperm traits, such as the problems of using the diploid methods, the differences and relation between the diploid and triploid methods, the genetic variance components of endosperm traits, and the problems if QTL effects are present and ignored, are also investigated. A series of simulation studies was performed to further investigate these issues, to examine the relative efficiency of different experimental designs, and to evaluate the performance of the MIM-based method as compared to the current methods in mapping endosperm traits.

GENETIC MODEL OF ENDOSPERM TRAITS

Genetic model: For individuals in a backcross or F_2 population of autogamous plants, the endosperm tissues of their seeds can have four possible genotypes, *QQQ*, *QQq*, *Qqq*, and *qqq*, if only one QTL *Q* is considered (appendix b). Some genetic models for defining the genetic parameters and modeling the relationship between their genotypic values and the genetic parameters already exist (*e.g.*, Gale 1976; Mo 1987; Bogyo *et al.* 1988; Pooni *et al.* 1992; Zhu and Weir 1994). Here, the genetic model by Bogyo *et al.* is adopted for modeling, and it can be expressed in matrix notation as

$$\begin{bmatrix} G_1 \\ G_2 \\ G_3 \\ G_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} \frac{3}{2} & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \\ -\frac{3}{2} & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ d_1 \\ d_2 \end{bmatrix}, \qquad (1)$$

where the notations G_1, G_2, G_3, and G_4 denote the genotypic values of genotypes *QQQ*, *QQq*, *Qqq*, and *qqq*, respectively, and a, d_1, and d_2 are the genetic parameters. In Equation 1, the matrix with 4 × 3 dimension is called a genetic design matrix as it specifies the relationship between the genotypic values and genetic parameters, and it is symbolized by D. The unique solutions of a, d_1, and d_2 in terms of the genotypic values are

$$\mu = \frac{G_1}{2} + \frac{G_4}{2},$$

$$a = \frac{G_1}{3} - \frac{G_4}{3},$$

$$d_1 = -\frac{2G_1}{3} + G_2 - \frac{G_4}{3} = a - (G_1 - G_2),$$

$$d_2 = -\frac{G_1}{3} + G_3 - \frac{2G_4}{3} = (G_3 - G_4) - a.$$

The parameter μ obviously is not a measure of mean genotypic values as the genotypic values of *AAa* and *Aaa* are ignored. The parameter a, which measures the average effect of substituting Q for q, is defined as the additive effect, and the parameter d_1 (d_2), which measures the departure of the substitution effect in *QQ* (*qq*) background, is defined as the first (second) dominance effect. The genetic model can be expressed more succinctly as

$$G_i = \mu + ax + d_1 z_1 + d_2 z_2, \quad i = 1, 2, 3, 4, \qquad (2)$$

where the coded variables are defined as

$$x = \begin{cases} ^{3}/_{2} & \text{if } Q \text{ is } QQQ \\ ^{1}/_{2} & \text{if } Q \text{ is } QQq \\ -^{1}/_{2} & \text{if } Q \text{ is } Qqq \\ -^{3}/_{2} & \text{if } Q \text{ is } qqq, \end{cases} \qquad z_1 = \begin{cases} 1 & \text{if } Q \text{ is } QQq \\ 0 & \text{otherwise}, \end{cases}$$

$$z_2 = \begin{cases} 1 & \text{if } Q \text{ is } Qqq \\ 0 & \text{otherwise}, \end{cases}$$

such that each genotype corresponds to its genotypic value. If different genetic models are used for modeling, they can be also expressed as in Equations 1 and 2, but note that the parameters may have different meanings and the variance component may have different structure.

The extension of the one-locus genetic model in Equation 1 to multiple, say m, loci is straightforward. Consider m QTL, Q_1, Q_2, . . . , and Q_m, each with four genotypes and three genetic parameters. Together, for m QTL, there are 4^m possible different QTL genotypes and $3m$ parameters if epistasis between QTL is not considered or $3m(3m - 1)/2$ parameters if only up to digenic epistasis is considered. The columns for epistasis can easily be obtained from the product of columns of marginal effects. By expanding the genetic design matrix D of Equation 1 to a $4^m \times 3m$ or $4^m \times 3m(3m - 1)/2$ matrix (see THE MIM MODEL FOR MAPPING ENDOSPERM TRAITS), the genetic model for m QTL in matrix notation can be obtained. The genetic design matrix D plays an important role in the estimation of the QTL effects in the triploid MIM model. The corresponding multiple-QTL model in the form of Equation 2 can be easily obtained using a regression principle. Following the regression principle, the genetic model of m QTL by considering up to digenic epistasis can be written as

$$\begin{aligned} G_i = \mu &+ \sum_{j=1}^{m} a_j x_j + \sum_{j=1}^{m} d_{j1} z_{j1} + \sum_{j=1}^{m} d_{j2} z_{j2} \\ &+ \sum_{j<k}^{m} i_{a_j a_k}(x_j x_k) + \sum_{j \neq k}^{m} i_{a_j d_{k1}}(x_j z_{k1}) + \sum_{j \neq k}^{m} i_{a_j d_{k2}}(x_j z_{k2}) \\ &+ \sum_{j<k}^{m} i_{d_{j1} d_{k1}}(z_{j1} z_{k1}) + \sum_{j<k}^{m} i_{d_{j1} d_{k2}}(z_{j1} z_{k2}) + \sum_{j<k}^{m} i_{d_{j2} d_{k1}}(z_{j2} z_{k1}) \\ &+ \sum_{j<k}^{m} i_{d_{j2} d_{k2}}(z_{j2} z_{k2}), \quad i = 1, 2, \ldots, 4^m, \end{aligned} \qquad (3)$$

where μ is the intercept; a_j, d_{j1}, and d_{j2} are the additive and dominance effects of Q_j; $i_{a_j a_k}$, $i_{a_j d_{k1}}$, $i_{a_j d_{k2}}$, $i_{d_{j1} d_{k1}}$, $i_{d_{j1} d_{k2}}$, $i_{d_{j2} d_{k1}}$, and $i_{d_{j2} d_{k2}}$ denote the epistatic effects between QTL; and x_j, z_{j1}, and z_{j2} are the coded variables of the additive and dominance effects for Q_j.

Variance components: Consider only one QTL in the genetic model. It is easy to show that the variance of the additive variable, $V(x)$, is $^{19}/_{16}$, and the variances of the two dominance variables, $V(z_1)$ and $V(z_2)$, are $^{7}/_{64}$, in a backcross population. In an F_2 population, these variances are $^{7}/_{4}$, $^{7}/_{64}$, and $^{7}/_{64}$, respectively. The covariances between the variables, $\mathrm{Cov}(x, z_1)$, $\mathrm{Cov}(x, z_2)$, and $\mathrm{Cov}(z_1, z_2)$, are $^{5}/_{32}$, $^{1}/_{32}$, and $-^{1}/_{64}$, respectively, in the backcross population, and they are $^{1}/_{16}$, $-^{1}/_{16}$, and $-^{1}/_{64}$, respectively, in the F_2 population. Therefore, the genetic variance components of an endosperm trait are

$$\sigma_G^2 = \frac{19}{16}a^2 + \frac{7}{64}d_1^2 + \frac{7}{64}d_2^2 + \frac{5}{16}ad_1 + \frac{1}{16}ad_2 - \frac{1}{32}d_1 d_2 \qquad (4)$$

in the backcross population, and they are

$$\sigma_G^2 = \frac{7}{4}a^2 + \frac{7}{64}d_1^2 + \frac{7}{64}d_2^2 + \frac{1}{8}ad_1 - \frac{1}{8}ad_2 - \frac{1}{32}d_1 d_2 \qquad (5)$$

in the F_2 population. It shows that each effect contributes not only to its variance but also to the covariances with other effects, and that the relative importance of effects in contributing to the total genetic variance depends not only on their sizes but also on their associated coefficients (the variance or covariance of their coded variables). When m QTL each with complete effects are considered together, the genetic variance has $[9m^2(3m - 1)^2 + 6m(3m - 1)]/8$ components. For example, the total genetic variance has 120 components for $m = 2$ in both populations (not shown), and it reduces to 111 components in the backcross population and 83 components in the F_2 population when the two QTL are unlinked (APPENDIX A). Among the coefficients of the variances involving the epistatic effects, the coefficients associated with the additive-by-additive effect ($i_{a_1 a_2}$) are relatively much larger than those of other variances and covariances. For example, in the F_2 population, the coefficient of $i_{a_1 a_2}^2$ (the variance of $x_1 x_2$) is $^{49}/_{16}$ (APPENDIX A); *i.e.*, the variance contributed by $i_{a_1 a_2}$ is $^{49}/_{16} \times i_{a_1 a_2}^2$, the coefficients of the other four epistatic variances involving the additive effects are $^{7}/_{32}$, and the coefficients of the remaining four different types of epistatic variance are $^{63}/_{4096}$. The coefficients of the covariances between the additive effects and the epistatic effects involving the additive effects are $^{7}/_{32}$, and the coefficients of the covariances between $i_{a_1 a_2}$ and the other epistatic effects involving the additive effects are $^{7}/_{64}$. The other covariances are relatively smaller. Therefore, it implies that, for the same order of the epistatic effects, the epistatic effects involving the additive effects, especially the additive-by-additive effect, are relatively easy to detect, and the other epistatic effects are relatively difficult to detect in practical QTL mapping (with a limited sample size). A similar pattern can also be found in the backcross population. For two nonepistatic QTL, the variance components are

$$
\begin{aligned}
\sigma_G^2 = {} & \frac{7}{4}a_1^2 + \frac{7}{64}d_{11}^2 + \frac{7}{64}d_{12}^2 + \frac{7}{4}a_2^2 + \frac{7}{64}d_{21}^2 + \frac{7}{64}d_{22}^2 \\
& + \frac{1}{8}a_1d_{11} - \frac{1}{8}a_1d_{12} - \frac{1}{32}d_{11}d_{12} + \frac{1}{8}a_2d_{21} - \frac{1}{8}a_2d_{22} - \frac{1}{32}d_{21}d_{22} \\
& + \frac{1}{4}[9(1-2r_{12}) + 5(1-2r_{12})^2]a_1a_2 + \frac{1}{8}(1-2r_{12})^2a_1d_{21} \\
& - \frac{1}{8}(1-2r_{12})^2a_1d_{22} + \frac{1}{8}(1-2r_{12})^2d_{11}a_2 \\
& + \left\{\frac{1}{4}[r_{12}^4 + (1-r_{12})^4] - \frac{1}{32}\right\}d_{11}d_{21} \\
& + \left\{\frac{1}{2}[r_{12}(1-r_{12})]^2 - \frac{1}{32}\right\}d_{11}d_{22} \\
& - \frac{1}{8}(1-2r_{12})^2a_2d_{12} + \left\{\frac{1}{2}[r_{12}(1-r_{12})]^2 - \frac{1}{32}\right\}d_{12}d_{21} \\
& + \left\{\frac{1}{4}[r_{12}^4 + (1-r_{12})^4] - \frac{1}{32}\right\}d_{12}d_{22},
\end{aligned}
$$

where r_{12} is the recombination fraction between the two QTL, in the F_2 population. Similarly, the variance components for the backcross population also have 21 terms (not shown). If the two nonepistatic QTL are unlinked, the variance components reduce to a much simpler form with the first 12 components.

THE RELATION BETWEEN THE DIPLOID AND TRIPLOID METHODS

The traditional QTL mapping methods are usually designed to map for QTL controlling traits in diploid organisms (Lander and Botstein 1989; Haley and Knott 1992; Jansen 1993; Zeng 1994; Kao *et al.* 1999; Kao and Zeng 2002). These diploid methods classify the genotypes of each QTL into two groups, *QQ* (*qq*) and *Qq*, for the backcross population or three groups, *qq*, *Qq*, and *QQ*, for the F_2 population, and they detect the association between the QTL genotype and the trait value both measured at the same generation for QTL mapping. Although the endosperms are known to be triploid and represent the next generation, some researchers have applied these diploid methods to mapping endosperm traits of the backcross or F_2 individuals (Tan *et al.* 1999; Wang *et al.* 2001; Wu *et al.* 2002a). Therefore, it is important to investigate the problems of using the diploid methods and the relation between the diploid and triploid methods in mapping endosperm traits.

Diploid methods: When applying the diploid methods to mapping endosperm traits and only one QTL is considered, the statistical model for n endosperms in the backcross population can be written as

$$y_i = \mu + bw_i^* + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{6}$$

where w_i^* is coded as

$$w_i^* = \begin{cases} \tfrac{1}{2} & \text{if } Q \text{ is } Qq, \\ -\tfrac{1}{2} & \text{if } Q \text{ is } qq; \end{cases}$$

y_i is the endosperm trait value; μ is the intercept; b is the QTL effect measuring the genotypic difference between *Qq* and *qq*. The statistical model for n endosperms in the F_2 population can be expressed as

$$y_i = \mu + b_a w_{ai}^* + b_d w_{di}^* + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{7}$$

where w_{ai}^* and w_{di}^* are defined as

$$w_{ai}^* = \begin{cases} 1 & \text{if } Q \text{ is } QQ, \\ 0 & \text{if } Q \text{ is } Qq, \\ -1 & \text{if } Q \text{ is } qq, \end{cases} \qquad w_{di}^* = \begin{cases} \tfrac{1}{2} & \text{if } Q \text{ is } Qq, \\ -\tfrac{1}{2} & \text{otherwise;} \end{cases}$$

b_a and b_d denote the additive and dominance effects. The residual error ε_i in the above two models is assumed to have a normal distribution with mean zero and variance σ_ε^2. As QTL may not be coincident with markers, the QTL genotype is usually unobservable. Therefore, the likelihood of the diploid model is known as a mixture of normals,

$$L(\theta|\mathbf{Y}, \mathbf{X}) = \prod_{i=1}^{n}\left[\sum_{j=1}^{k} p_{ij}N(\mu_j, \sigma_\varepsilon^2)\right], \tag{8}$$

where μ_j's correspond to the genotypic values of the k different QTL genotypes ($k = 2$ for the backcross model and $k = 3$ for the F_2 model), and the mixing proportions, p_{ij}'s, are the conditional probabilities of QTL genotypes (see Tables 1 and 2 in Kao and Zeng 1997). The maximum-likelihood estimate (MLE) of the QTL effects and their asymptotic variance-covariance can be obtained using the EM algorithm (Dempster *et al.* 1977) and Louis's (1982) method by treating the normal mixture model as an incomplete-data problem.

The relation between the diploid and triploid models: When applying the diploid models to mapping endosperm traits, it is generally assumed that the endosperm traits are directly controlled by the diploid genomes of the backcross or F_2 individuals. This assumption, however, violates the fact that the triploid endosperms represent the genetic composition of the next generation, which, in fact, is mainly responsible for the trait variation. Consequently, as compared to the use of the triploid model, some problems, such as less power and precision in QTL detection, will occur in the diploid model as shown below.

When an endosperm trait affected only by one QTL, Q, is regressed on a marker M along the genome to infer Q, the regression coefficient of M in the backcross diploid model is

$$b_M = (1 - 2r_{QM})\left[\frac{3}{2}a + \frac{1}{4}(d_1 + d_2)\right], \tag{9}$$

where r_{QM} is the recombination fraction between M and

Q, in the backcross population (APPENDIX B). If the marker M is coincident with $Q(r_{QM} = 0)$, the coefficient reduces to $b_M = {}^3\!/_2 a + {}^1\!/_4(d_1 + d_2)$. The estimated coefficient of the backcross diploid model is composed of the additive effect and two dominance effects. In the F_2 diploid model, the regression coefficient for the additive effect of M is

$$b_{M_a} = \frac{3(1 - 2r_{QM})}{2}a, \tag{10}$$

and the coefficient for the dominance effect is

$$b_{M_d} = \frac{1 - 2r_{QM}}{4}(d_1 + d_2). \tag{11}$$

If M and Q are coincident, the additive coefficient reduces to $b_{M_a} = 3a/2$ and the dominance coefficient reduces to $b_{M_a} = (d_1 + d_2)/4$. The additive coefficient estimated in the F_2 diploid model is 1.5 times the additive effect, and the estimated dominance coefficient is one-quarter of the sum of the two dominance effects. When both of the additive and dominance variables are fitted in the model, the partial regression coefficients are the same as Equations 10 and 11 because of orthogonality. The above derivations present the relation of parameters between the diploid and triploid models and show that the diploid models cannot directly estimate the QTL effects in mapping endosperm traits.

The phenotypic variance conditional on the marker M in the backcross diploid model is

$$\sigma^2_{y.M} = \sigma^2 + (1 - 2r_{QM})^2 \times \left[\frac{5}{8}a^2 + \frac{3}{32}(d_1^2 + d_2^2) + \frac{1}{8}(ad_1) - \frac{1}{8}(ad_2) - \frac{1}{16}(d_1d_2)\right] \tag{12}$$

(APPENDIX B). It shows that the genetic variances and covariances contributed by the additive and dominance effects cannot be fully controlled in the model. The percentages of additive and dominance variances uncontrolled by the diploid model are ~47.4% (9/19) and 14.3% (1/7), respectively. For the F_2 population, the phenotypic variance conditional on the additive and dominance variables of marker M is the same as that in the backcross model (APPENDIX B). The percentages of uncontrolled additive and dominance variances are ~63.4% (9/14) and 14.3% (1/7), respectively. In addition, a part of the genetic covariances is also uncontrolled by the diploid model. The uncontrolled variances and covariances will become a part of the genetic residual, causing inflation of the sampling variance of the coefficients. The sampling variance of the regression coefficient of the backcross model is $\sim V(\hat{b}_M) = n^{-1} \times \sigma^2_{y.M}/\sigma^2_M$, where σ^2_M is the variance of the coded variable of M, in a large sample with size n (STUART and ORD 1991). Using the approximation, the sampling variances of the regression coefficients between the diploid and triploid models can be compared when M and Q are coincident ($r_{QM} = 0$). Taking a QTL with no dominance ($a = 1$, $d_1 = d_2 = 0$) and contributing 10% of the trait variation as an example, the conditional phenotypic variance roughly equals to σ^2 for the triploid model, and it is $\sim{}^{181}\!/_{171} \times \sigma^2$ for the diploid model. The variances σ^2_M for the two different models are ${}^1\!/_4$ and ${}^{19}\!/_{16}$, respectively. Consequently, the sampling variance of the regression coefficient for the diploid model is ~5.03 times that for the triploid model in the backcross population. It is ~3.64 times that for the same setting in the F_2 population. The sampling variances of the regression coefficients in the diploid models are larger than those in the triploid model.

On the basis of the above findings, two problems will occur if the diploid models are applied to mapping endosperm traits. First, the estimates in the diploid models are generally confounded by the additive and dominance effects of endosperm QTL (Equations 9–11). Second, the sampling variances of the estimates will inflate because the genetic variances and covariances contributed by QTL are not fully controlled in the model. Consequently, the diploid models cannot directly estimate the effects of the endosperm QTL, and they have the confounding problems in estimation and will decrease the power in endosperm QTL detection.

THE MIM MODEL FOR MAPPING ENDOSPERM TRAITS

Endosperm trait multiple-interval mapping: Assume an endosperm trait is controlled by m QTL, Q_1, Q_2, . . . , and Q_m, located at positions p_1, p_2, . . . , and p_m, in m different marker intervals, I_1, I_2, . . . , and I_m, along the genome. If only up to digenic epistasis is considered, the value of an endosperm trait, y_i, in the backcross or F_2 population can be related to the m putative QTL by the model

$$\begin{aligned} y_i = \mu &+ \sum_{j=1}^{m} a_j x^*_{ij} + \sum_{j=1}^{m} d_{j1} z^*_{ij1} + \sum_{j=1}^{m} d_{j2} z^*_{ij2} + \sum_{j<k}^{m} i_{a_j a_k}(x^*_{ij} x^*_{ik}) \\ &+ \sum_{j\neq k}^{m} i_{a_j d_{k1}}(x^*_{ij} z^*_{ik1}) + \sum_{j\neq k}^{m} i_{a_j d_{k2}}(x^*_{ij} z^*_{ik2}) + \sum_{j<k}^{m} i_{d_{j1} d_{k1}}(z^*_{ij1} z^*_{ik1}) \\ &+ \sum_{j<k}^{m} i_{d_{j1} d_{k2}}(z^*_{ij1} z^*_{ik2}) + \sum_{j<k}^{m} i_{d_{j2} d_{k1}}(z^*_{ij2} z^*_{ik1}) + \sum_{j<k}^{m} i_{d_{j2} d_{k2}}(z^*_{ij2} z^*_{ik2}) \\ &+ \varepsilon_i, \quad i = 1, 2, \ldots, n, \end{aligned} \tag{13}$$

where the parameters and coded variables have the same definitions as those in the genetic model in Equation 3, and the residual error ε_i is assumed to follow normal distribution with mean zero and variance σ^2. In QTL mapping, the endosperm QTL genotype of any putative QTL, say Q_j, $j = 1, 2, \ldots, m$, is usually not observable and could be $Q_jQ_jQ_j$, $Q_jQ_jq_j$, $Q_jq_jq_j$, or $q_jq_jq_j$ with different (conditional) probabilities for different endosperm i. The conditional probabilities (distribution) for each Q_j under different experimental designs can be derived by using its flanking marker information from the maternal plants (and their embryos) as shown below, and then the normal mixture likelihood of the model can be constructed. As multiple (m) intervals are used to infer the conditional distribution of the (m) endosperm QTL for modeling, this approach is called multiple-interval mapping in QTL mapping (Kao and Zeng 1997; Kao *et al.* 1999), and this model is a MIM-based triploid model. By specifying appropriate conditional probabilities to the 4^m endosperm QTL genotypes of the m QTL, this triploid MIM model can be applied widely to mapping endosperm traits using data from various designs and populations.

Likelihood: For any interval, I_j, flanked by the two markers, M_j and N_j, the maternal plants or their embryos can have four and nine different marker genotypes in the backcross and F_2 populations, respectively. If both the plants and embryos are considered together, their marker genotypes can have 16 and 25 combinations in the two different populations, respectively (appendix c). For any Q_j in I_j, the (conditional) probabilities of the four endosperm QTL genotypes can be inferred only from the maternal plants (one-stage design) or both from the maternal plants and their embryos (two-stage design) as shown in appendix c. To assist with explaining the parameter estimation, these conditional probabilities are extracted to form a matrix $\mathbf{Q}_j$, $j = 1, 2, \ldots, m$. The dimension of $\mathbf{Q}_j$ is 25×4 (16×4) for a two-stage design in the F_2 (backcross) population, it is 9×4 (4×4) for a one-stage design in the F_2 (backcross) population (note that Q denotes QTL, and $\mathbf{Q}$ denotes the conditional probability matrix). For the total m QTL in the m different intervals, there are 4^m possible endosperm QTL genotypes in each of 25^m (16^m, 9^m, or 4^m) possible marker genotypes. The 4^m joint conditional probabilities of endosperm QTL genotypes can be obtained by the product of individual conditional probabilities for each QTL using the property of conditional independence among different QTL (Kao and Zeng 1997), and they play the role of mixing proportions in the normal mixture likelihood. Let the conditional probabilities of 4^m possible QTL genotypes for endosperm i from designs and populations be denoted as p_{ij}'s, $j = 1, 2, \ldots, 4^m$ (note that p_j's denote QTL positions, and p_{ij}'s denote the conditional probabilities). The likelihood of the triploid MIM model for the n endosperms is a mixture of 4^m normals as

$$L(\theta|\mathbf{Y}, \mathbf{X}) = \prod_{i=1}^{n}\left[\sum_{j=1}^{4^m} p_{ij}N(\mu_j, \sigma^2)\right], \tag{14}$$

where μ_j's correspond to the genotypic values of the 4^m different QTL genotypes, and the mixing proportions, p_{ij}'s, are the corresponding joint conditional probabilities. The density of each individual i is a mixture of 4^m possible normal densities with different means, μ_j's, and mixing proportions, p_{ij}'s. The general formulas proposed by Kao and Zeng (1997) are used to obtain the MLE of the effects and their asymptotic variance-covariance matrix.

Parameter estimation: The application of the general formulas to obtain the MLE and the asymptotic variance-covariance matrix for the triploid MIM model is based on the construction of the two matrices D and Q, where D is the genetic design matrix for characterizing the QTL effects, and Q is the conditional probability matrix containing the mixing proportions of QTL genotypes. Given the two matrices, the MLE of QTL effects and their asymptotic variance-covariance matrix of the triploid model can be easily obtained. The construction of the D and Q matrices is described below.

For one QTL ($m = 1$) in the model, there are four endosperm QTL genotypes and three genetic effects, and the genetic design matrix is a 4×3 matrix as shown in Equation 1. For m QTL in the model, if epistasis between QTL effects is not considered, there are 4^m endosperm QTL genotypes and $3m$ genetic effects (m additive effects, m first dominance effects, and m second dominance effects), and the genetic design matrix is then a $4^m \times 3m$ matrix. If all the possible digenic epistases between QTL are considered, the column dimension of D becomes $3m(3m - 1)/2$. An example of genetic design matrix with $m = 2$ and all possible effects (with dimension 16×15) can be found in Wu *et al.* (2002a). The joint conditional probability matrix $\mathbf{Q}$ for the m QTL has a dimension $9^m \times 4^m$ ($4^m \times 4^m$) or $25^m \times 4^m$ ($16^m \times 4^m$) under the one- or two-stage design in the F_2 (backcross) population, and they can be obtained by $\mathbf{Q} = \mathbf{Q}_1 \otimes \mathbf{Q}_2 \otimes \ldots \otimes \mathbf{Q}_m$, where $\otimes$ denotes the Kronecker product. The 4^m mixing proportions of any endosperm i, p_{ij}'s, in the likelihood can be found to be one of the rows in $\mathbf{Q}$ according to the marker genotypes of the plants (and embryos). By applying the matrices $\mathbf{D}$ and $\mathbf{Q}$ to the general formulas, the MLE of the effects and their asymptotic variance-covariance matrix can be readily obtained.

The problems if effects are present and ignored: Three marginal genetic effects are associated with each endosperm QTL. In practice, QTL may display all or some of the effects (see Wu *et al.* 2002b as an example), and, before mapping, it is not known which effects are present or absent. The possible drawback of fitting the absent effects (overfitting) in the model is the loss of power in QTL detection, as higher critical value is usually required to claim the significance of QTL. If some displayed effects are ignored in the model, not only the power of detection will be affected but also the confounding problem will occur as discussed below.

Assume the endosperm trait value y is affected by two nonepistatic endosperm QTL, Q_1 and Q_2. When the trait value is regressed on Q_1 by fitting only its additive variable x_1 into the model, the regression coefficients in terms of the QTL effects and linkage parameter for the backcross and F_2 populations are shown in appendix d. It shows that the estimate of the additive effect of Q_1

is not unbiased for a_1 and is confounded by its other effects and the effects of Q_2. The confounding of Q_2 effects is through linkage parameter. If Q_1 and Q_2 are unlinked, the regression coefficients reduce to much simpler forms without the confounding of Q_2. For example, if $r_{12} = 0.5$, $b_{yx_1} = a_1 + 5d_{11}/38 + d_{12}/38$ for the backcross population, and $b_{yx_1} = a_1 + d_{11}/28 - d_{12}/28$ for the F_2 population. The confounding of Q_2 disappears, and the coefficient is confounded only by its dominance effects. The same confounding problem can also be found for the estimate of the dominance effect if fitting only its dominance variable z_1 in the model (APPENDIX D). If epistasis is present and ignored in the model, most of the epistatic effects will be confounded in the estimation as most of the covariances between the marginal and epistatic effects are not zero whether they are linked or not (result not shown). To avoid the confounding problem and enhance the detection power, it is desirable to fit only those displayed effects into the model in QTL mapping.

SIMULATION STUDY

A series of simulations was performed to achieve three purposes: (1) to verify the derived relations and compare the differences between the diploid and triploid models, (2) to examine the performance of the triploid method in different experimental designs and populations, and (3) to evaluate the performance of the proposed MIM-based triploid method as compared to the current methods in mapping endosperm traits. The simulation study includes two parts. The first part is to achieve the first two purposes, and the second part is to achieve the third purpose. In each part, the sample size is assumed to be 200. The first part assumes one QTL affecting the endosperm trait with two levels of heritability (h^2), 0.1 and 0.2. It includes four different parameter settings: (1) $a = 1$, $d_1 = -2$, $d_2 = -2$ ($G_1 = 3/2$, $G_2 = -3/2$, $G_3 = -5/2$, and $G_4 = -3/2$); (2) $a = 1$, $d_1 = 2$, $d_2 = 2$ ($G_1 = 3/2$, $G_2 = 5/2$, $G_3 = 3/2$, and $G_4 = -3/2$); (3) $a = 1$, $d_1 = -2$, $d_2 = 0$; ($G_1 = 3/2$, $G_2 = -3/2$, $G_3 = -1/2$, and $G_4 = -3/2$); and (4) $a = 1$, $d_1 = 0$, $d_2 = 0$ ($G_1 = 3/2$, $G_2 = 1/2$, $G_3 = -1/2$, and $G_4 = -3/2$). Among the four settings, the QTL genotypes are complete-recessive type in the first and third settings, and they are complete-dominance type in the second setting. For each setting, the QTL is placed in the middle of a chromosome with six 20-cM equally spaced markers, and data from both the one- and two-stage designs in the backcross and F_2 populations were generated. The number of simulation replicates is 500. Both the diploid and triploid methods were used to detect the QTL using the generated data sets. The results are shown in Tables 1 and 2. The second part assumes three chromosomes each with six 20-cM equally spaced markers, and each chromosome contains only one QTL. The three unlinked QTL, Q_A, Q_B, and Q_C, are assumed to contribute 40% to the total trait variation together and to be located in the middle of the chromosomes. Data from the two-stage design in the F_2 population were generated. The parameter setting is $a_1 = 3$, $d_{11} = -3$, and $d_{12} = -3$ for Q_A; $a_2 = 2.5$, $d_{21} = 4$, and $d_{22} = 4$ for Q_B; and $a_3 = 1.5$, $d_{31} = 0$, and $d_{32} = 0$ for Q_C. There is additive-by-additive interaction between Q_B and Q_C, and the epistatic effect $i_{a_2a_3}$ is assumed to be 1. Under the parameter setting, the genetic and environmental variances are ~38.37 and 51.66, respectively. In the total genetic variance, the marginal effects of the three QTL contribute ~45.44, 36.32, and 10.26%, respectively, and the epistatic effect contributes ~7.98%. In the genetic variance contributed by Q_A (Q_B), the variance contributed by the two dominance effects is ~11.29% (25.11%). The number of simulation replicates is 100. Both the current triploid method considering only one QTL, *i.e.*, the interval-mapping (IM)-based method, and the proposed MIM-based method were used to analyze the data. The results are shown in Table 3. In each scenario, permutation tests proposed by CHURCHILL and DOERGE (1994) were used to determine the critical values for power calculation.

Tables 1 and 2 show the results of the first part of the simulation. The relationship between the estimates of the diploid and triploid models corresponds very well with the derived prediction (Equations 9–11). For the backcross population, the effects of the diploid models in the four settings are expected to be 0.5, 2.5, 1.0, and 1.5, according to Equation 9. The means of the estimates are found to be 0.610, 2.516, 1.040, and 1.521, respectively, for $h^2 = 0.1$ (Table 2), and they are 0.599, 2.489, 1.005, and 1.475, respectively, for $h^2 = 0.2$ (Table 3). For the F_2 population, the means of the estimated additive and dominance effects in the diploid model are also found to be very close to the predicted values in both levels of heritability. For example, the mean of the estimated additive effects for the first setting with $h^2 = 0.1$ is 1.499 (predicted value 1.5), and the mean of the estimated dominance effects for the second setting with $h^2 = 0.2$ is 1.010 (predicted value 1.0). The estimated residual variance by the diploid model is found to be upwardly biased in all cases as expected by Equation 12.

The most striking differences in power and estimation between the diploid and triploid models are found in the first parameter setting when the additive and dominance effects are in the opposite direction and $h^2 = 0.2$ (Table 2). The detecting powers of the diploid model are 0.160 and 0.100, respectively, in the two different populations. The detecting powers of the triploid model are 0.508 and 0.926, respectively, under the one-stage design, and they increase to 0.980 and 0.998, respectively, under the two-stage design. For QTL position, the means of position estimates by the diploid model are 46.46 (SD 28.58) and 49.63 (SD 11.10), respectively, in the two populations. The means of position estimates provided by the triploid model under the two-stage design are 49.77 (SD 7.08) and 50.21 (SD 5.68), respectively,

TABLE 1

Simulation results of using diploid and triploid methods under different experimental designs and parameter settings with heritability 0.1

	Backcross population							F_2 population						
Method	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = -2$	$d_2 = -2$	$\sigma^2 = 10.69$	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = -2$	$d_2 = -2$	$\sigma^2 = 22.5$
Diploid	0.116	45.30 (31.46)	−1.255 (0.249)	0.610 (0.742)	—	—	11.58 (1.17)	0.626	50.96 (18.21)	−0.508 (0.354)	1.499 (0.646)	−0.886 (1.137)		23.36 (2.42)
Triploid (one stage)	0.190	47.48 (29.98)	−1.231 (0.566)	0.199 (0.452)	1.587 (2.342)	−0.589 (2.202)	9.89 (1.42)	0.596	51.98 (18.36)	−0.102 (0.648)	0.994 (0.430)	−0.625 (3.056)	−2.501 (2.465)	20.65 (2.91)
Triploid (two stage)	0.730	49.39 (15.10)	−0.067 (0.469)	0.962 (0.324)	−1.479 (1.501)	−2.244 (1.260)	10.20 (1.19)	0.796	50.58 (12.84)	−0.027 (0.453)	1.010 (0.357)	−1.239 (1.866)	−2.524 (1.745)	22.11 (3.62)
Method	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = 2$	$d_2 = 2$	$\sigma^2 = 24.19$	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = 2$	$d_2 = 2$	$\sigma^2 = 22.5$
Diploid	0.774	49.26 (16.36)	−0.252 (0.718)	2.516 (0.718)	—	—	24.72 (2.60)	0.668	49.66 (19.16)	0.495 (0.353)	1.504 (0.637)	—	1.018 (1.056)	23.37 (2.38)
Triploid (one stage)	0.748	48.34 (17.59)	−0.001 (0.654)	1.006 (0.445)	2.814 (1.990)	1.010 (2.439)	22.46 (2.84)	0.648	49.85 (19.05)	0.038 (0.615)	0.985 (0.426)	2.793 (2.352)	0.895 (3.045)	20.54 (2.94)
Triploid (two stage)	0.850	50.00 (13.04)	0.007 (0.673)	1.001 (0.452)	2.677 (1.903)	1.622 (2.032)	22.92 (2.66)	0.796	49.86 (12.35)	0.021 (0.460)	1.013 (0.358)	2.530 (1.724)	1.114 (1.911)	21.53 (2.54)
Method	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = -2$	$d_2 = 0$	$\sigma^2 = 9.00$	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = -2$	$d_2 = 0$	$\sigma^2 = 17.44$
Diploid	0.400	49.03 (23.68)	−1.005 (0.232)	1.040 (0.528)	—	—	9.49 (1.00)	0.692	48.74 (17.28)	−0.261 (0.312)	1.518 (0.514)	−0.503 (0.847)	—	17.77 (1.75)
Triploid (one stage)	0.406	50.33 (23.67)	−1.035 (0.591)	0.306 (0.465)	2.077 (2.176)	—	8.47 (1.09)	0.758	49.09 (17.26)	−0.057 (0.499)	1.017 (0.344)	−1.487 (3.173)	—	16.16 (2.02)
Triploid (two stage)	0.808	49.15 (12.38)	−0.047 (0.407)	0.966 (0.260)	−1.797 (1.604)	—	8.64 (0.95)	0.872	50.00 (11.55)	−0.033 (0.394)	1.015 (0.281)	−1.726 (1.995)	—	16.83 (1.78)
Method	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = 0$	$d_2 = 0$	$\sigma^2 = 10.69$	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = 0$	$d_2 = 0$	$\sigma^2 = 15.75$
Diploid	0.674	50.23 (18.89)	−0.752 (0.246)	1.521 (0.493)	—	—	11.03 (1.17)	0.840	50.87 (15.25)	−0.005 (0.292)	1.541 (0.467)	—	—	16.37 (1.70)
Triploid (one stage)	0.678	51.00 (18.43)	0.008 (0.345)	1.015 (0.320)	—	—	10.33 (1.19)	0.842	51.42 (15.05)	−0.007 (0.293)	1.040 (0.314)	—	—	15.60 (1.74)
Triploid (two stage)	0.950	49.85 (11.71)	0.020 (0.298)	1.001 (0.232)	—	—	10.43 (1.13)	0.918	50.39 (11.96)	−0.008 (0.288)	1.020 (0.290)	—	—	15.71 (1.64)

For each parameter setting, 500 replicates, each with sample size 200, were analyzed with one QTL controlling the trait variation and located in the middle of a chromosome. Permutation tests were used to determine the critical value at 0.05 significance level in each parameter setting. The critical values for the diploid methods in the backcross (F_2) population are 6.989 (9.815), 6.570 (9.772), 6.770 (9.809), and 6.678 (6.423), respectively, in the four parameter settings. The critical values for the triploid methods under the one-stage design in the backcross (F_2) population are 9.096 (11.515), 8.667 (10.756), 10.577 (9.252), and 6.759 (6.573), respectively. The critical values for the triploid methods under the two-stage design in the backcross (F_2) population are 11.399 (11.955), 11.465 (11.300), 11.485 (9.893), and 6.901 (7.210), respectively. The numbers in the parentheses denote standard deviations. The data from the one-stage and two-stage designs were analyzed by the triploid method. Posi, position.

TABLE 2

Simulation results of using the diploid and triploid methods under different experimental designs and parameter settings with heritability 0.2

	Backcross population							F_2 population						
Method	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = -2$	$d_2 = -2$	$\sigma^2 = 4.75$	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = -2$	$d_2 = -2$	$\sigma^2 = 10$
Diploid	0.160	46.46 (28.58)	−1.254 (0.177)	0.599 (0.460)	—	—	5.75 (0.59)	0.100	49.63 (11.10)	−0.506 (0.246)	1.492 (0.398)	−0.941 (0.690)		11.07 (1.17)
Triploid (one stage)	0.508	49.00 (24.26)	−1.312 (0.445)	0.143 (0.344)	2.117 (1.628)	−0.766 (1.547)	4.50 (0.66)	0.926	51.03 (11.19)	−0.057 (0.395)	0.993 (0.273)	−1.176 (2.035)	−2.212 (1.427)	9.37 (1.32)
Triploid (two stage)	0.980	49.77 (7.08)	−0.037 (0.295)	0.980 (0.200)	−1.781 (0.889)	−2.063 (0.762)	4.60 (0.54)	0.988	50.21 (5.68)	−0.015 (0.294)	1.012 (0.223)	−1.604 (1.165)	−2.211 (1.100)	9.65 (1.17)
Method	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = 2$	$d_2 = 2$	$\sigma^2 = 10.75$	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = 2$	$d_1 = 2$	$\sigma^2 = 10$
Diploid	0.976	48.33 (10.45)	−0.257 (0.251)	2.489 (0.522)	—	—	11.77 (1.22)	0.928	49.34 (11.21)	0.500 (0.242)	1.494 (0.391)	1.010 (0.622)	—	11.07 (1.15)
Triploid (one stage)	0.972	49.20 (10.68)	−0.038 (0.459)	0.970 (0.302)	2.358 (1.302)	1.636 (1.530)	10.17 (1.30)	0.938	50.70 (11.65)	0.030 (0.376)	0.989 (0.266)	2.326 (1.391)	1.462 (1.861)	9.37 (1.36)
Triploid (two stage)	0.998	50.17 (6.80)	−0.008 (0.430)	0.988 (0.290)	2.286 (1.208)	2.047 (1.180)	10.25 (1.21)	0.986	49.77 (6.37)	0.010 (0.301)	1.010 (0.223)	2.224 (1.119)	1.529 (1.223)	9.65 (1.16)
Method	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = -2$	$d_2 = 0$	$\sigma^2 = 1.9375$	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = -2$	$d_2 = 0$	$\sigma^2 = 7.75$
Diploid	0.932	48.98 (12.26)	−1.003 (0.121)	1.005 (0.233)	—	—	2.640 (0.272)	0.970	49.52 (10.31)	−0.253 (0.215)	1.492 (0.341)	−0.455 (0.584)	—	8.466 (0.896)
Triploid (one stage)	0.950	49.97 (11.16)	−1.091 (0.285)	0.278 (0.238)	2.352 (0.682)	—	1.925 (0.241)	0.974	51.01 (10.37)	−0.060 (0.340)	1.004 (0.233)	−1.393 (2.149)	—	7.383 (0.965)
Triploid (two stage)	0.998	50.07 (4.53)	−0.026 (0.184)	0.986 (0.124)	−1.791 (0.689)	—	1.916 (0.217)	0.992	50.48 (5.41)	−0.011 (0.256)	1.012 (0.192)	−1.903 (1.246)	—	7.595 (0.860)
Method	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = 0$	$d_2 = 0$	$\sigma^2 = 1.75$	Power	Posi = 50	$\mu = 0$	$a = 1$	$d_1 = 0$	$d_2 = 0$	$\sigma^2 = 7.00$
Diploid	1.000	48.75 (6.56)	−0.752 (0.116)	1.475 (0.235)	—	—	2.361 (0.243)	1.000	49.40 (8.97)	−0.002 (0.202)	1.497 (0.320)	—	—	7.658 (0.812)
Triploid (one stage)	1.000	49.57 (5.82)	−0.012 (0.161)	0.991 (0.122)	—	—	1.736 (0.235)	1.000	50.71 (8.72)	−0.011 (0.202)	1.012 (0.210)	—	—	6.958 (0.850)
Triploid (two stage)	1.000	50.29 (3.59)	0.023 (0.134)	1.004 (0.103)	—	—	1.727 (0.202)	1.000	50.37 (5.16)	−0.005 (0.196)	1.012 (0.181)	—	—	6.996 (0.768)

For each parameter setting, 500 replicates, each with sample size 200, were analyzed with one QTL controlling the trait variation and located in the middle of a chromosome. Permutation tests were used to determine the critical value at 0.05 significance level in each parameter setting. The critical values for the diploid methods in the backcross (F_2) population are 7.171 (9.646), 6.607 (9.682), 7.169 (9.631), and 6.799 (6.402), respectively, in the four parameter settings. The critical values for the triploid methods under the one-stage design in the backcross (F_2) population are 8.785 (11.186), 8.801 (10.695), 12.497 (9.203), and 6.824 (6.531), respectively. The critical values for the triploid methods under the two-stage design in the backcross (F_2) population are 11.396 (11.418), 11.690 (11.388), 11.629 (10.416), and 7.540 (7.140), respectively. The numbers in the parentheses denote standard deviations. The data from the one-stage and two-stage designs were analyzed by the triploid method. Posi, position.

and they are 49.00 (SD 24.26) and 51.03 (SD 11.19) under the one-stage design, respectively. Therefore, the triploid model performs significantly better than the diploid model in this setting. In other settings, the triploid model under the two-stage design is also found to be much more powerful and precise than the diploid model, but the triploid model under the one-stage design seems to provide power and precision (in position estimation) similar to the diploid model. For example, in the third setting of the backcross population with $h^2 = 0.1$, the diploid model has power 0.400 and mean estimated position 49.03 (SD 23.68; Table 1). For the triploid model, they are 0.406 and 50.33 (SD 23.67) under the one-stage design, and they are 0.800 and 49.15 (SD 12.38) under the two-stage design. In the second setting of the F_2 population with $h^2 = 0.1$, the diploid model has power 0.668 and mean estimated position 49.66 (SD 19.16). For the triploid model, they are 0.648 and 49.85 (SD 19.05) under the one-stage design, and they are 0.796 and 49.86 (SD 12.35) under the two-stage design. A similar pattern can also be found for the other settings in Tables 1 and 2.

The triploid model is found to have better performance under the two-stage design than under the one-stage design in this study. Under the two-stage design, the triploid model can provide higher power for QTL detection and more precise estimates for positions and effects. For example, in the first setting with $h^2 = 0.1$ in the backcross population, the powers are 0.190 and 0.730, respectively (Table 1), and the means of the position estimates are 47.48 (SD 29.98) and 49.39 (SD 15.10), respectively, under the two different designs. In the second setting with $h^2 = 0.2$ in the F_2 population, the powers are 0.938 and 0.986, respectively (Table 2), and the means of the position estimates are 50.70 (SD 11.65) and 49.77 (SD 6.37), respectively, under the two different designs. Besides, the triploid model under the one-stage design seems to have problems in correctly estimating the effects in the backcross population when the additive and dominance effects are in opposite direction. For example, in the first setting ($a = 1$, $d_1 = -2$, and $d_2 = -2$), the means of the effect estimates by the triploid model under the one-stage design are 0.199 (SD 0.452), 1.587 (SD 2.342), and −0.589 (SD 2.202), respectively, for $h^2 = 0.1$ (Table 1), and they are 0.143 (SD 0.344), 2.117 (SD 1.628), and −0.766 (SD 1.547), respectively, for $h^2 = 0.2$ (Table 2). These estimates are highly biased and imprecise under the one-stage design. Similar problems can also be found in the third setting ($a = 1$, $d_1 = -2$, and $d_2 = 0$) for the backcross population. Such estimation problems, however, do not occur in the F_2 population or under the two-stage design (see Tables 1 and 2), which may suggest that the F_2 population is a better population than the backcross population and the two-stage design might be a more suitable design than the one-stage design for mapping endosperm traits.

TABLE 3

Simulation results of QTL mapping using the IM- and MIM-based triploid methods

	Q_A					Q_B					Q_C				
Method	Posi = 50	$a_1 = 3$	$d_{11} = -3$	$d_{12} = -3$	Power	Posi = 50	$a_2 = 2.5$	$d_{21} = 4$	$d_{22} = 4$	Power	Posi = 50	$a_3 = 1.5$	Power	$i_{a_2 a_3} = 1$	$\sigma^2 = 51.66$
IM	48.94 (7.86)	3.003 (0.554)	−1.449 (3.451)	−3.955 (2.585)	0.97	50.95 (6.50)	2.543 (0.543)	4.967 (2.901)	2.662 (3.408)	0.94	50.83 (20.43)	1.612 (0.811)	0.41	—	—
MIM (without epistasis)	49.37 (6.34)	3.003 (0.551)	−2.033 (2.836)	−3.698 (2.737)	1.00	50.60 (6.63)	2.546 (0.530)	4.630 (2.580)	2.958 (3.346)	1.00	49.40 (18.29)	1.591 (0.716)	0.57	—	53.61 (9.171)
MIM (with epistasis)	49.37 (6.34)	2.986 (0.545)	−2.291 (2.745)	−3.695 (2.675)	1.00	50.60 (6.63)	2.527 (0.531)	4.153 (3.159)	3.538 (4.079)	1.00	48.97 (17.18)	1.510 (0.698)	0.71	0.904 (0.510)	50.39 (8.26)

The number of simulated replicates is 100. The three unlinked QTL, Q_A, Q_B, and Q_C, contribute 40% of the trait variation and are located in the middle of the chromosomes. The average LRT statistics provided by IM are 31.12 (SD 10.00), 25.99 (SD 8.82), and 9.26 (SD 6.45) for Q_A, Q_B, and Q_C, respectively. The average LRT statistics provided by MIM without epistasis are 35.18 (SD 10.57), 30.07 (SD 10.42), and 11.36 (SD 6.97), respectively. The average LRT statistic provided by MIM with epistasis is 16.84 (SD 7.79) for Q_C. The experimentwise critical values at 0.05 significance level based on 1000 permutations are 13.48, 12.57, and 9.36 for the additive-, one-dominant-, and complete-effect models. Posi, QTL position.

The simulation in the second part aims to evaluate and compare the differences between the proposed MIM-based and the current IM-based methods in mapping endosperm traits. The results are shown in Table 3. When the IM-based method is used to detect QTL, three different models, the additive-effect model (with a only), the one dominant-effect model (with a and d_1), and the complete-effect model (with a, d_1, and d_2), will be implemented in the search. The experimentwise critical values at 0.05 significance level are found to be 9.36, 12.57, and 13.48 for the three different models, respectively, by 1000 permutations. For the additive-effect model, the powers to detect Q_A, Q_B, and Q_C are 0.97, 0.96, and 0.41, respectively. For the one dominant-effect model, the powers to detect the three QTL are 0.97, 0.95, and 0.31, respectively. For the complete-effect model, the powers are 0.97, 0.94, and 0.31, respectively. The three models have similar powers to detect Q_A and Q_B, and the additive-effect model has greater power than the other two models to detect Q_C. Among the 100 replicates, the three models can detect either both or one of Q_A and Q_B in each replicate. The results of mapping Q_A and Q_B by the complete-effect model and mapping Q_C by the additive-effect model are presented in Table 3. In Table 3, the means of the position estimates for the three QTL are 48.94 (SD 7.86), 50.95 (SD 6.50), and 50.83 (SD 20.43), respectively. The average LRT statistics are 31.12 (SD 10.00), 25.99 (SD 8.82), and 9.26 (SD 6.45), respectively, for the three QTL. This shows that the larger QTL, Q_A and Q_B, can be detected with higher power and better precision as compared to the small QTL, Q_C. Besides, the estimates of additive effects generally are more precise than those of dominance effects. For example, the mean of $\hat{a}_1$ is 3.003 (SD 0.554), and the means of $\hat{d}_{11}$ and $\hat{d}_{12}$ are -1.449 (SD 3.451) and -3.995 (SD 2.585), respectively. One of the advantages of the MIM-based method is that it is capable of fitting the detected QTL into the model in further searching for the other QTL. When the MIM-based method considers only one QTL in the model ($m = 1$), the mapping results are identical to those obtained by the IM-based method. Among the 100 replicates analyzed by the IM-based method, most of the replicates (91 replicates) have both Q_A and Q_B detected. For the remaining 9 replicates, either Q_A or Q_B is detected. If the detected Q_A (Q_B) is fitted into the MIM-based model in the search ($m = 2$), the undetected Q_B (Q_A) in the 9 replicates can be identified and the already detected Q_B (Q_A) in the other replicates will have a larger LRT statistic by including either their partial or complete effects in the model (that is, the power for detecting Q_A and Q_B is 1.0 for MIM with $m = 2$). To shorten the article, only the results of considering complete effects of Q_A and Q_B in the analysis are presented (Table 3). The average (partial) LRT statistics of Q_A and Q_B increase to 35.18 (SD 10.57) and 30.07 (SD 10.42), respectively. Further, if these two detected QTL are fitted into the MIM model for QTL search along the third chromosome ($m = 3$), the power to detect Q_C is 0.57 (average LRT statistic 11.36 with SD 6.97) if only the additive effect (a_3) is considered (Table 3). The power decreases to 40% (36%) if the one-dominant-effect (complete-effect) model is considered (not shown). The means of the position estimates are 49.37 (SD 6.34), 50.60 (SD 6.63), and 49.40 (SD 18.29) for the three QTL, respectively, which become more precise as compared to those by the IM-based method. If epistasis is taken into account to search for the third chromosome, many different types of epistasis can be considered. For illustration, only the additive-by-additive epistatic effect between QTL is considered (see also GENETIC MODEL OF ENDOSPERM TRAITS for first taking the additive-by-additive effect into account). Among the three possible additive-by-additive effects, only the consideration of $i_{a_2a_3}$ improves the QTL detection. The power increases to 71% (Table 3) when $i_{a_2a_3}$ is considered in the MIM model ($m = 3$ with epistasis) to search for Q_C (critical value 12.57 by permutation tests; average partial LRT statistic 16.84 with SD 7.79). The mean estimate of $i_{a_2a_3}$ is 0.904 (SD 0.510), and the mean estimate of σ^2 is 50.39 (SD 8.26). The mean of position estimate for Q_C becomes 48.97 (SD 17.18), and the mean of the estimated effect is 1.510 (SD 0.698), which is more precise than that obtained by ignoring epistasis.

CONCLUSION AND DISCUSSION

The endosperm of a seed is a triploid tissue and has a more complicated genetic mechanism than the diploid tissues. Therefore, the traditional QTL mapping methods (LANDER and BOTSTEIN 1989; HALEY and KNOTT 1992; JANSEN 1993; ZENG 1994; CHURCHILL and DOERGE 1994; KAO *et al.* 1999; KAO and ZENG 2002) designed for traits under diploid control are not appropriate approaches to map for QTL underlying the endosperm traits because they ignore the triploid nature of endosperms. WU *et al.* (2002a,b) and XU *et al.* (2003) first considered the triploid inheritance of endosperms to propose IM-based triploid methods in the detection of the underlying QTL. In this article, a new triploid approach based on the MIM method is developed to take multiple QTL into account in the model for mapping endosperm traits. The proposed method can be implemented to analyze data from either the one-stage design using only maternal genotypes or the two-stage design using both maternal and embryo genotypes in the backcross and F_2 populations. As shown in this article, the triploid MIM method can provide better detection power and estimation precision, and it can analyze and search for epistatic QTL directly in comparison with the current IM-based methods when mapping endosperm traits. Some important issues in mapping endosperm traits, such as the problems of using the diploid mapping methods, the relation between the diploid and triploid methods, the variance components of genetic variance, the problems if effects are present and ignored, and the relative efficiency of the diploid and triploid models under different experimental designs, are also investigated analytically or by simulation.

The triploid mapping method can provide better power in detection and more precise estimation under the two-stage design than under the one-stage design in mapping endosperm traits as shown in the simulation study (Tables 1 and 2) and also demonstrated by Wu *et al.* (2002b). This is because the two-stage design, which provides both the maternal and embryo marker genotypes, is more informative than the one-stage design, which offers only the maternal marker genotype, in inferring the conditional probabilities of the endosperm QTL genotypes (see the website http://www.stat.sinica.edu.tw/chkao/ for the conditional probabilities under different experimental designs). In the backcross population, the one-stage design provides only 4 different marker genotypes, and these marker genotypes are noninformative in inferring QQQ, QQq, and Qqq as equal conditional probabilities are assigned to them. The two-stage design, however, can provide 16 different marker genotypes, and the marker genotypes are not informative only for QQq and Qqq. In the F_2 population, the one- and two-stage designs can provide 9 and 25 marker genotypes, respectively, and each marker genotype in either design is noninformative only for the genotypes QQq and Qqq. Therefore, the two-stage design is generally more informative than the one-stage design, and the F_2 population is generally more informative than the backcross design in inferring the conditional probabilities. As these conditional probabilities are the mixing proportions in the normal mixture likelihood, they play a very important role in the quality estimation of QTL parameters for the model. A more informative design or population can provide more detailed information in inferring the conditional probabilities and thus can help improve the estimation of QTL parameters. This argument can explain the reasons why the performance of the triploid method is generally poor under the one-stage design in the backcross population as compared to the performance under another data structure (see, for example, the simulation results in Tables 1 and 2 when the additive and dominance effects are in the opposite directions) and why the triploid method under the two-stage design can perform well with satisfactory power and precision in all the parameter settings. The two-stage design generally requires more genotyping work as both the genomes of the plants and their seeds need to be genotyped, and different sampling strategies for allocations of a given sample size between the two generations should be considered for cost control. Besides, Wu *et al.* (2002b) also pointed out that the different sampling strategies for allocations can affect the parameter estimation. Therefore, the best strategy of allocation for the two-stage design under the consideration of cost and estimation deserves further investigation in practical QTL mapping.

The traditional diploid methods proposed for mapping diploid traits have been applied to mapping endosperm traits by several researchers (Tan *et al.* 1999; Wang and Larkins 2001; Wang *et al.* 2001). Such applications generally violate the traditional belief that the endosperm traits are under the control of triploid mechanisms (Benner *et al.* 1989; Zhu and Weir 1994; Wu *et al.* 2002a,b). If the diploid methods are applied to mapping endosperm traits, the confounding problem in estimation will occur (Equations 9–11), and the sampling variances of the estimates will inflate. Consequently, the diploid methods can cause some problems, such as bias in estimation and loss in power, in mapping endosperm traits. Although the diploid methods have these problems, the simulation study indicates that, in some parameter settings, its performance (in power and position estimate) can be similar to the triploid method under the one-stage design (Tables 1 and 2) due mainly to the correlation between the genomes of the maternal plant and its endosperms. Therefore, the diploid method can still be used as a preliminary method in mapping endosperm traits. By taking the triploid mechanism into account, the triploid method, especially under the two-stage design, can effectively solve the problems and significantly improve the mapping of endosperm traits.

The proposed MIM-based triploid method is a multiple-QTL model. This multiple-QTL approach distinguishes itself from the current IM-based methods of Wu *et al.* (2002a,b) and Xu *et al.* (2003) by the ability to use multiple-marker intervals simultaneously to fit multiple QTL into the model in mapping endosperm traits. As a result, the proposed method can provide greater power and precision, and it can readily analyze and search for epistatic QTL in endosperm trait mapping. Besides, the estimation procedures between these methods are different. The likelihood of the MIM-based method is a mixture of 4^m normals and will become increasingly unwieldy in maximization as the number of QTL (m) fitted into the model increases. To solve the maximization problem with large m, the general formulas proposed by Kao and Zeng (1997) are applied to obtain the MLE of QTL effects as well as their variance-covariance matrix (see the mim model for mapping endosperm traits). The procedure of the general formulas is a maximum-likelihood approach based on the EM algorithm. The method by Xu *et al.* uses an iteratively reweighted least squares (IRWLS) procedure, which is a second-order approximation to the maximum likelihood, and it has problems in estimating the two dominance effects separately as pointed out by Xu *et al.* The estimation procedure in Wu *et al.* also implements a maximum-likelihood approach via the EM algorithm, but it needs additional procedures in the M-step to obtain the MLE if some QTL effects are not considered in the model (see appendix b in Wu *et al.* 2002b). The general formulas, however, do not have these problems and are relatively straightforward and simple to maximize. An initial version of the triploid MIM program source code (written in Fortran 77 language) is available on the worldwide web (http://www.stat.sinica.edu.tw/chkao/).

It has been pointed out that the critical value for claiming QTL detection is a very complicated issue and deserves further investigation (Lander and Botstein 1989; Jansen 1993; Zeng 1994; Kao *et al.* 1999). Generally, the critical value depends on the number and size of intervals, different levels of heritability (size of QTL), different numbers of (linked or unlinked) QTL, and linked QTL in the same or opposite direction of effects. Visscher and Haley (1996) pointed out that the critical value should be reduced after a QTL of large effect has been detected. The determination of critical value in mapping endosperm traits will be more complicated as each QTL can have three possible effects and many different types of epistasis, and more different experimental designs (the one-stage or two-stage design with different allocations in the backcross or F_2 population) can be considered. In this article, the permutation tests by Churchill and Doerge (1994) are used to determine the critical value for claiming QTL detection in endosperm trait mapping. It is found that the critical value for the triploid model in the two-stage design is larger than that in the one-stage design (Tables 1 and 2). Given the same heritability, the critical value in the F_2 population is larger than that in the backcross population except for the third setting. More efforts are needed to unravel the issue of critical value in mapping endosperm traits. The understanding of QTL underlying the endosperm traits is very important to cereal breeding in improving yield potential and grain quality. This MIM-based triploid method can serve as an effective tool to estimate the parameters associated with the underlying QTL in mapping endosperm traits. Another important issue worth pursuing is to investigate the properties of different genetic models in mapping endosperm traits. Besides, several researchers (Zhu and Weir 1994; Mazur *et al.* 1999; van der Meer *et al.* 2001; Wu *et al.* 2002b; Xu *et al.* 2003) have pointed out that the maternal and offspring genomes could jointly affect the seed- or endosperm-specific traits. Therefore, it is important to take the genome information about the two successive generations into account in mapping those traits and, more importantly, to do so on the basis of a multiple-QTL model approach.

The author is grateful to two anonymous reviewers for helpful comments and to Pei-Ying Shih and Chu-Chun Chen for the derivation of variance components. This work was supported by grants NSC92-2118-M-001-038 from the National Science Council, Taiwan, Republic of China.

LITERATURE CITED

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

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 967–971.

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

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

Haldane, J. B. S., 1919 The combination of linkage values and the calculation of distances between the loci of linked factors. J. Genet. **8:** 299–309.

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

Jansen, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135:** 205–211.

Kao, C.-H., and Z-B. Zeng, 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:** 359–371.

Kao, C.-H., and Z-B. Zeng, 2002 Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics **160:** 1243–1261.

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

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

Lou, X.-Y., and J. Zhu, 2002 Analysis of genetic effects of major genes and polygenes on quantitative traits. II. Genetic models for seed traits of crops. Theor. Appl. Genet. **105:** 964–971.

Louis, T. A., 1982 Finding the observed information matrix when using the EM algorithm. J. R. Stat. Soc. Ser. B **44:** 226–233.

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

Mo, H.-D., 1987 Genetic expression for endosperm traits, pp. 478–487 in *Proceedings of the Second International Conference on Quantitative Genetics*, edited by B. Weir, E. J. Eisen, M. M. Goodman and G. Namkoong. Sinauer Associates, Sunderland, MA.

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

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

Stuart, A., and J. K. Ord, 1991 *Kendall's Advanced Theory of Statistics*, Ed. 5, Vol. 2. Oxford University Press, New York .

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

Visscher, P. M., and C. S. Haley, 1996 Detection of the putative quantitative trait loci in line crosses under infinitesimal genetic models. Theor. Appl. Genet. **93:** 691–702.

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.

Wu, R.-L., X.-Y. Lou, C.-X. Ma, X. Wang, B. A. Larkins *et al.*, 2002a An improved genetic model generates high-resolution mapping of QTL for protein quality in maize endosperm. Proc. Natl. Acad. Sci. USA **99:** 11281–11286.

Wu, R.-L., C.-X. Ma, M. Gallo-Meagher, R. C. Littell and G. Casella, 2002b Statistical methods for dissecting triploid endosperm traits using molecular markers: an autogamous model. Genetics **162:** 875–892.

Xu, C., X. He and S. Xu, 2003 Mapping quantitative trait loci underlying triploid endosperm traits. Heredity **90:** 228–235.

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

Zeng, Z-B., C.-H. Kao and C. Basten, 1999 Estimating the genetic architecture of quantitative traits. Genet. Res. **74:** 279–289.

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

Communication editor: A. Paterson

APPENDIX A: THE GENETIC VARIANCE COMPONENTS OF ENDOSPERM TRAITS

When m QTL with complete marginal and epistatic effects are considered together, the genetic variance of an endosperm trait can be decomposed into $4^m \times (4^m - 1)/2$ variance and covariance components. Taking $m = 2$ as an example, the genetic variance can have 120 variance and covariance components in the backcross and F_2 populations (not shown). If the two QTL are unlinked, the genetic variance reduces to 83 and 111 components in the two populations. For the F_2 population, these components are

$$\begin{aligned}
V_G = {} & \frac{7}{4}a_1^2 + \frac{7}{64}d_{11}^2 + \frac{7}{64}d_{12}^2 + \frac{7}{4}a_2^2 + \frac{7}{64}d_{21}^2 + \frac{7}{64}d_{22}^2 + \frac{49}{16}i_{a_1a_2}^2 + \frac{7}{32}i_{d_{11}a_2}^2 + \frac{7}{32}i_{d_{12}a_2}^2 \\
& + \frac{7}{32}i_{a_1d_{21}}^2 + \frac{63}{4096}i_{d_{11}d_{21}}^2 + \frac{63}{4096}i_{d_{12}d_{21}}^2 + \frac{7}{32}i_{a_1d_{22}}^2 + \frac{63}{4096}i_{d_{11}d_{22}}^2 + \frac{63}{4096}i_{d_{12}d_{22}}^2 \\
& + \frac{1}{8}(a_1d_{11} + a_2d_{21} - a_1d_{12} - a_2d_{22}) + \frac{7}{16}(a_1i_{a_1d_{21}} + a_1i_{a_1d_{22}} + a_2i_{d_{11}a_2} + a_2i_{d_{12}a_2}) \\
& + \frac{1}{64}(a_1i_{d_{11}d_{21}} + a_1i_{d_{11}d_{22}} - a_1i_{d_{12}d_{21}} - a_1i_{d_{12}d_{22}} + a_2i_{d_{11}d_{21}} + a_2i_{d_{12}d_{21}} \\
& \qquad - a_2i_{d_{11}d_{22}} - a_2i_{d_{12}d_{22}}) - \frac{1}{32}(d_{11}d_{12} + d_{21}d_{22}) \\
& + \frac{1}{64}(d_{11}i_{a_1d_{21}} + d_{11}i_{a_1d_{22}} - d_{12}i_{a_1d_{21}} - d_{12}i_{a_1d_{22}} + d_{21}i_{d_{11}a_2} \\
& \qquad + d_{21}i_{d_{12}a_2} - d_{22}i_{d_{11}a_2} - d_{22}i_{d_{12}a_2}) \\
& + \frac{7}{256}(d_{11}i_{d_{11}d_{21}} + d_{11}i_{d_{11}d_{22}} + d_{12}i_{d_{12}d_{21}} + d_{12}i_{d_{12}d_{22}} + d_{21}i_{d_{11}d_{21}} \\
& \qquad + d_{21}i_{d_{12}d_{21}} + d_{22}i_{d_{11}d_{22}} + d_{22}i_{d_{12}d_{22}}) \\
& + \frac{1}{256}(d_{11}i_{d_{12}d_{21}} + d_{11}i_{d_{12}d_{22}} + d_{12}i_{d_{11}d_{21}} + d_{12}i_{d_{11}d_{22}} \\
& \qquad + d_{21}i_{d_{11}d_{22}} + d_{21}i_{d_{12}d_{22}} + d_{22}i_{d_{11}d_{21}} + d_{22}i_{d_{12}d_{21}}) \\
& + \frac{7}{32}(i_{a_1a_2}i_{d_{11}a_2} + i_{a_1a_2}i_{a_1d_{21}} - i_{a_1a_2}i_{d_{12}a_2} - i_{a_1a_2}i_{a_1d_{22}}) \\
& + \frac{1}{128}(i_{a_1a_2}i_{d_{11}d_{21}} + i_{a_1a_2}i_{d_{12}d_{22}} - i_{a_1a_2}i_{d_{12}d_{21}} - i_{a_1a_2}i_{d_{11}d_{22}}) \\
& + \frac{1}{128}(i_{d_{11}a_2}i_{a_1d_{21}} + i_{d_{12}a_2}i_{a_1d_{22}} - i_{d_{11}a_2}i_{a_1d_{22}} - i_{d_{12}a_2}i_{a_1d_{21}}) \\
& + \frac{1}{64}(i_{d_{11}a_2}i_{d_{11}d_{21}} + i_{d_{12}a_2}i_{d_{12}d_{21}} - i_{d_{11}a_2}i_{d_{11}d_{22}} - i_{d_{12}a_2}i_{d_{12}d_{22}} + i_{a_1d_{21}}i_{d_{11}d_{21}} \\
& \qquad + i_{a_1d_{22}}i_{d_{11}d_{22}} - i_{a_1d_{21}}i_{d_{12}d_{21}} - i_{a_1d_{22}}i_{d_{12}d_{22}}) \\
& - \frac{1}{2048}(i_{d_{11}d_{21}}i_{d_{12}d_{21}} + i_{d_{11}d_{21}}i_{d_{11}d_{22}} + i_{d_{11}d_{21}}i_{d_{12}d_{22}} + i_{d_{12}d_{21}}i_{d_{11}d_{22}} + i_{d_{12}d_{21}}i_{d_{12}d_{22}}) \\
& - \frac{1}{2048}\, i_{d_{11}d_{22}}i_{d_{12}d_{22}}.
\end{aligned}$$

Likewise, the components of variance and covariance for the backcross population can be also obtained.

APPENDIX B: THE RELATION BETWEEN THE PARAMETERS OF THE DIPLOID AND TRIPLOID MODELS IN MAPPING ENDOSPERM TRAITS

To simplify the argument, assume that an endosperm trait value, y, measured in the backcross or F_2 population is affected only by a single QTL, Q. The backcross individuals can have two possible QTL genotypes, Qq ($w = \frac{1}{2}$) and qq ($w = -\frac{1}{2}$), each with frequency 1/2. The F_2 individuals can have three possible QTL genotypes, QQ ($w_1 = 1$, $w_2 = -\frac{1}{2}$), Qq ($w_1 = 0$, $w_2 = \frac{1}{2}$), and qq ($w_1 = -1$, $w_2 = -\frac{1}{2}$), with frequencies 1/4, 1/2, and 1/4, respectively. For autogamous plants, the individuals with QQ or qq genotype can produce only one endosperm genotype, QQQ or qqq. The individuals with Qq genotype can produce four kinds of endosperm genotype, QQQ ($x = \frac{3}{2}$, $z_1 = 0$, $z_2 = 0$), QQq ($x = \frac{1}{2}$, $z_1 = 1$, $z_2 = 0$), Qqq ($x = -\frac{1}{2}$, $z_1 = 0$, $z_2 = 1$), and qqq ($x = -\frac{3}{2}$, $z_1 = 0$, $z_2 = 0$), each with frequency 1/4. The frequencies of the four triploid QTL genotypes are 1/8, 1/8, 1/8, and 5/8, respectively, in the backcross population, and they are 3/8, 1/8, 1/8, and 3/8, respectively, in the F_2 population. The covariances between the coded variables for the QTL genotypes of a diploid individual and its triploid endosperm are found to be $\mathrm{Cov}(x, w) = \frac{3}{8}$, $\mathrm{Cov}(z_1, w) = \frac{1}{16}$, $\mathrm{Cov}(z_2, w) = \frac{1}{16}$ in the backcross population, and they are $\mathrm{Cov}(x, w_1) = \frac{3}{4}$, $\mathrm{Cov}(z_1, w_1) = 0$, $\mathrm{Cov}(z_2, w_1) = 0$, $\mathrm{Cov}(x, w_2) = 0$, $\mathrm{Cov}(z_1, w_2) = \frac{1}{16}$, and $\mathrm{Cov}(z_2, w_2) = \frac{1}{16}$ in the F_2 population.

If the diploid models in Equation 6 or 7 are applied to analyze a marker, M, to infer Q along the genome, the regression coefficient of y on the marker M (coded by w_M) in the backcross model is given by $b_{yM} = \mathrm{Cov}(y, w_M)/V(w_M)$, where $\mathrm{Cov}(y, w_M)$ is the covariance between the endosperm trait and the marker variable, and $V(w_M)$ is the variance of the marker variable. It is easy to obtain

$$\begin{aligned}
\mathrm{Cov}(y, w_M) &= \mathrm{Cov}(\mu + ax + d_1z_1 + d_2z_2 + \varepsilon, w_M) \\
&= a\mathrm{Cov}(x, w_M) + d_1\mathrm{Cov}(z_1, w_M) + d_2\mathrm{Cov}(z_2, w_M) \\
&= (1 - 2r_{QM})\left[\frac{3}{8}a + \frac{1}{16}(d_1 + d_2)\right]
\end{aligned}$$

if there is no covariance between the residual error and marker variable. The regression coefficient is $b_{yM} = (1 - 2r_{QM})[3a/2 + (d_1 + d_2)/4]$ because $V(w_M) = \frac{1}{4}$. Similarly, the two regression coefficients for the additive and dominance effects of M in the F_2 diploid model can be obtained. The regression coefficient of the additive variable is

$$b_{yM_a} = \frac{\mathrm{Cov}(y, w_{M_a})}{V(w_{M1})} = \frac{3(1 - 2r_{QM})a}{2},$$

and regression coefficient of the dominance variable is

$$b_{yM_d} = \frac{\mathrm{Cov}(y, w_{M_d})}{V(w_{M2})} = \frac{(1 - 2r_{QM})(d_1 + d_2)}{4}.$$

Note that the partial regression coefficients for the additive and dominance effects are the same as b_{yM_a} and b_{yM_d}, as w_{M_a} and w_{M_d} are orthogonal in the F_2 population.

The conditional phenotypic variance on the marker M for the backcross diploid model is $\sigma_{y.M}^2 = \sigma_y^2 - b_M \times$

σ_{yM}, where σ_y^2 is the phenotypic variance, and σ_{yM} denotes the covariance between y and M. The conditional phenotypic variance is

$$\sigma_{y.M}^2 = (\sigma^2 + \sigma_G^2) - (1 - 2r_{QM})^2 \times \left[\frac{3}{2}a + \frac{1}{4}(d_1 + d_2)\right] \times \left[\frac{3}{8}a + \frac{1}{16}(d_1 + d_2)\right]$$

$$= \sigma^2 + (1 - 2r_{QM})^2 \times \left[\frac{5}{8}a^2 + \frac{3}{32}(d_1^2 + d_2^2) + \frac{1}{8}(ad_1) - \frac{1}{8}(ad_2) - \frac{1}{16}(d_1 d_2)\right],$$

where σ^2 is the variance of residual error. For the F_2 diploid model, the conditional phenotypic variance on the marker M is $\sim\sigma_{y.M}^2 = \sigma_y^2 - (b_{yM\mathrm{a}} \times \sigma_{yM\mathrm{a}} + b_{yM\mathrm{d}} \times \sigma_{yM\mathrm{d}})$. The conditional phenotypic variance is

$$\sigma_{y.M}^2 = (\sigma^2 + \sigma_G^2) - (1 - 2r_{QM})^2 \times \left[\frac{3}{2}a \times \frac{3}{4}a + \frac{1}{4}(d_1 + d_2) \times \frac{1}{16}(d_1 + d_2)\right]$$

$$= \sigma^2 + (1 - 2r_{QM})^2 \times \left[\frac{5}{8}a^2 + \frac{3}{32}(d_1^2 + d_1^2) + \frac{1}{8}(ad_1) - \frac{1}{8}(ad_2) - \frac{1}{16}(d_1 d_2)\right].$$

The conditional phenotypic variances are the same for the backcross and F_2 models.

APPENDIX C: CONDITIONAL PROBABILITIES OF ENDOSPERM QTL GENOTYPES

Consider a marker interval, I_j, flanked by markers, M_j and N_j, on a linkage group. For the plants in the F_2 population, there are nine observable genotypes for markers M_j and N_j. They are M_jN_j/M_jN_j, M_jN_j/M_jn_j, M_jn_j/M_jn_j, M_jN_j/m_jN_j, $M_jm_jN_jn_j$ (M_jN_j/m_jn_j or M_jn_j/m_jN_j), M_jn_j/m_jn_j, m_jN_j/m_jN_j, m_jN_j/m_jm_j, and m_jn_j/m_jm_j with proportions $(1 - r)^2/4$, $r(1 - r)/2$, $r^2/4$, $r(1 - r)/2$, $(1 - r)^2/2 + r^2/2$, $r(1 - r)/2$, $r^2/4$, $r(1 - r)/2$, and $(1 - r)^2/4$, respectively. For the plants in the backcross population, there are four observable genotypes, M_jN_j/M_jN_j, M_jN_j/M_jn_j, M_jN_j/m_jN_j, and M_jN_j/m_jn_j, with proportions $(1 - r)/2$, $r/2$, $r/2$, $(1 - r)/2$, respectively. For autogamous plants, the plants with genotypes M_jN_j/M_jN_j, M_jn_j/M_jn_j, m_jN_j/m_jN_j, and m_jn_j/m_jm_j each can produce only one progeny (embryo) genotype. The plants with genotypes M_jN_j/M_jn_j, M_jN_j/m_jN_j, M_jn_j/m_jn_j, and m_jN_j/m_jm_j each can produce three different embryo genotypes. For example, the three embryo genotypes produced by plants with genotype M_jN_j/M_jn_j are M_jN_j/M_jN_j, M_jN_j/M_jn_j, and M_jn_j/M_jn_j. The plants with genotype M_jN_j/m_jn_j (M_jn_j/m_jN_j) can produce nine different embryo genotypes. A total of 25 and 16 different combinations of the plant and embryo genotypes are in the F_2 and backcross populations, respectively.

If an unobservable QTL, Q_j, is located in I_j, among the seeds (progeny) collected from the F_2 plants, there are three possible embryo genotypes, Q_jQ_j, Q_jq_j, and q_jq_j, and four possible endosperm genotypes, $Q_jQ_jQ_j$, $Q_jQ_jq_j$, $Q_jq_jq_j$, and $q_jq_jq_j$. The conditional distribution of these endosperm genotypes given the observable marker genotypes of the F_2 plant (t) and embryo ($t + 1$) can be derived on the basis of Haldane's mapping function (Haldane 1919) assuming no crossover interference. For example, the conditional probabilities of the endosperm genotype, $Q_jQ_jQ_j$, given the plant genotype $M_jN_j/M_jn_j^{(t)}$ and its embryo genotype $M_jN_j/M_jN_j^{(t+1)}$ are calculated as

$$\mathrm{Prob}\left(Q_jQ_jQ_j | \frac{M_jN_j^{(t)}}{M_jn_j}, \frac{M_jN_j^{(t+1)}}{M_jN_j}\right) = \frac{\mathrm{Prob}(Q_jQ_jQ_j,\ M_jN_j^{(t)}/M_jn_j,\ M_jN_j^{(t+1)}/M_jN_j)}{\mathrm{Prob}(M_jN_j^{(t)}/M_jn_j,\ M_jN_j/M_jN_j^{(t+1)})}. \quad \text{(C1)}$$

The probability in the denominator of Equation C1 is $r(1 - r)/8$. As the QTL endosperm genotype $Q_jQ_jQ_j$ implies the embryo genotype Q_jQ_j, it ensures that the marker and QTL genotype of the embryo is $M_jQ_jN_j/M_jQ_jN_j^{(t+1)}$. The possible F_2 plants that can produce such an embryo genotype should be from one of the three genotypes, $M_jQ_jN_j/M_jq_jn_j^{(t)}$, $M_jq_jN_j/M_jQ_jn_j^{(t)}$, and $M_jQ_jN_j/M_jQ_jn_j^{(t)}$. It is easy to obtain that the probabilities of the F_2 plants with these three genotypes are $r_1(1 - r_1)(1 - r_2)^2/2$, $r_1(1 - r_1)r_2^2/2$, and $(1 - r_1)^2r_2(1 - r_2)/2$, respectively, and that their chances to produce seeds with embryo genotype $M_jQ_jN_j/M_jQ_jN_j^{(t+1)}$ are $(1 - r_2)^2/4$, $r_2^2/4$, and $1/4$, respectively. This allows calculation of the numerator of Equation C1 as the sum of the following three probabilities:

$$\mathrm{Prob}\left(\frac{M_jQ_jN_j^{(t)}}{M_jq_jn_j}, \frac{M_jQ_jN_j^{(t+1)}}{M_jQ_jN_j}\right) + \mathrm{Prob}\left(\frac{M_jq_jN_j^{(t)}}{M_jQ_jn_j}, \frac{M_jQ_jN_j^{(t+1)}}{M_jQ_jN_j}\right) + \mathrm{Prob}\left(\frac{M_jQ_jN_j^{(t)}}{M_jQ_jn_j}, \frac{M_jQ_jN_j^{(t+1)}}{M_jQ_jN_j}\right)$$

$$= \frac{r_1(1 - r_1)(1 - r_2)^2}{2} \times \left(\frac{1 - r_2}{2}\right)^2 + \frac{r_1(1 - r_1)r_2^2}{2} \times \left(\frac{r_2}{2}\right)^2 + \frac{(1 - r_1)^2r_2(1 - r_2)}{2} \times \frac{1}{4}$$

$$= \frac{r_1(1 - r_1)[(1 - r_2)^4 + r_2^4] + (1 - r_1)^2r_2(1 - r_2)}{8}.$$

Therefore, the conditional probability of the endosperm genotype, $Q_jQ_jQ_j$, given the plant marker genotype, $M_jN_j/M_jn_j^{(t)}$, and its embryo marker genotype, $M_jN_j/M_jN_j^{(t+1)}$, is

$$\mathrm{Prob}\left(Q_jQ_jQ_j | \frac{M_jN_j^{(t)}}{M_jn_j}, \frac{M_jN_j^{(t+1)}}{M_jN_j}\right) = \frac{r_1(1 - r_1)[(1 - r_2)^4 + r_2^4] + (1 - r_1)^2r_2(1 - r_2)}{r(1 - r)}.$$

The same argument leads the other three conditional probabilities of the endosperm genotypes, $Q_jQ_jq_j$, $Q_jq_jq_j$, and $q_jq_jq_j$, to

$$\mathrm{Prob}\left(Q_jQ_jq_j | \frac{M_jN_j^{(t)}}{M_jn_j}, \frac{M_jN_j^{(t+1)}}{M_jN_j}\right) = \frac{r_1(1 - r_1)r_2(1 - r_2)[r_2^2 + (1 - r_2)^2]}{r(1 - r)}$$

$$\mathrm{Prob}\left(Q_jq_jq_j | \frac{M_jN_j^{(t)}}{M_jn_j}, \frac{M_jN_j^{(t+1)}}{M_jN_j}\right) = \frac{r_1(1 - r_1)r_2(1 - r_2)[r_2^2 + (1 - r_2)^2]}{r(1 - r)}.$$

$$\text{Prob}\left(q_jq_jq_j|\frac{M_jN_j^{(t)}}{M_jn_j}, \frac{M_jN_j^{(t+1)}}{M_jN_j}\right) = \frac{r_1r_2(1-r_2)[2(1-r_1)(1-r_2)r_2+r_1]}{r(1-r)}.$$

Similarly, the conditional probabilities of endosperm QTL genotypes given the other combinations of the F_2 (backcross) plant and embryo genotypes (the two-stage design) can be derived. If only the plant marker genotype (the one-stage design) is available for inference, the derivation for the conditional probabilities of endosperm QTL genotypes is simpler and can be also obtained. These conditional probabilities under the one- and two-stage designs in the backcross and F_2 populations are placed on the website (http://www.stat.sinica.edu.tw/chkao/) or a part of them can be found in Wu *et al.* (2002a,b) and Xu *et al.* (2003).

APPENDIX D: THE PROBLEMS IF EFFECTS ARE PRESENT AND IGNORED IN MAPPING ENDOSPERM TRAITS

For simplicity, assume that an endosperm trait is controlled by two QTL, Q_1 and Q_2, without epistasis. It can be found that the covariances between the coded variables for the effects of different QTL are

$$\text{Cov}(x_1, x_2) = \frac{9}{8}(1-r_{12}) + \frac{5}{8}(1-r_{12})(1-2r_{12}) - \frac{9}{16},$$

$$\text{Cov}(x_1, z_{21}) = -\frac{3r_{12}}{16} + \frac{(1-r_{12})(1-2r_{12})}{16} + \frac{3}{32},$$

$$\text{Cov}(x_1, z_{22}) = -\frac{3r_{12}}{16} - \frac{(1-r_{12})(1-2r_{12})}{16} + \frac{3}{32},$$

where r_{12} is the recombination fraction betwen Q_1 and Q_2 in the backcross population. In the F_2 population, these covariances become

$$\text{Cov}(x_1, x_2) = \frac{9}{8}(1-2r_{12}) + \frac{5}{8}(1-2r_{12})^2,$$

$$\text{Cov}(x_1, z_{21}) = \frac{(1-2r_{12})^2}{16},$$

$$\text{Cov}(x_1, z_{22}) = -\frac{(1-2r_{12})^2}{16}.$$

If Q_1 and Q_2 are unlinked ($r_{12} = 0.5$), the covariances are all zeros. In the backcross population, if a single-QTL model considering only the additive effect is used to analyze Q_1, the regression coefficient is

$$b_{yx_1} = a_1 + \frac{1}{19}[18(1-r_{12}) + 10(1-r_{12})(1-2r_{12}) - 9]a_2 + \frac{5}{38}d_{11} + \frac{1}{38}d_{12}$$
$$- \frac{1}{19}\left[3r_{12} - (1-r_{12})(1-2r_{12}) - \frac{3}{2}\right]d_{21}$$
$$- \frac{1}{19}\left[3r_{12} - (1-r_{12})(1-2r_{12}) - \frac{3}{2}\right]d_{22}.$$

If only a dominance effect, say d_1, is considered, the regression coefficient is

$$b_{yz_1} = \frac{10}{7}a_1 - \frac{1}{7}[12r_{12} - 4(1-r_{12})(1-2r_{12}) - 6]a_2 + d_{11} - \frac{1}{7}d_{12}$$
$$+ \frac{1}{7}[8(1-r_{12})^3 - 1]d_{21} + \frac{1}{7}[8r_{12}^2(1-r_{12}) - 1]d_{22}.$$

In the F_2 population, the two coefficients are

$$b_{yx_1} = a_1 + \frac{1}{14}[9(1-2r_{12}) + 5(1-2r_{12})^2]a_2$$
$$+ \frac{1}{28}d_{11} - \frac{1}{28}d_{12} + \frac{(1-2r_{12})^2}{28}d_{21} - \frac{(1-2r_{12})^2}{28}d_{22}$$

and

$$b_{yz_1} = \frac{4}{7}a_1 + \frac{4(1-2r_{12})^2}{7}a_2 + d_{11} - \frac{1}{7}d_{12}$$
$$+ \frac{1}{7}\{8[r_{12}^4 + (1-r_{12})^4] - 1]\}d_{21} + \frac{1}{7}[16r_{12}^2(1-r_{12})^2 - 1]d_{22}.$$

They show that the estimate of the additive (dominance) effect of Q_1 is confounded by its other effects and the effects of Q_2.