Model selection for quantitative trait locus analysis in polyploids

R. W. Doerge*†‡ and Bruce A. Craig†

Departments of *Agronomy and †Statistics, Purdue University, West Lafayette, IN 47907

Communicated by John D. Axtell, Purdue University, West Lafayette, IN, February 24, 2000 (received for review January 20, 1999)

Over the years, substantial gains have been made in locating regions of agricultural genomes associated with characteristics, diseases, and agroeconomic traits. These gains have relied heavily on the ability to statistically estimate the association between DNA markers and regions of a genome (quantitative trait loci or QTL) related to a particular trait. The majority of these advances have focused on diploid species, even though many important agricultural crops are, in fact, polyploid. The purpose of our work is to initiate an algorithmic approach for model selection and QTL detection in polyploid species. This approach involves the construction of all possible chromosomal configurations (models) that may result in a gamete, model reduction based on estimation of marker dosage from progeny data, and lastly model selection. While simplified for initial explanation, our approach has demonstrated itself to be extendible to many breeding schemes and less restricted settings.

Detecting and locating genomic regions associated with quantitative traits is known as quantitative trait locus (QTL) mapping. The statistical methods (1–8) employed to identify QTL are numerous and rely heavily on the fact that the organism is diploid (i.e., homologous pairs of chromosomes arranged in sets). In the QTL analysis framework, diploidy ensures that the outcome of meiosis is predictable and that in most breeding schemes, molecular markers are at most single dose (one copy) and observable, and thus segregate in the ususal manner.

When there are more homologous chromosomes per set, the species is referred to as polyploid. While most animal species are diploid, there are many important agricultural crops such as sugarcane, cotton, banana, alfalfa, potato, coffee, and wheat that are polyploid. Among natural species of flowering plants, nearly half are polyploid (9). Even in animals, polyploidy is known to exist. Salmonid fish and specific amphibians display doubling and tripling of their ploidy level (9).

In some cases, such as the potato, a polyploid species is closely related to a diploid and standard diploid QTL analysis can be successful. In other cases, such as sugarcane, there is no closely related diploid species making QTL analysis very difficult. This difficulty is due to several inherent factors. First, the number of possible genotypes per marker and/or QTL are much greater in polyploids than diploids simply because of the increased number of chromosomes in the homologous set. Second, the numbers of copies of each marker and/or QTL (known as the dosage) in the parents and progeny is not obvious and are often not observable. Third, the additional doses (copies) of a marker can mask recombination information; and fourth, the meiosis process (i.e., pairing behavior and outcome of meiosis) of the species is usually unknown. Our task in this paper is to identify each of these important aspects of polyploidy and incorporate them into an algorithmic model selection process which will be used in a single marker analysis for QTL detection.

The two main characteristics that describe a polyploid are the number of chromosomes in each homologous set (ploidy level) and the pairing mechanism during meiosis. The pairing of chromosomes can range from preferential (always pairing with the same chromosome in the set) to completely random (equally likely to pair with any other chromosome in the set). Throughout the remainder of this paper, the term preferential pairing will be used, and it holds the same meaning as disomic pairing. Similarly, the term random pairing will be employed to mean nonpreferential and/or polysomic pairing. Unlike the diploid situation, where the meiotic process is known to involve the pairing of two homologous chromosomes, the process in a polyploid is unpredictable. A common assumption, and the one we will use throughout this paper, is that meiosis is simply an extension of the diploid case and involves multiple pairings of homologous chromosomes (i.e., preferential bivalent pairing). During polyploid meiosis, pairs of chromosomes in each homologous set align and possibly exchange genetic material (i.e., crossover). Each (bivalent) chromosomal pair then contributes one chromosome to the chromosomal set in each gamete.

The probability of each type of gamete depends on the specific set of homologous chromosomes (configuration), the ploidy level, and the pairing mechanism of the organism. Unlike the diploid case, the pairing mechanism is important because there are more than two chromosomes in a set. Species that display preferential pairing are known as allopolyploids, whereas species displaying random pairing are referred to as autopolyploids. Species intermediate to preferential pairing and random pairing are often represented as % polyploid/random. Our work will be based on a preferential pairing mechanism, thereby reducing the complexity of polyploidy to essentially that of a diploid. However, as noted throughout, our methods are directly extendible to more complicated % polyploid/random pairing mechanisms or the random pairing situation.

In addition to determining the probabilities of each chromosomal pairing during meiosis, the ploidy level, k, is important because it determines the possible dosage levels of the marker and QTL in both parents and progeny. The dosage, denoted by d, is the number of copies of a particular marker/QTL in a homologous set of chromosomes. If we consider a standard diploid backcross experimental design, there is at most one possible dose of each marker and/or QTL. For our polyploid situation, as many as $\frac{k}{2}$ copies of a genetic marker and/or QTL can be passed to the gamete. For example, in a tetraploid ($k = 4$), as many as $d = 2$ copies of the marker or QTL can be passed to a gamete. One key issue when considering dosage of QTL and/or marker for a polyploid is the fact that standard laboratory procedures cannot determine the genotypic state (dosage) of either, and they are instead estimated via segregation ratios. With this as a consideration, we restrict our attention to the even ploidy levels of 4, 6, and 8. Although there are species with an odd number of chromosomes in a homologous set, these species are characteristically of reduced fertility and of limited general interest in the QTL setting.

Abbreviation: QTL, quantitative trait locus

‡To whom reprint requests should be addressed at: Department of Statistics, 1399 Mathematical Sciences Building, Purdue University, West Lafayette, IN 47907-1399. E-mail: doerge@stat.purdue.edu

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

The complications of polyploidy have restricted the use of DNA markers for genetic mapping, as well as for identifying genomic regions responsible for quantitative traits. Wu *et al.* (10) derived a theoretical approach for mapping single dose DNA markers in polyploids under the assumption of nonpreferential bivalent pairing. da Silva and Sorrells (11) published similar theory with data in 1996. Ripol *et al.* (12) later developed theory for placing multiple dose markers on previously estimated maps comprised of single dose markers, by first estimating the dosage of the molecular marker and then relying on this information to determine its chromosomal pairing and relationship to known single dose markers. Furthermore, both da Silva and Sorrells (11) and Guimaraes and Sobral (13) pointed out that the use of multiple dose markers improves the accuracy of detection of pairing homologs and their organization into homology groups. Each of these works (10–13) is an important contribution toward understanding genome organization and evolution. Equally important in understanding history and its organization is the detection of QTL in association with multiple dose markers. Some of the first efforts to map QTL in polyploids (sugarcane) were performed by Sills *et al.* (14), and later extended by Guimaraes *et al.* (15). In these studies, various agronomically important traits were associated with single dose markers by means of multiple regression model building and maximum likelihood methods. In each of these QTL analyses, the model used to develop the likelihood function was limited to single dose markers. To date, no effort has been made to employ multiple dose markers for QTL analyses.

The potential for widening the study of polyploids was recently explored by Galitski *et al.* (16) through a microarray-based investigation of ploidy regulation of gene expression in *Saccharomyces cerevisiae* where different ploidies demonstrated different patterns of gene expression. Hieter and Griffiths (9) discuss the implications that are increasing our understanding of polyploids by pointing out that whatever the molecular reason, gene regulation depends on ploidy levels. Fueled by recent technologies, new questions are now being asked and directed toward the genetics and genomics of polyploids (17). The basis of many of these studies is deeply rooted in investigations involving the evolutionary significance of multiple origins of polyploid species. Each investigation holds promise toward moving our understanding of complex genomes to that of diploid species, and each has the long-term hope that conservation of QTL across species will supply additional knowledge to and from polyploid species.

Model Selection for QTL Analysis in Polyploids

The Experimental Model. Let us consider a pseudo-doubled backcross population (18) that is the result of selecting an informative parent, doubling half of its chromosomes to create a noninformative parent, and then crossing the two parental lines. By construction of the experiment, when the informative parent is crossed to the noninformative parent, pseudo-double backcross progeny result (18, 19). It is important to realize that the informative parent's genetic constitution (i.e., dosage of markers) is not known but may later be inferred from the pseudo-backcross progeny. For our purposes, we assume the noninformative parent marker and QTL dosages are zero. From this point forward, we concentrate on one homologous set of chromosomes taken from a pseudo–doubled backcross polyploid organism. The extension to the remainder of the chromosome sets is obvious, and direct.

Similar to the diploid QTL analysis, we assume that there are only two alleles at each marker and QTL, and we denote a molecular marker by M and a QTL by Q. Because we focus on a single marker and single QTL analysis, each homologous set is a mixture of only four types of chromosomes. These types are denoted as MQ (both present), M (only the marker present), Q (only the QTL present), and $\emptyset$ (neither M nor Q present). The number of each type of chromosome will depend on the ploidy level.

The Diploid Model. In a diploid, the pseudo–doubled backcross suits a standard backcross design initiated from two inbred parental lines that differ in the trait of interest. The basic idea of QTL analysis using single markers in diploid organisms is to associate observable marker genotypes with measurable quantitative traits. Marker genotypes are observable, dosage of the marker and unobservable QTL are known to be at most one, and quantitative traits are scored. The statistical methodology for doing single marker QTL analysis includes t-tests, regression, and likelihood ratio tests (see ref. 20 for review). When the likelihood is employed, it is a function of marker genotypes and varying mixtures of normal distributions that are controlled in number by the possible genotypes of the unknown QTL, as well as the mating design. Because the diploid meiotic process (e.g., chromosomal pairing, crossing over, gametic probabilities) is well understood, the likelihood function is easily stated as a function of marker genotype classification probability distributions, and numerically maximized with respect to parental means, variances, and recombination between the marker and QTL. A test statistic can then be calculated for the purpose of detecting/locating QTL.

All Possible Polyploid Models. In the diploid, there is only one model to consider; however, in the polyploid setting, one must model aspects of the chromosomal pairing, all possible gametic configurations that may result from chromosomal pairing, segregation, and independent assortment, as well as all possible dosages for both the marker and QTL. To consider all of the possible polyploid models, we break down this process, first focusing on a single homologous pair and then combining the chromosomal contributions of each pair. In anticipation of later, more complicated expressions, matrix representations of QTL and marker probabilities are used.

For each pair of homologous chromosomes, the probability of its contribution to the gamete can be expressed using a matrix of the form

$$\mathbf{C} = \begin{bmatrix} P(\emptyset) & P(Q) \\ P(M) & P(MQ) \end{bmatrix}. \qquad [1]$$

The rows and columns of the matrix **C** represent the possible dosage levels of the marker and QTL, respectively. The elements of the matrix **C** are probabilities that depend on the configuration of the paired chromosomes, and thus they are functions of recombination r.

Extending to the polyploid case, there are $\frac{k}{2}$ pairs of chromosomes in each homologous set, making the probabilities of the overall contribution a function of $\frac{k}{2}$ **C** matrices. Because each pair contributes to one of four possible gametes independently, the Kronecker product of the $\mathbf{C}_i$; $i = 1, \ldots, \frac{k}{2}$ matrices yields a $2^{\frac{k}{2}} \times 2^{\frac{k}{2}}$ probability matrix for each order-specific contribution. The Kronecker product, or the direct product, is a matrix algebraic mechanism that consists of all possible products of an element of a matrix multiplied by the elements of a second matrix. Because we are not interested in what each chromosomal pair specifically contributes to the gamete but rather the overall contribution of all $\frac{k}{2}$ pairs, we simplify this matrix such that its rows and columns represent the gamete's possible dosage levels for the genetic marker and QTL. Because each chromosomal pair can contribute at most one copy of the marker and QTL, the collapsed (or, simplified) matrix will be of dimension $(\frac{k}{2}+1) \times (\frac{k}{2}+1)$, instead of $2^{\frac{k}{2}} \times 2^{\frac{k}{2}}$.

The general algorithmic reduction of the full $2^{\frac{k}{2}} \times 2^{\frac{k}{2}}$ probability matrix is accomplished by multiplying each successive Kronecker product by a matrix $\mathbf{A}_i; i = 1, \ldots, \frac{k}{2}$ and its transpose. Each $\mathbf{A}_i$ is of dimension $2i \times i+1$ and consists of i $\mathbf{I}_{2\times 2}$ matrices along the main diagonal. The elements of $\mathbf{A}_i$ may be generalized by

$$a_{rc} = \begin{cases} 1 \text{ if } & r = 2c-1 \text{ or } 2c-2 \\ 0 & \text{otherwise} \end{cases}$$
$$r = 1, 2, ..., 2i \text{ and } c = 1, 2, ..., i+1.$$

For any allopolyploid, the following expression generates all gametic probabilities for allowable configurations of maximum dosage $\frac{k}{2}$

$$\mathbf{G} = \mathbf{A}_{\frac{k}{2}}^T(\ldots(\mathbf{A}_2^T((\mathbf{A}_1^T\mathbf{C}_1\mathbf{A}_1) \otimes \mathbf{C}_2)\mathbf{A}_2) \otimes \ldots \otimes \mathbf{C}_{\frac{k}{2}})\mathbf{A}_{\frac{k}{2}}.$$

Preferential pairing, like diploids, provides the most straightforward calculations as there is only one set of homologous pairs, $\mathbf{C}_i$, so the matrix $\mathbf{G}$ represents the gametic probabilities for specific ploidy and dosage levels. However, when pairing is random, there is more than one set of chromosomal pairs possible, and the gametic probabilities for all configurations are more extensive. For each set, one could construct the matrices $\{\mathbf{C}_i\}$ and produce the gametic probabilities as described. The overall gametic probabilities are then obtained by multiplying each $\mathbf{G}$ matrix by the probability that that set of chromosomal pairs occurs, and summing the matrices together. A simple example of this process is described in *Appendix A*, which is published as supplemental data on the PNAS web site (www.pnas.org).

With an end goal of assessing all possible polyploid models for the situation we are considering, we assume the ploidy level of the species has been previously studied and is known in advance, and that the dosage of both the marker and QTL is unknown. Realizing that the dosage levels regulate the final gametic probabilities, it is necessary to compute the resultant gametic probabilities for each possible dosage level of both QTL and marker and then, attempt to find the best model via model selection. For the pseudo–doubled backcross under consideration, the maximum dosage of either QTL or marker is $\frac{k}{2}$ and these copies are restricted to $\frac{k}{2}$ of the chromosomes in the homologous set. The remaining $\frac{k}{2}$ chromosomes are $\emptyset$ (null) chromosomes. For each combination of dosage levels, there is often more than one configuration of chromosomes possible. We, however, consider only the configuration that maximizes the number of MQ chromosomes in the homologous set. Under our preferential pairing mechanism this configuration maximizes the information concerning the recombination fraction (see *Appendix A* at www.pnas.org). Thus, the number of models that we consider is $(\frac{k}{2})^2$ which is less than the total number of models. For example, with $k = 8$, we consider 16 models instead of the full 26 models (*Appendix A*, www.pnas.org; see Table 6). The extension to all models, however, is straightforward, direct, and practical since the approach is algorithmic.

Polyploid Model Reduction. Having formulated all possible polyploid models, we now reduce the potential pool of models by estimating the dosage of the observable marker in the informative parent. The progeny that result from the pseudo–doubled backcross could be easily described solely by what was passed to them from the informative parent, if that information were observable. Even though we know that the informative parent has a marker, we do not know the dosage of that marker, denoted d_M. Relying on the backcross offspring, we can infer the dosage of the marker in the informative parent, which in turn provides additional information that reduces the pool of models from which we will eventually select the best model. Letting n denote the number of progeny, the probability of observing $n_\emptyset$ progeny with no marker given the informative parent dosage d_M is

$$\Pr(n_\emptyset | n, d_M) = \text{Bin}(n_\emptyset; n, p_{d_M}) = \binom{n}{n_\emptyset} p_{d_M}^{n_\emptyset}(1 - p_{d_M})^{n-n_\emptyset},$$

where $p_{d_M} = (1/2)^{d_M}$ and represents the probability of a progeny having zero copies of the marker when the informative parent has d_M. This conditional probability is a result of our pseudo–doubled backcross design and our preferential pairing mechanism. Under a random pairing situation, this procedure would follow similarly, except

$$p_{d_M} = \binom{k - d_M}{k/2} \Big/ \binom{k}{k/2}.$$

Employing this probability allows us to infer the dosage, d_M, of a marker in the informative parent, via a Bayesian approach. *A priori* we assume each possible dosage level ($d = 1, \ldots, \frac{k}{2}$, k assumed known) is equally likely, and use Bayes theorem (21) to compute the posterior probability of each dosage level,

$$\Pr(d_M | n, n_\emptyset) = \frac{\text{Bin}(n_\emptyset; n, p_{d_M})}{\sum_{d=1}^{k/2} \text{Bin}(n_\emptyset; n, p_{d_M})}.$$

If a particular dosage level has a posterior probability greater than an arbitrary cutoff, in our case we selected 90%, we then restrict our attention to only those models with that dosage level. If no dosage has probability greater than 90%, we select successive dosage levels based on the largest posterior probability until the sum of the probabilities is greater than 90%. By eliminating models that are highly unlikely, given the observed number with no marker present, we have reduced the potential models that need to be considered.

Model Selection and Parameter Estimation. With the dosage of the marker at least partially resolved, and a potential set of models available, the aim becomes selecting the single best model that will in turn provide the maximum likelihood estimates in the single marker QTL analysis. The form of the likelihood is similar to that of the diploid case except that there are now $\frac{k}{2}+1$ dosage levels of the QTL that lend $\frac{k}{2}+1$ possible phenotypic means. For example, assuming an additive dosage effect on the phenotypic mean and using $I_{m,i}$ to indicate presence of the marker for each individual i,

$$I_{m,i} = \begin{cases} 1 \text{ if individual } i \text{ has the marker} \\ 0 \qquad\qquad \text{otherwise} \end{cases}$$

the likelihood is

$$L = \prod_{I_{m,i}=0} \left(\sum_{j=0}^{k/2} \text{P}(Q_j) N(y_i; \mu_j, \sigma^2) \right) \times \prod_{I_{m,i}=1} \left(\sum_{j=0}^{k/2} \text{P}(MQ_j) N(y_i; \mu_j, \sigma^2) \right),$$

where $j = 0, \ldots, k/2$ represents the dosage of the QTL, $\text{P}(Q_j)$ is the gametic probability of no marker and j copies of the QTL, $\text{P}(MQ_j)$ is the gametic probability of at least one copy of the marker and j copies of the QTL, y_i is the quantitative trait value for the ith individual, and $N(y_i; \mu_j, \sigma^2)$ denotes a normal distribution. Recall that the probabilities $\text{P}(Q_j)$ and $\text{P}(MQ_j)$ are, respectively, elements and sums of elements of the matrix $\mathbf{G}$ and these elements are a function of the recombination fraction r (*Appendix A*, www.pnas.org). For the additive dosage effect, the

mean of the quantitative trait distribution for a specified QTL dosage is

$$\mu_j = \frac{(k-j)\mu_2 + j\mu_1}{k},$$

and depends on the ploidy level of the organism, as well as the means of the informative ($d_Q = k$) and noninformative parents, μ_1 and μ_2, respectively. The variance, σ^2, is assumed equal in both parents, but could easily be considered as two separate parameters. Use of the EM-algorithm (22) maximizes the likelihood function in a fashion similar to the diploid situation, only now the expectation step, or the E-step, involves a multinomial rather than binomial distribution.

Simulated Example Demonstrating the Model Selection Process

As an example of the model selection process, we detail, by using simulated data for 50 progeny, the algorithmic approach put forth in this work. The informative parent was double coupled (two copies of both marker and QTL on the same chromosome) with a recombination fraction of $r = 0.25$. We also assumed the quantitative trait, y, was normally distributed with mean $4d_Q$, where d_Q is the dosage of the QTL, and variance 1.0. The simulated data for this example can be found in Table 1.

Steps Involved in the Model Selection Process. *1. Estimating the marker dosage.* For this data set, 41 of the 50 progeny have at least one copy of the marker. The posterior probability of marker dosage 1 through 4 is summarized in Table 2. Recall that the expected fraction of progeny with no copies of the marker, when the parental dosage is d_M, is $(1/2)^{d_M}$. Because the sum of the posterior probabilities of marker dosage 2 and 3 is greater than 0.90, we restrict our search to just these two dosages.

2. Computing the likelihood for each model. Standard EM–algorithm (22) procedures are used to compute the maximum likelihood estimates for each of the models considered. The likelihood value and parameter estimates are summarized in Table 3. The parameter μ_0 is the mean of the quantitative trait when the QTL dosage is zero (the true value is 0). The parameter a represents the shift in the mean for an additional copy of the QTL (the true value is 4). For these data, the model ($d_M = 2$, $d_Q = 2$) has the highest likelihood so it would be selected as the model. It should be noted that this configuration is just slightly better than the model ($d_M = 3$, $d_Q = 2$), with very little difference in the parameter estimates.

Single Marker QTL Analysis in Polyploids

As demonstrated, the real complication arising from polyploidy is not the QTL analysis itself, but rather the model upon which the likelihood function is based. Selection of the single best model to represent the polyploid situation under investigation allows one to proceed with such a formulation of the likelihood function. This likelihood function, when coupled with a standard test statistic (i.e., LOD score or likelihood ratio test) can then be used to test various statistical hypotheses concerning QTL detection and effect, as well as QTL location. Relying on Monte Carlo resampling procedures, the distribution of the test statistic can be estimated and the meaning of statistical significance understood for the polyploid at hand. For a review of single marker analyses and Monte Carlo methods for estimating significance thresholds in a QTL setting see Doerge *et al.* (20). It is expected that ploidy level, marker dosage, and pairing mechanism of homologous chromosomes will add to the genetic specificity that complicates the asymptotic distribution of the test statistic.

Table 1. Fifty simulated progeny from a pseudo–doubled backcross

y	I_M	y	I_M	y	I_M	y	I_M	y	I_M
4.190	1	4.070	1	3.387	1	−0.156	1	−0.448	0
−0.324	0	2.091	1	5.753	1	−1.146	1	6.886	1
4.620	1	2.367	0	2.541	1	−1.483	1	3.446	1
1.286	1	7.638	1	4.866	1	7.718	1	7.662	1
6.795	1	3.098	1	3.185	1	2.156	1	7.808	1
4.542	1	1.218	1	4.967	1	3.309	1	3.449	1
0.480	1	3.674	1	4.146	0	1.338	1	0.212	0
3.864	0	8.481	1	8.249	1	8.130	1	3.389	1
2.404	0	8.417	1	7.424	1	1.069	1	6.855	1
4.380	0	7.875	1	3.890	1	4.439	1	3.856	0

y denotes the quantitative trait, and I_M indicates presence of the marker.

Table 2. Posterior probabilities of parental marker dosage for simulated data

Marker dosage of parent			
1	2	3	4
0.000	0.471	0.512	0.017

Results

A simulation study was performed to assess the power of this model selection procedure. Motivated by an example in sugarcane, an octaploid ($1 \leq d_M, d_Q \leq 4$) was simulated by using the pseudo–modified–doubled backcross, as previously described. For each combination of d_M, d_Q, r, and n (number of progeny), we generated 1,000 data sets which contained the quantitative trait value and the marker genotype for each progeny. The quantitative trait distribution had a common variance of $\sigma^2 = 1.0$ and a mean which depended on the dosage of the QTL. The noninformative parental mean was set to −2.0 and each dose of the QTL increased the mean by 2.0 (additive). We investigated four progeny sizes $n = 50, 100, 200$, and 500 and two recombination rates $r = 0.01$ and 0.25. In total, $16 \times 4 \times 2 = 128$ different parameter combinations were investigated.

For each parameter combination, the percentage of data sets that resulted in the correct d_M and d_Q estimate (Table 4), and the percentage of sets that also correctly estimated the recombination fraction (Table 5) were recorded. Tables 4 and 5 summarize the results for $d_M = 1, 2, 3$, and 4; and $d_Q = 1, 2, 3$, and 4; and $r = 0.01$ and 0.25. We assumed the recombination fraction was estimated successfully when the maximum likelihood estimate was less than 0.05 when the true value was $r = 0.01$ and between 0.125 and 0.375 for true value $r = 0.25$.

Each of the 1,000 simulated data sets per parameter combination and sample size was analyzed, via the procedure described, for the purpose of selecting the best model, and thus formulating the likelihood function. Because the estimation of the dosage level is the limiting factor in the process, we first spend some time considering the effect of dosage estimation on the general

Table 3. Maximum likelihood results for simulated data

d_M	d_Q	$log(L)$	a	μ_0	σ	r
2	1	−96.210	3.906	2.061	4.162	0.0001
2	2	−89.096	3.708	0.082	0.782	0.3291
3	1	−95.612	4.232	2.202	3.616	0.0001
2	3	−96.619	2.573	−0.516	0.894	0.2610
3	2	−89.191	3.697	0.120	0.778	0.3049
3	3	−96.622	2.560	0.247	1.975	0.2765
2	4	−95.408	3.203	−3.459	0.797	0.0417
3	4	−95.976	2.596	−1.135	0.779	0.2589

Table 4. Simulation results for a pseudo–doubled backcross of sample size $n = 50, 100, 200,$ and 500

		Progeny size							
d_M	r	50				100			
1	.01	0.992	0.983	0.956	0.890	0.999	1.000	1.000	0.996
	.25	0.968	0.933	0.910	0.863	0.997	0.995	0.996	0.982
2	.01	0.863	0.955	0.925	0.843	0.958	0.981	0.977	0.972
	.25	0.828	0.839	0.817	0.744	0.936	0.952	0.952	0.944
3	.01	0.664	0.722	0.778	0.815	0.805	0.881	0.903	0.942
	.25	0.616	0.632	0.615	0.722	0.786	0.823	0.813	0.928
4	.01	0.799	0.805	0.827	0.763	0.882	0.908	0.939	0.925
	.25	0.779	0.793	0.751	0.428	0.880	0.891	0.889	0.588
		Progeny size							
d_M	r	200				500			
1	.01	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	.25	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000
2	.01	0.994	0.994	1.000	0.996	1.000	1.000	1.000	1.000
	.25	0.992	0.995	0.990	0.992	1.000	1.000	1.000	1.000
3	.01	0.946	0.963	0.973	0.986	0.994	0.995	0.998	0.998
	.25	0.930	0.926	0.939	0.978	0.993	0.993	0.993	0.997
4	.01	0.955	0.957	0.987	0.964	0.997	0.999	1.000	0.997
	.25	0.942	0.943	0.950	0.803	0.992	0.991	0.994	0.977

Each cell (d_M, r, and n) represents $d_Q = 1, 2, 3,$ and 4, respectively, and contains the percentage of correct dosage level identifications.

Table 5. Simulation results of sample size $n = 50, 100, 200,$ and 500

		Progeny Size							
d_M	r	50				100			
1	.01	0.971	0.909	0.825	0.727	0.995	0.968	0.927	0.857
	.25	0.920	0.771	0.666	0.516	0.991	0.950	0.868	0.795
2	.01	0.742	0.929	0.867	0.733	0.900	0.978	0.951	0.901
	.25	0.583	0.772	0.666	0.553	0.821	0.935	0.902	0.840
3	.01	0.607	0.641	0.751	0.736	0.688	0.833	0.894	0.907
	.25	0.327	0.453	0.518	0.525	0.537	0.723	0.758	0.826
4	.01	0.729	0.713	0.684	0.507	0.806	0.793	0.834	0.735
	.25	0.142	0.357	0.429	0.259	0.405	0.583	0.692	0.474
		Progeny Size							
d_M	r	200				500			
1	.01	1.000	0.993	0.979	0.928	1.000	1.000	0.999	0.983
	.25	1.000	0.994	0.976	0.938	1.000	1.000	1.000	0.999
2	.01	0.967	0.994	0.997	0.975	0.999	1.000	1.000	0.999
	.25	0.949	0.993	0.983	0.974	1.000	1.000	1.000	0.999
3	.01	0.865	0.945	0.973	0.981	0.981	0.993	0.998	0.997
	.25	0.781	0.894	0.930	0.950	0.968	0.991	0.993	0.997
4	.01	0.841	0.893	0.951	0.881	0.944	0.979	0.996	0.986
	.25	0.619	0.774	0.882	0.776	0.868	0.966	0.988	0.977

Each cell (d_M, r, and n) represents $d_Q = 1, 2, 3,$ and 4, respectively, and contains the percentage of correct dosage level and recombination fraction identifications.

process of model selection. For all marker dosage, d_M, and QTL dosage, d_Q, combinations, the probability of correctly identifying the dosage levels was 97% or higher when $n = 500$ and 80% or higher when $n = 200$ (Table 4). When in fact the sample size is 50 or 100 our ability to correctly estimate dosage of marker and/or QTL greatly decreased as the dosage level of both marker and QTL increase. This point emphasizes the importance of sample size when mapping in polyploids. If one is going to rely on multiple dose markers and multiple dose QTL, large sample sizes must be employed. In general, as the dosage level of the marker increases, a corresponding doubling of the sample size maintains the same level of power to detect the correct model. In this simulation, when $d_M = 4$, there was some increase in power over $d_M = 3$ strictly because only models with $d_M \leq 4$ were considered (border effect). In situations where the dosage levels were not identified correctly, there was a tendency to overestimate both d_M and d_Q, with the QTL dosage more likely to be identified correctly. This overestimation can largely be attributed to the fact that $p_{d_M} = (1/2)^{d_M}$. For a given d_M, p_{d_M+1} is much closer to p_{d_M} than p_{d_M-1}. Lastly, as the distance or recombination, r, increases between the QTL and marker, the probability of correctly identifying the dosage levels decreases.

When the motivation for model selection in polyploids is to test for QTL detection and/or location, the estimate of recombination when coupled with an appropriate map function will supply a relational distance between the marker and QTL (i.e., how far the QTL is from the marker). As with all maximum likelihood estimation, estimates of r tend to be underestimated when the sample sizes are small, and in polyploids this situation is even more pronounced when $d_M \gg d_Q$, and when the linkage is weak ($r = 0.35$) (not shown). When sample sizes increase, the power to estimate r correctly is greater when, in fact, $d_Q \geq d_M$. As is the case in this simulation, preferential pairing ensures that each informative chromosome from the informative parent is paired with a null chromosome, and as a result, only chromosomes which contain both a marker and QTL provide information on recombination. When the QTL and marker dosage levels are unequal, there will be some chromosomes containing just an M or Q, and thus provide no information about r. Unequal dosage levels for the QTL and marker can even mask recombination, the effect of which is even more severe when there are additional copies of the marker (i.e., increased marker dosage) since d_Q is observed in the quantitative trait distribution means. Lastly, as the linkage between the marker and QTL weakens (i.e., the QTL is further in location from the marker), regardless of marker and/or QTL dosage, the power to estimate r decreases dramatically.

Discussion

Model selection for QTL analysis using a single marker has been presented for a pseudo–doubled backcross polyploid organism demonstrating preferential pairing during meiosis. Clearly, the assumption of preferential pairing and known ploidy level affects the power by increasing or decreasing the number of potential models. Thus, for a polyploid with a smaller ploidy, the power for all possible parameter configurations will be higher than what has been described. When the assumption of preferential pairing is lifted to accommodate random pairing, the results may be very different in that, the ploidy level not only alters the number of potential models, it can also affect the probability of an informative pairing. Extensions to include this work are in progress.

Given our mating design and simulation, we assumed an additive QTL mean model with the effect of the QTL being a

single value, and a variance of 1.0. In doing so, we realize that we have limited our simulation space, and for completion, a range of QTL effects, along with varying variance parameter values must be considered. We fully expect the statistical power of what we described to be affected as both QTL effect and variance change. Clearly, as the QTL dose means become more disparate it will be easier to estimate the correct dosage of the QTL. Additionally, our model selection process is simplified because the number of parameters for each configuration is the same. A more flexible approach is to use only an order restriction on the means. In other words $\mu_0 < \mu_1 < ... < \mu_{d_Q}$, where the subscript represents the dosage of the QTL; however, this alters the number of parameters in each configuration. If a non-additive model is employed, a model selection criterion such as the BIC (23) could be used to select the model.

As demonstrated by Ripol *et al.* (12), placing multiple dose markers on an existing framework of single dose markers allows the estimation of a genetic map for any polyploid. As shown in many diploid studies, given that a genetic map exists, the genetic distances between markers can easily be exploited for the purpose of QTL mapping by using interval mapping methodology (3). The limiting factor in extending what has been successful in diploid QTL mapping, to what needs to be done in polyploids, has been the development of models that reflect the polyploid nature of more complex organisms. Our goal in this paper has been to describe all the tools necessary to investigate QTL mapping in polyploids by initiating the simplest situation of single marker QTL mapping, and setting the stage for interval mapping or composite interval mapping (6, 7). We anticipate that (composite) interval mapping will present an entirely new set of challenges when coupled with the complexities of random pairing and non-additive models.

Finally, in addition to the particularities of the polyploidy and the complications that arise in attempts to model it for QTL mapping, questions with regard to linkage between markers and QTL arise. These questions have great potential to further our understanding of genome organization within and between species, as well as provide us with an evolutionary time line for polyploidization. Some of these questions are: if a molecular marker is found to be tightly linked to a QTL, should the dosage of the marker agree with the dosage of the QTL? In which situations is the linkage more strongly affected? Should the models which are controlled by dosage levels be weighted for the purpose of representing more realistic results? Would models with dosage levels more similar to each other be more likely, especially with close linkage? Answers to these questions may aid in our understanding of the genetics, evolution, and comparative organization between well mapped diploids and sparsely investigated polyploids. QTL mapping in polyploids may enable us to create links between evolutionarily related species, many of which are diploid, which in turn will allow us to broaden our understanding of genetically diverse and distantly related species.

We thank Maria Ripol, and Drs. George McCabe, Bruno Sobral, and Mark Sorrells, and an anonymous reviewer for providing useful publications, comments, and discussions. Additionally, we thank Dr. Christie Williams for a detailed review of this manuscript, helpful comments, and invaluable discussions on the complexities of polyploidy.

1. Sax, K. (1923) *Genetics* **8,** 552–560.
2. Soller, M., Brody, T. & Genizi, A. (1976) *Theor. Appl. Genet.* **47,** 35–39.
3. Lander, E. S. & Botstein, D. (1989) *Genetics* **121,** 185–199.
4. Haley, C. S. & Knott, S. (1992) *Heredity* **69,** 315–324.
5. Martínez, O. & Curnow, R. N. (1992) *Theor. Appl. Genet.* **85,** 480–488.
6. Zeng, Z.-B. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 10972–10976.
7. Zeng, Z.-B. (1994) *Genetics* **136,** 1457–1468.
8. Whittaker, J. C., Thompson, R. & Visscher, P. M. (1996) *Heredity* **77,** 23–32.
9. Hieter, P. & Griffiths, T. (1999) *Science* **285,** 210–211.
10. Wu, K. K., Burnquist, W., Sorrells, M. E., Tew, T. L., Moore, P. H. & Tanksley, S. D. (1992) *Theor. Appl. Genet.* **83,** 294–300.
11. da Silva, J. A. G. & Sorrells, M. E. (1996) *Methods of Genome Analysis in Plants* (CRC, New York).
12. Ripol, M. I., Churchill, G. A., da Silva, J. A. G. & Sorrells, M. (1999) *Gene* **235** (1–2), 31–41.
13. Guimaraes, C. T. & Sobral, B. W. (1999) *Plant Breeding Reviews* **16,** 1–13.
14. Sills, G. R., Bridges, W., Aljanabi, S. M. & Sobral, B. W. S. (1995) *Molecular Breeding* **1**(4), 355–363.
15. Guimaraes, C. T., Sills, G. R. & Sobral, B. W. S. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 14261–14266.
16. Galitski, T., Saldanha, A. J., Styles, C. A., Lander, E. S. & Fink, G. R. (1999) *Science* **285,** 251–254.
17. Soltis, D. E. & Soltis, P. S. (1999) *Trends Ecol. Evol.* **14,** 348–352.
18. Grattapaglia, D. & Sederoff, R. (1994) *Genetics* **137,** 1121–1137.
19. da Silva, J. A. G. & Sobral, B. W. S. (1996) in *The Impact of Plant Molecular Genetics*, ed. Sobral, B. W. S. (Birkhäuser, Boston), pp. 3–38.
20. Doerge, R. W., Zeng, Z.-B. & Weir, B. S. (1997) *Stat. Sci.* **12**(3), 195–219.
21. Degroot, M. H. (1986) *Probability and Statistics* (Addison-Wesley, Reading, MA).
22. Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) *J. R. Stat. Soc.* **39,** 1–38.
23. Kass, R. E. & Raftery, A. E. (1995) *J. Am. Stat. Soc.* **90,** 773–795.