AN INTRODUCTION TO MODEL SELECTION FOR QUANTITATIVE TRAIT LOCUS ANALYSIS IN POLYPLOIDS

R. W. Doerge

Bruce A. Craig

AN INTRODUCTION TO MODEL SELECTION FOR QUANTITATIVE TRAIT LOCUS ANALYSIS IN POLYPLOIDS

R.W. Doerge[1],[2],[3] and Bruce A. Craig [3]

[1] Departments of Agronomy, Purdue University, West Lafayette, IN 47907

[2] Computational Genomics, Purdue University, West Lafayette, IN 47907

[3]Department of Statistics, Purdue University, West Lafayette, IN 47907

Abstract

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 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 (two homologous chromosomes per set), even though many important agricultural crops are, in fact, polyploid (more than two homologous chromosomes per set). The purpose of our work is to initiate an algorithmic approach for model selection and QTL detection in polyploid species. This approach involves the enumeration 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 as being extendible to many breeding schemes and less restricted settings.

1 Introduction

Detecting and locating genomic regions associated with quantitative traits is known as quantitative trait locus (QTL) mapping. The statistical methods (for review see Doerge *et al.*, 1997) employed to identify QTL are numerous, and rely heavily on the fact that the organism is diploid (*i.e.*, has homologous pairs of chromosomes). In the framework of QTL analysis, 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 usual Mendelian manner.

When there are more than two homologous chromosomes per set, the species is referred to as polyploid. While most animal species are diploid, many important agricultural crops such as sugarcane, cotton, banana, alfalfa, potato, coffee, and wheat are polyploid. Among natural species of flowering plants, nearly half are polyploid (Hieter and Griffiths, 1999). Even in animals, polyploidy exists. Salmonid fish and specific amphibians display doubling and tripling of their ploidy level (Hieter and Griffiths, 1999).

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 situations, such as sugarcane, there is no closely related diploid species making QTL analysis 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 number 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. 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 algorithm for model selection process to be used in a single marker analysis for QTL detection.

1.1 Characteristics of Polyploidy

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). 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 used throughout this paper, is that meiosis is simply an extension of the diploid case and involves multiple pairings of homologous chromosomes. During polyploid meiosis, pairs of chromosomes in each homologous set align and possibly exchange genetic material (*i.e.*, crossover). Each 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, while species displaying random pairing are referred to as autopolyploids. Our work will be based on a preferential pairing mechanism, thereby reducing the complexity of polyploidy to essentially that of a diploid.

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 dose of each marker and/or QTL. For the polyploid situation, as many as $\frac{k}{2}$ copies of a genetic marker and/or QTL can be passed to the gamete.

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.* (1992) derived a theoretical approach for mapping single dose DNA markers in polyploids under the assumption of random pairing. Ripol *et al.* (1997) 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 (1996) and Guimaraes and Sobral (1999) pointed out that the use of multiple dose markers improves the accuracy of detection of pairing homologs and their organization into homology groups. Wu *et al.* (1992), da Silva and Sorrells (1996), Ripol *et al.* (1997), and Guimaraes and Sobral (1999) have each

made a 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 made by Sills *et al.* (1995), and later extended by Guimaraes *et al.* (1997). In those studies various agronomically important traits were associated with single dose markers by means of multiple regression model building and maximum likelihood methods. In these QTL analyses the model used to develop the likelihood function was limited to single dose markers. To date, no attempt has been made to employ multiple dose markers for QTL analyses.

2 Model Selection for QTL Analysis in Polyploids

2.1 The Experimental Model

Let us consider a pseudo–doubled backcross population (Grattapaglia and Sederoff, 1994) that is the result of selecting an informative parent, doubling half of its chromosomes to create a non-informative parent, and then crossing the two parental lines (Figure 1) so that pseudo–double backcross progeny result (Grattapaglia and Sederoff, 1994; da Silva and Sobral, 1996). It is important to realize that the informative parent's genetic constitution (*i.e.*, dosages) is not known, but may later be inferred from the pseudo-backcross progeny. For our purposes, we assume the non-informative 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.

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 single, and quantitative traits are scored. The statistical methodology for doing single marker QTL analysis includes t-tests, regression, and likelihood ratio tests. 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. With the diploid meiotic process (*e.g.*, chromosomal pairing, crossing over, gametic probabilities) well understood, the likelihood function is easily stated as a function of probability distributions of marker genotype classification, 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.

Similar to the diploid QTL analysis, we assume only two alleles at each marker and QTL, and denote a molecular marker by M and a QTL by Q. Since 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 M present), Q (only QTL present), and $\emptyset$ (neither M nor Q present). The number of each type of chromosome will depend on the ploidy level.

2.2 All Possible Polyploid Models

In the diploid setting, 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 this process down, 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

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

The rows and columns of the matrix $\mathbf{C}$ represent the possible dosage levels of the marker and QTL, respectively. The elements of the matrix $\mathbf{C}$ are probabilities that depend on the configuration of the paired chromosomes, and thus are functions of the recombination fraction r.

Since there are $\frac{k}{2}$ pairs of chromosomes in each homologous set, the probabilities of the overall contribution are a function of $\frac{k}{2}$ $\mathbf{C}$ matrices. Because each pair, i, independently contributes one

chromosome to a gamete, the Kronecker product of the $\frac{k}{2}$ $\mathbf{C}_i$ matrices yields a $2^{\frac{k}{2}} \times 2^{\frac{k}{2}}$ probability matrix of each order-specific contribution. Since we are solely interested in the overall contribution, we simplify this matrix such that its rows and columns represent the gamete's possible dosage levels for the genetic marker and QTL. Each chromosomal pair can contribute at most one copy of the marker and QTL, therefore 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

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

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

$$\mathbf{G} = \mathbf{A}_{\frac{k}{2}}^T(\ldots(\mathbf{A}_3^T((\mathbf{A}_2^T((\mathbf{A}_1^T\mathbf{C}_1\mathbf{A}_1)\otimes\mathbf{C}_2)\mathbf{A}_2)\otimes\mathbf{C}_3)\mathbf{A}_3)\otimes\ldots\otimes\mathbf{C}_{\frac{k}{2}})\mathbf{A}_{\frac{k}{2}}.$$

Preferential pairing, like diploids, allows the most straightforward calculations as there is only one set of chromosomal pairs, or equivalently one set of $\{\mathbf{C}_i\}$, so the matrix $\mathbf{G}$ represents the gametic probabilities for specific ploidy and dosage levels. However, when pairing is random, more than one set of chromosomal pairs is possible, and the gametic probabilities for all configurations are more extensive. For each set of $\{\mathbf{C}_i\}$, 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 the set of chromosomal pairs occurs, and summing these matrices together.

With the 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.

2.3 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 copies. This conditional probability is a result of our pseudo–doubled backcross design and our assumption of preferential pairing. Under a random pairing situation, the procedure would follow similarly, except $p_{d_M} = \binom{k-d_M}{k/2}/\binom{k}{k/2}$.

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 ($d = 1, \ldots, \frac{k}{2}$) is equally likely and compute the posterior probability of each dosage level, $\Pr(d_M|n, n_\emptyset) =$
$\text{Bin}(n_\emptyset; n, p_{d_M}) / \sum_{d=1}^{k/2} \text{Bin}(n_\emptyset; n, p_d)$.

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

2.4 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 to select the single best model that will in turn provide 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 provide for $\frac{k}{2}+1$ possible

phenotypic means. For example, using $I_{m,i}$ to indicate presence of the marker (1, if individual i has the marker and 0, otherwise), and denoting y_i as the trait values for individual i, the likelihood is

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

where $j = 1, \ldots, k/2$ represents the range of dosage for the QTL, $\mathrm{P}(Q_j)$ is the gametic probability of no marker and j copies of the QTL, and $\mathrm{P}(MQ_j)$ is the gametic probability of at least one copy of the marker and j copies of the QTL. For an additive dosage effect, the mean of the quantitative trait distribution for a specified QTL dosage is $\mu_j = \mu_2 + j\alpha$, where μ_2 is the mean of the non-informative parent and α is the additive contribution for a single dosage of the QTL. The variance, σ^2, is assumed equal in both parents, but could easily be considered as two separate parameters. Utilization of the EM-algorithm (Dempster, Laird, and Rubin, 1977) maximizes the likelihood function in a fashion similar to the diploid situation, only now the expectation step (the E-step), involves a multinomial rather than binomial distribution.

3 Simulated Example Demonstrating the Model Selection Process

As an example of the model selection process, we detail the algorithmic approach using simulated data for 50 progeny (Table 1) of an octaploid. The informative parent was double coupled (2 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.

3.1 Steps Involved in the Model Selection Process

3.1.1 Estimating the Marker Dosage

For this data set, 41 of the 50 progeny had at least one copy of the marker. The posterior probability of marker dosage 1 through 4 is 0.000, 0.471, 0.512, 0.017, respectively. Recall that the expected fraction of progeny with no copies of the marker, when the parental dosage is d_M, is $(1/2)^{d_M}$. Since 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.

3.1.2 Computing the Likelihood For Each Model

Standard EM–algorithm (Dempster, Laird, and Rubin, 1977) procedures were used to compute the maximum likelihood estimates for each of the models considered. The likelihood values and parameter estimates were calculated (Table 2). For these data, the model ($d_M = 2$, $d_Q = 2$) had the highest likelihood so it would be selected as the model. This configuration is just slightly better than the model ($d_M = 3$, $d_Q = 2$), with little difference in the parameter estimates.

4 Single Marker QTL Analysis in Polyploids

As demonstrated, the real challenge arising from polyploidy is not the QTL analysis itself, but rather the model on 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 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.* (1997). Ploidy level, marker dosage, and pairing mechanism of homologous chromosomes are expected to add to the genetic specificity that complicates the asymptotic distribution of the test statistic.

5 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 using the pseudo–modified–doubled backcross. For each combination of d_M, d_Q, r, and n (number of progeny), we generated 1000 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 non-informative 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 three recombination rates $r = 0.01, 0.25$, and 0.35. In total, $16 \times 4 \times 3 = 192$ different parameter combinations were investigated.

Each of the 1000 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. Since the estimation of the dosage level is the limiting factor in the process, we first consider the effect of dosage estimation on the general 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 $n = 200$. When in fact the sample size was 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 increased. This result 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 $d_M >> d_Q$, and when the linkage is weak ($r = 0.35$). 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 masking 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 farther from the marker), regardless of marker and/or QTL dosage, the power to estimate r decreases dramatically.

6 Summary

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 assumptions of preferential pairing and known ploidy level affect 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, but can also affect the probability of an informative pairing. Extensions to include that case are in progress.

With 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 completeness, 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 (Kass and Raftery, 1995) could be used to select the model.

As demonstrated by Ripol *et al.* (1997) 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. The limiting factor in extending to polyploids what has been successful in diploid QTL mapping, has been the development of models which reflect the polyploid nature of more complex organisms. Our goal in this paper has been to describe the tools necessary to investigate QTL mapping in polyploids by beginning with the simplest situation of single marker QTL mapping, and setting the stage for more advanced investigations. Currently, we are extending this algorithm to interval mapping which involves two markers and a single QTL. While the steps are similar, the increased number of potential models makes this a more complex problem.

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. Exploration of these questions have great potential to further our understanding of genome organization within and between species, as well as to 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 what 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 understanding of the genetics, evolution, and comparative organization between well mapped diploids and sparsely investigated polyploids. Mapping of QTL 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.

7 Cited Literature

da Silva, J.A.G. and Sorrells, M.E. 1996. Methods of Genome Analysis in Plants. in The Impact of Plant Molecular Genetics, B.W.S. Sobral ed, Boston: Birkhäuser, 3–38. CRC Press, New York.

da Silva, J.A.G. and Sobral, B.W.S. 1996. Genetics of Polyploids in The Impact of Plant Molecular Genetics. B.W.S. Sobral ed, Boston: Birkhäuser, 3–38.

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

Doerge, R.W., Zeng, Z-B, and Weir, B.S. 1997. Statistical issues in the search for genes affecting quantitative traits in experimental populations. Statistical Science. 12(3):195–219.

Grattapaglia, D. and Sederoff, R. 1994. Linkage maps of *eucalyptus-grandis* and *eucalyptus-urophylla* using a pseudo-testcross-mapping strategy and rapid markers. Genetics. 137(4):1121–1137.

Guimaraes, C.T., Sills, G.R., and Sobral, B.W.S. (1997) Comparative mapping of *Andropogoneae: Saccharum L.* (sugarcane) and its relation to sorghum and maize', *Proceedings of the National Academy of Science USA*, 94(26):14261–14266.

Guimaraes, C.T. and Sobral, B.W. 1999. Plant Breeding Reviews. The *Saccharum* Complex and its relation to other andropogoneae, Plant Breeding Reviews, Vol 16.

Hieter, P. and Griffiths, T. 1999. Polyploidy –More is More or Less. Science 285:210–211.

Kass, R.E. and Raftery, A.E. 1995. Bayes Factors. J. the Amer. Stat. Soc. 90:773 –795.

Ripol, M.I., Churchill, G.A., da Silva, J.A.G., and Sorrells, M. 1997. Statistical aspects of genetic mapping in autopolyploids. Gene-Combis.

Sills, G.R., Bridges, W., Aljanabi, S.M. and Sobral, B.W.S. 1995. Analysis of agronomic traits in a cross between sugarcane (*Saccharum - Officinarium L.*) and its presumed progenitor (*S-Robustum Brandes* and *Jesw Ex Grassl*) Molecular Breeding 1(4):355-363.

Wu, K.K., Burnquist, W., Sorrells, M.E., Tew, T.L., Moore, P.H., and Tanksley, S.D. 1992. The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theor. and Appl. Genet. 83:294–300.

Table 1: Simulated octaploid progeny ($n = 50$) from a pseudo–doubled backcross where y denotes the quantitative traits, and I_M indicates presence of the marker.

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

Table 2: Maximum likelihood results for simulated data where d_M is the dosage of the marker, d_Q is the dosage of the QTL, L is the likelihood described in the text, a is the additive effect of the respective QTL dose, μ_0 is the mean of the quantitative trait when the QTL dosage is 0.0, σ is the standard deviation of the quantitative trait distribution, and r is the recombination fraction between the marker and the QTL.

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

Figure 1: A pseudo-doubled backcross experimental mating design which is the result of selecting an informative parent, doubling half of its chromosomes to create a non-informative parent, then crossing the two parental lines to produce a pseudo-doubled backcross population.