INTERVAL MAPPING FOR AUTOPOLYPLOIDS

Dachuang Cao

Bruce A. Craig

R. W. Doerge

Recommended Citation

Cao, Dachuang; Craig, Bruce A.; and Doerge, R. W. (2003). "INTERVAL MAPPING FOR AUTOPOLYPLOIDS," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1175

Interval mapping for autopolyploids

Dachuang Cao[1], Bruce A. Craig[1] & R.W. Doerge[1,2,3]

[1] Department of Statistics, Purdue University, West Lafayette, IN 47907
[2] Computational Genomics, Purdue University, West Lafayette, IN 47907
[3] Department of Agronomy, Purdue University, West Lafayette, IN 47907

Abstract

While extensive progress has been made in quantitative trait locus (QTL) mapping of diploid species, the progress of QTL mapping in polyploids has been limited due to the polyploid's complex genetic architecture. To date, QTL mapping in polyploids has focused primarily on tetraploids with dominant markers and/or codominant markers. In this paper, we extend the interval mapping methodology to any autopolyploid of even ploidy level. Our approach selects a set of likely parental chromosomal configurations (models) using a Bayesian model reduction step. The EM algorithm is then employed to estimate each model's parameters including QTL location, marker dosages, QTL dosages, and the trait effect.

Key words: autopolyploid, QTL mapping, interval mapping,

1. Introduction

QTL (quantitative trait locus/loci) mapping detects and identifies regions of a genome associated with the variation of a quantitative trait of interest. Molecular markers have been used extensively to construct genetic maps for diploid species (Koornneef et al. 1983; Dietrich et al. 1996), and act as the foundation for further QTL analysis. Based on genetic maps, many statistical methods have been developed for QTL mapping, namely interval mapping (Lander and Botstein1989), composite interval mapping (Zeng 1993, 1994), and multiple QTL mapping (Jansen 1993). The statistical issues involved in QTL mapping are reviewed in Doerge (2002).

Polyploids are organisms having more than two complete sets of chromosomes (genomes) in a cell. Polyploidy is most common in plants, especially in agriculture plants, and is found in some insects, amphibians, and reptiles. It is also an important evolutionary force that is the basis of many investigations (Soltis and Soltis 2000; Osborn et al. 2003). Due to the different approaches of current QTL mapping methods for polyploids, we need to classify polyploids according to the homology between genomes. A polyploid with genomes all derived from the same species is called an autopolyploid. Otherwise, if the multiple sets of chromosomes are derived from different species, the polyploid is called an allopolyploid.

For allopolyploids, such as bread wheat and potato, meiotic pairing is restricted in ancestral parental homologues; therefore, diploid QTL mapping methods can be utilized.

However, for autopolyploids the high homology between the genomes creates additional complexities in the meiotic process. Specifically, autopolyploids may undergo either bivalent pairing (two homologs pair) or multivalent pairing (more than two homologues pair), and it varies in different species (Rieseberg and Doyle 1989; Sybenga 1995). Furthermore, the manner in which the paired chromosomes segregate during meiosis, especially for multivalent pairing, also varies among species (Jackson and Jackson 1996). Finally, the number of alleles for each locus, how many copies of each allele, and the linkage phase between loci for the parents and the progeny are unknown.

For autopolyploids, the multiple copies of markers or QTL may be biallelic or multiallelic. Here, we assume they are biallelic (or dominant) loci, which holds for experiments based on doubled-haploids (Guha and Maheshwari 1964), such as pseudo-doubled backcross experiments (Grattapaglia and Sederoff 1994) in inbred populations. Wu et al. (1992) proposed the method of estimating a genetic map for autopolyploids with simplex markers, or single dose restriction fragment (SDRF) markers, which represent only one homologue and segregates 1:1 in the progeny. Ripol et al. (1999) extended the method of Wu et al. (1992) to any dominant marker with an unobservable dosage level by first estimating marker dosages and linkage phase, and then constructing a genetic map by computing the maximum likelihood estimates of recombination based on estimated parental marker configurations. Doerge and Craig (2000) developed an algorithmic model selection process for a single marker QTL analysis with dominant markers for autopolyploids with any even ploidy level.

As a continuation of the work by Doerge and Craig (2000), we propose a maximum likelihood based interval mapping method using available genetic maps to increase the power of detecting and estimating QTL locations within an autopolyploid bivalent pairing framework. Our work is based on a pseudo-doubled backcross experiment (Grattapaglia and Sederoff 1994) and employs model selection for interval mapping to simultaneously estimate model parameters including QTL location, parental marker and QTL dosages, and the QTL effect given marker presence/absence data and quantitative trait data for the progeny. We first estimate the parental marker configuration (i.e., marker dosages and their arrangements) to reduce the number of potential parental configurations (i.e., models). Based on each putative parental configuration, interval mapping is used to estimate QTL location and QTL effect. We limit our approach to even ploidy levels with multiple-dose dominant markers since odd ploidy levels are often highly infertile. After describing the methodology, simulation studies are presented to investigate how QTL or marker dosages and their linkage affect the performance of our algorithm.

2. Method

An example of a pseudo-doubled backcross experiment for a tetraploid with two markers is shown in Figure 1. In a pseudo-doubled backcross experiment, after an informative parent P_1 is selected, half of its chromosomes are doubled to create a non-informative parent P_2, and the progeny F_1 is produced by crossing P_1 and P_2 under the following assumptions. First, by definition, the informative parent has at least one, and at most, half the ploidy dose of the dominant allele at each locus. Second, the pairing mechanism in the meiosis process is either preferential pairing (select homologs always pair together) or random pairing (equally likely to pair with each homologue). In this work, the examples and simulation studies will be presented using preferential pairing mechanism, in which each informative homologue always pairs with a non-informative homologue. Last, assuming an additive QTL effect, the trait Y has a normal distribution with mean $\mu_j = ja + b$, where j is the dosage of the QTL, and common variance σ^2. Let $\theta = (a, b, \sigma)$ denote the vector of model parameters, n denote the number of progeny, m denote the number of markers, and k the ploidy level. In what follows, for each locus the upper case is used to denote both the locus name and its dominant allele, and the lower case stands for the recessive allele (e.g., A and a). The dosage of a locus denotes the dosage of the dominant allele at that locus. When a marker is present in an individual, at least one dose of the dominant allele for that marker is observed.

2.1. Interval mapping

Given a genetic map (i.e., the recombination fractions or genetic distances between the markers) and a parental configuration C, interval mapping can be applied to estimate $\theta = (a, b, \sigma)$. A parental configuration includes the dosage at each locus and the linkage phase between the loci. Consider two markers M_1 and M_2 $(m = 2)$, and one QTL, Q. Let $x = (x_1, x_2, \ldots, x_n)$ denote all the observable data. For the i^{th} individual, $x_i = (y_i, o_i)$, where y_i is the trait value, and $o_i = (I_i^{M_1}, I_i^{M_2})$ are the marker presence/absence indicators with $I_i^{M_h} = 1$ if marker M_h is present and 0 otherwise. With a fixed putative position of the QTL and parental configuration C, the likelihood function of the trait is a mixture of normal distributions,

$$L(\theta \mid x, C) = \prod_{i=1}^{n} \sum_{j=0}^{k/2} P(Q^j \mid I_i^{M_1}, I_i^{M_2}, C) \phi(y_i; \mu_j = ja + b, \sigma^2), \qquad (1)$$

where $P(Q^j \mid I_i^{M_1}, I_i^{M_2}, C)$ is the probability of the i^{th} progeny having j copies of the QTL with marker presence status $\{ I_i^{M_1}, I_i^{M_2} \}$, and $\phi(y_i; \mu_j = ja + b, \sigma^2)$ is the normal density function valued at y_i with mean $\mu_j = ja + b$ and variance σ^2. The $P(Q^j \mid I_i^{M_1}, I_i^{M_2}, C)$ is a function of the recombination fractions between the QTL and markers, and is a known quantity.

In the interval mapping framework, the putative QTL position is fixed at incremental positions between the flanking markers for the purpose of evaluating a test statistic (commonly, the log likelihood ratio). At each evaluation position, let r_h^a $(h = 1, 2)$ denote the recombination fractions between marker M_h and the QTL, where the superscript a stands for alternative hypothesis. To test the hypothesis

$$H_0 : r_1 = r_2 = 0.5 \qquad \text{vs.} \qquad H_a : r_1 = r_1^a, r_2 = r_2^a$$

where the null hypothesis assumes that the QTL is present, but unlinked to both markers, and the alternative hypothesis assumes that the QTL is present and linked to M_1 and M_2 with recombination fractions less than or equal to r_h^a, the log likelihood ratio test statistic LRT is

$$LRT = -2 \ln \frac{L(\hat{\theta}_0, r_1 = r_2 = 0.5)}{L(\hat{\theta}_a, r_1 = r_1^a, r_2 = r_2^a)},$$

with $\hat{\theta}_0$ and $\hat{\theta}_a$ representing the estimated parameters under the null and alternative hypotheses, respectively. A permutation test can be performed to estimate the significance threshold for the test statistic (Churchill and Doerge 1994}. If the test statistic is significant then the corresponding position with the largest *LRT* statistic is referred to as the estimated QTL position, with the corresponding maximum likelihood estimate (MLE) of θ at this position.

Due to the summation over QTL dosages in the likelihood function (1), the EM algorithm (Dempster et al. 1977) is employed to estimate model parameters. The complete likelihood function is

$$f(\theta \mid x, C) = \prod_{i=1}^{n} \prod_{j=0}^{k/2} [P(Q^j \mid I_i^{M_1}, I_i^{M_2}, C)\phi(y_i; \mu_j = ja + b, \sigma^2)]^{z_{ij}} . \qquad (2)$$

Here, the unobservable data are the individual QTL dosages. For the i^{th} individual, define $z_i = (z_{io}, z_{i1}, \ldots, z_{i,k/2}), i = 1,2,\ldots,n,$ be such that z_{ij} is an indicator variable for the QTL dosage of j for the i^{th} progeny. To implement the EM algorithm, the z_{ij} are estimated in the E step by

$$\hat{z}_{ij} = \hat{E}[z_{ij}] = P_\theta(d_Q = j \mid y_{i,} I_i^{M_1}, I_i^{M_2}, C) = \frac{p_j \exp\{-0.5(\frac{y_i - j\hat{a} - \hat{b}}{\hat{\sigma}})^2\}}{\sum_{l=0}^{k/2} p_l \exp\{-0.5(\frac{y_i - l\hat{a} - \hat{b}}{\hat{\sigma}})^2\}},$$

where $p_j = P_\theta(d_Q = j \mid y_{i,} I_i^{M_1}, I_i^{M_2}, C)$. Given $\hat{z}_{ij}$, the MLE of $\theta = (a, b, \sigma)$ is

$$\hat{a} = \frac{n\sum_{i=1}^{n}\sum_{j=0}^{k/2} \hat{z}_{ij} \times j \times y_i - \sum_{i=1}^{n}\sum_{j=0}^{k/2} \hat{z}_{ij} \times y_i \sum_{i=1}^{n}\sum_{j=0}^{k/2} \hat{z}_{ij} \times j}{n\sum_{i=1}^{n}\sum_{j=0}^{k/2} \hat{z}_{ij} \times j^2 - (\sum_{i=1}^{n}\sum_{j=0}^{k/2} \hat{z}_{ij} \times j)^2},$$

$$\hat{b} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{k/2} \hat{z}_{ij}(y_i - j\hat{a})^2,$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{k/2} \hat{z}_{ij}(y_i - j\hat{a} - \hat{b})^2.$$

The E step and M step are iterated until a convergence criterion is satisfied.

As stated before, the parental configuration is unknown. If each possible parental configuration is viewed as a potential model, the model space tends to expand quickly as the loci number and ploidy level increase. For example, under a pseudo-doubled backcross experiment with only two markers, there are *14* possible parental configurations for a tetraploid, *91* for a hexaploid, and *390* for an octaploid. To limit the model search, a model reduction step is implemented.

2.2. Model Reduction

To reduce the model space, one can consider either a marginal method (1) or joint method (2). With method 1, we calculate the posterior probabilities of dosages for each marker individually assuming that all the possible dosages are equally likely *a priori*. If one particular dosage level has a posterior probability higher than a specified cutoff, only that marker dosage is considered in candidate parental configurations. Otherwise, we select the most likely dosage levels until the sum of posterior probabilities exceeds the cutoff and these dosages are used to form the candidate parental configuration set. With method 2, a similar approach is used but information on both markers is considered jointly. We can directly calculate parental marker configuration posterior probabilities and choose the candidates following the same rule as above.

2.2.1. Calculate parental marker dosage posterior probability

The posterior probabilities for parental marker dosages of each marker are calculated using the Bayes' rule based on the marginal marker presence/absence distribution (Ripol et al. 1999). This method will be referred to as the binomial method. For one marker M, the number of progeny with M absent, n_{null} has a binomial distribution $Bin\ (n,\ p_{d_M})$, where d_M denotes the marker dosage of M in the informative parent. Under the pseudo-doubled backcross experiment, $p_{d_M} = (0.5)^{d_M}$ for our preferential pairing system and $p_{d_M} = \binom{k-d_M}{k/2} \Big/ \binom{k}{k/2}$ for a random pairing system.

Thus, given the informative parent marker dosage d_M, the chance of observing n_{null} is

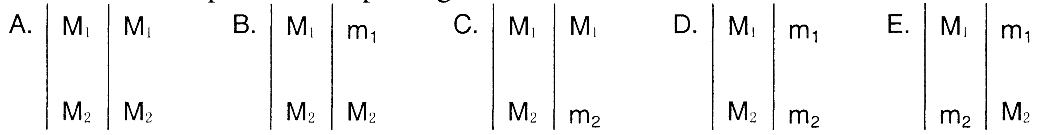
$$P(n_{null} \mid n, d_M) = \binom{n}{n_{null}} p_{d_M}{}^{n_{null}} (1 - p_{d_M})^{n-n_{null}},$$

Based on a discrete uniform prior on d_M, the posterior probability of each dosage level is

$$P(d_M \mid n_{null}, n) = \frac{p_{d_M}{}^{n_{null}} (1 - p_{d_M})^{n-n_{null}}}{\sum_{d=1}^{k/2} p_d{}^{n_{null}} (1 - p_d)^{n-n_{null}}},$$

2.2.2. Calculate parental marker configuration posterior probability

To calculate the posterior probability of parental marker configurations, we can use the joint marker information and a genetic map. This method will be referred to as the multinomial method. Suppose we have markers M_1 and M_2. All progeny can be classified into four sets according to the presence status of the two markers. Let $n_{obs} = (n_{00}, n_{01}, n_{10}, n_{11})$ stand for the observed frequency vector in the four sets, where the h^th^, ($h = 1, 2$) subscript is 1 if M_h is present, and 0 otherwise. Assuming no segregation distortion and a parental configuration C, n_{obs} follows a multinomial distribution with probability parameter vector $p^c = (p_{00}{}^c, p_{01}{}^c, p_{10}{}^c, p_{11}{}^c)$. As an example, p^c is listed in Table 1 for the five parental marker configurations (A--E) for a tetraploid with two markers under preferential pairing.



The probability of observing $n_{obs} = (n_{00}, n_{01}, n_{10}, n_{11})$ is

$$P(n_{obs} \mid n, C) = \frac{n!}{\prod_{i,j=0}^{1} n_{ij}!} \prod_{i,j=0}^{1} (p_{ij}^{\;c})^{n_{ij}},$$

Thus, the posterior probability for a parental marker configuration C_0 is

$$P(C_0 \mid n_{obs}, n) = \frac{\prod_{i,j=0}^{1} (p_{ij}^{\;c_0})^{n_{ij}}}{\sum_{C} \prod_{i,j=0}^{1} (p_{ij}^{\;c})^{n_{ij}}},$$

2.2.3. Comparison of the two model reduction methods

Among the criteria for a good model reduction method are ease of implementation, a high probability of selecting the true model, and efficiency in reducing the size of the candidate model space. The binomial method is relatively easy to be implemented and less computationally expensive than the multinomial method. But the binomial method only estimates parental marker dosages and there may be multiple marker configurations having the same marker dosages, especially if the marker dosages are low. On the other hand, the multinomial method directly estimates parental marker configuration posterior probabilities; therefore, the multinomial method will be more efficient in reducing the model space.

A simulation study was performed to compare the performance of the two methods for a tetraploid with two markers under all the possible parental marker configurations. The cutoff was set to be *0.90*. Marker genetic distance was from *10* cM (centi-Morgan) to *50* cM with increment *10* cM. Sample sizes ranged from *50* to *500* with increment *50*. For each combination of simulation setting, *10,000* data sets were generated under our preferential pairing mechanism. The probability of including the correct configuration in the candidate configuration space was estimated from the observed proportion, p_{inc}, for each parental configuration. For the binomial method, this proportion denotes the probability of selecting the correct marker dosages; while for the multinomial method, this means the probability of selecting the correct marker configuration. Also the probability of selecting a unique configuration was also estimated from the observed proportion, p_{uni}, as a measurement of the efficiency of reducing the model space. The results are listed in Table 2 with sample sizes *50*, *100*, and *150*, and marker distance *0.10* M, *0.3* M, and *0.5* M chosen for demonstration.

In general, both methods performed better with a larger sample size or shorter marker distance. With sample size *100* and above, both p_{inc} and p_{uni} were almost *1.0* for all of the simulation settings. With a small sample size, the extra information gained from the genetic map added strength to the multinomial method in selecting the correct marker configuration and reducing the model space. Both p_{inc} and p_{uni} of the multinomial method were higher than, or as large as, those of the binomial method except when the sample size was *50*, the marker distance was *0.5 M*, and both marker dosages were *1*.

Based on the simulation results, we suggest that these two methods can be used simultaneously especially if the map is sparse or the sample size is small. If the dosage configuration of the marker configuration selected by the multinomial method is in the candidate dosage configuration set selected by the binomial method, then we can use this selected marker configuration to do interval mapping, Otherwise, we need to use the whole set chosen by the binomial method. With more than two markers (m > *2)*, the binomial method can be naturally extended. The multinomial method can be carried out with all the markers considered jointly, and in that case, the dimensionality of n_{obs} and p^c is 2^m.

3. Simulation

Using a pseudo-doubled backcross experiment, a simulation study was performed for tetraploids with two markers and one QTL for all possible parental configurations. Among the factors affecting the worth of this algorithm, we chose to vary the location of the putative QTL in the marker interval, the parental configuration, and sample size to investigate how the QTL or marker dosages and their linkage affect the performance of our algorithm. The genetic distance between the two markers was set to be *50* cM. Let d_h, $h = 1, 2$ denote the distance between M_h and Q. The true location of Q was determined by the ratio of d_1 and d_2. Three ratios *1*:*9*, *3*:*7*, and *5*:*5* were used (e.g., if the ratio is *1*:*9*, then the QTL is *5* cM to the right of M_1 and *45* cM to the left of M_2). The trait distribution parameter vector was fixed at $\theta = (a, b, \sigma) = (2.0, 10.0, 1.0)$, and for each location, *100* data sets were generated, each having *500* progeny with marker presence/absence and trait data.

The simulation showed that if the correct configuration was selected, the estimates of the QTL position and $\theta = (a, b, \sigma)$ were close to the true values; otherwise, the estimates could be severely biased. Therefore, to be able to select the correct configuration is important. Configurations with higher dosages of marker and QTL, often have higher disequilibrium between the marker and the QTL and provide more information for recombination fraction compared with configurations with lower dosages of marker and QTL. This results in stronger linkage (or, stronger statistical association) between the QTL and markers, which in turn helps to capture the correct configuration and reduce the variation of the selected models. This is an interesting result because it implies that the statistical interpretation for linkage between markers and QTL is affected by the dosages of QTL and marker. An example is shown in Table 3, where the parental configuration $(M_1, Q, M_2 \mid M_1, Q, M_2)$ (Model I) has one more dose of QTL than the parental configuration $(M_1, Q, M_2 \mid M_1, q, M_2)$ (Model II). For all the three locations of the QTL, model I selected the correct configuration *100%*, while model II selected *99%* when the QTL is very close to M_1, and only *90%* when the QTL is at the center of the interval, which also demonstrates that the ratio of selecting the correct model is smaller if the linkage between the QTL and the markers is weaker.
The simulation showed that

4. Discussion

Model selection for QTL analysis using interval mapping for pseudo-double backcross experiments in autopolyploids is presented. Assuming a framework that includes a genetic map, progeny marker presence/absence data, and trait data, our approach first identifies the potential parental configurations (models), which fit the estimated parental marker dosages using a Bayesian approach. Based on the putative models, the QTL location and its effects are then estimated using likelihood ratio tests. Since only the presence or absence state of each marker is known, the number of potential parental configurations (models) increases dramatically as the number of marker and/or ploidy level increases. Estimating marker dosages using Bayesian methods is shown to be a useful way of reducing the model space.

Simulation studies demonstrate that marker dosage, QTL dosage, and QTL position affect the chance of selecting the correct configuration and thus, the accuracy in estimating parameters. A stronger linkage between the QTL and its markers is the key to increasing the power of detecting the QTL, and we have learned that dosage plays a part of this increased power. Our simulation study helps answer the questions raised in Doerge and Craig (2000). 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? Simulation shows the linkage between the marker and QTL is the strongest if they are in coupling and their dosages agree. In which situations is the linkage more strongly affected? We know from our simulation that higher dosages of marker and/or QTL strengthen the linkage. Would models with dosage levels more similar to each other be more likely, especially with close linkage? Yes, as demonstrated by simulation for a tetraploid, it is harder to choose among models with single dose markers than to choose among models with double dose markers simply because there are more possible models with single dose markers than those with double dose markers.

Hackett et al. (2001) also proposed a method for interval mapping in autotetraploids using both dominant markers and codominant markers, like SSR markers. Their method is based on weighted regression analysis and extends the allelic effect structure from the diploid setting to the polyploid setting. Since our method can be extended to include codominant markers, the strength of our maximum likelihood based method lies in the fact that it allows the development of more general QTL allelic effects than linear or additive effects as seen in regression-based methods. Rodzen and May (2002) suggested scoring multiallelic SSR markers as individual dominant markers unless the markers' underlying mode of inheritance are known because different loci may have different inheritance patterns.

The effect of the assumptions we make on the genetic map used for QTL mapping in polyploids is one area of research that remains unaddressed. Currently, our approach and the approach by Hackett et al. (2001) assume that the genetic distances are provided from a genetic map, but the parental configuration from which the progeny are generated remain unknown. Both methods select the potential pool of parental configurations based on observed progeny data. However, when a genetic map is estimated, the most likely parental configuration has to be estimated since the estimated recombination fractions are based on the estimated parental configuration. Therefore, the parental configuration used for estimating the genetic map should be consistent with the one used to locate QTL. This leads to the question of whether we should use the estimated parental configuration when we map QTL. How this question is addressed depends on our confidence in the estimated configuration. If the most likely parental configuration is not the correct configuration, the reliability of the map and the mapped QTL will be in question. Framework is currently under development to allow variability in estimated map, and to gain insight into these questions.

Acknowledgments

We thank Dr. Tom Osborn (Wisconsin) for helpful discussions and his continued work in this area.

References

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

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from imcomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B 39* (1), 1--38.

Dietrich, W. F., J. Miller, R. Steen, et al. (1996). A comprehensive genetic map of the mouse genome. *Nature 380*, 149--152.

Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics 3*, 43--52.

Doerge, R. W. and B. A. Craig (2000). Model selection for quantitative trait locus analysis in polyploids. *Proceedings of the National Acadmy of Sciences 97* (14), 7951--7956.

Grattapaglia, D. and R. Sederoff (1994). Genetic linkage maps of eucalyptus grandis and eucalyptus urophylla using a pseudo-testcross: Mapping strategy and rapd markers. *Genetics 137*, 1121--1137.

Guha, S. and S. C. Maheshwari (1964). In vitro production of embryos from amthers of datura. *Nature 204*, 497.

Jackson, R. C. and J. W. Jackson (1996). Gene segregation in autotetraploids: prediction from meiotic configurations. *American Journal of Botany 83* (6), 673--678.

Jansen, R. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics 135*, 205--211.

Koornneef, M., J. van Eden, C. J. Hanhart et al. (1983). Linkage map of *Arabidopsis haliana*. *Journal of Heredity 74*, 265--272.

Lander, E. S. and D. Botstein (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics 121*, 185--199.

Osborn, T., J. Pires, J. Birchler et al. (2003). Understanding mechanisms of novel gene expression in polyploids. *Trends in Genetics 19* (3), 141--147.

Rieseberg, L. H. and M. F. Doyle (1989). Tetrasomic segregation in the naturally occurring autotetraploid *Allium nevii* (aliaceae). *Hereditas 111*, 31--36.

Ripol, M., G. Churchill, J. da Silva, and M. Sorrells (1999). Statistical aspects of genetic mapping in autopolyploids. *Gene 235*, 31--41.

Rodzen, J. A. and B. May (2002). Inheritance of microsatellite loci in the white sturgeon Acipenser transmontanus. *Genome 45*, 1064--1076.

Soltis, P. S. and D. E. Soltis (2000). The role of genetic and genomic attributes in the success of polyploids. *Proceedings of the National Academy of Sciences if the United States of America 97* (13), 7051--7057.

Sybenga, J. (1995). Meiotic pairing in autohexaploid {\it lathyrus}: a mathematical model. *Heredity 75*, 343--350.

Wu, K. K., W. Burnquist, M. Sorrells et al. (1992). The detection and estimation of linkage in polyploids using single-dose restriction fragments. *Theoretical and Applied Genetics 83*, 294--300.

Zeng, Z.-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of National Academy of Sciences of the United States of America 90* (23), 10972--10976.

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

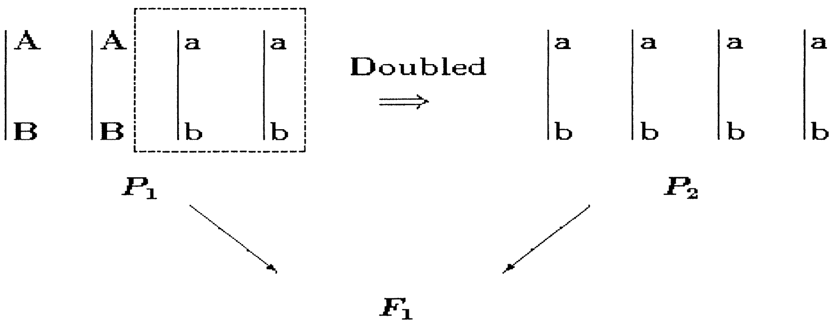


Figure 1. A pseudo-doubled backcross experiment for a tetraploid with two loci *A/a* and *B/b*. The upper case denotes the dominant allele. P_1 is the informative parent with two dose of each dominant allele. P_2 is the non-informative doubled-haploid produced by doubling a haploid of P_1, which only contains recessive alleles.

	$(M_1M_2\|M_1M_2)$	$(M_1M_2\|M_1m_2)$	$(M_1M_2\|m_1M_2)$	$(M_1M_2\|m_1m_2)$	$(m_1M_2\|M_1m_2)$
p_{00}	$0.25(1-r)^2$	$0.25(1-r)$	$0.25(1-r)$	$0.5(1-r)$	0.25
p_{01}	$0.25(1-(1-r)^2)$	$0.25(1+r)$	$0.25r$	$0.5r$	0.25
p_{10}	$0.25(1-(1-r)^2)$	$0.25r$	$0.25(1+r)$	$0.5r$	0.25
p_{11}	$0.5+0.25(1-r)^2$	$0.5-0.25r$	$0.5-0.25r$	$0.25(1-r)$	0.25

Table 1. Multinomial distribution probability parameters $p^c = (p_{00}^c, p_{01}^c, p_{10}^c, p_{11}^c)$ for parental marker configurations in a tetraploid with two markers. The recombination fraction between the two markers is denoted as *r*.

	n	50			100			150		
Binomial	d	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
$(M_1M_2\|M_1M_2)$	p_{inc}	0.989	0.989	0.988	0.997	0.997	0.997	0.997	0.999	0.999
	p_{uni}	0.866	0.841	0.835	0.974	0.970	0.972	0.997	0.997	0.997
$(M_1M_2\|m_1M_2)$	p_{inc}	0.985	0.987	0.984	0.998	0.997	0.997	1.000	0.999	0.999
	p_{uni}	0.816	0.822	0.825	0.976	0.977	0.975	0.997	0.996	0.996
$(M_1M_2\|M_1m_2)$	p_{inc}	0.987	0.986	0.986	0.997	0.998	0.998	0.999	0.999	0.999
	p_{uni}	0.812	0.823	0.813	0.976	0.976	0.974	0.996	0.997	0.996
$(M_1M_2\|m_1m_2)$	p_{inc}	0.988	0.987	0.985	0.998	0.999	0.999	0.999	0.999	0.999
	p_{uni}	0.860	0.842	0.841	0.983	0.982	0.979	0.996	0.997	0.997
$(m_1M_2\|M_1m_2)$	p_{inc}	0.983	0.987	0.987	0.998	0.998	0.997	0.999	0.999	0.999
	p_{uni}	0.821	0.821	0.819	0.982	0.982	0.981	0.996	0.996	0.996
Multinomial		0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
$(M_1M_2\|M_1M_2)$	p_{inc}	0.993	0.990	0.988	0.999	0.999	0.998	1.000	1.000	0.999
	p_{uni}	0.902	0.854	0.830	0.991	0.987	0.982	0.999	0.998	0.998
$(M_1M_2\|m_1M_2)$	p_{inc}	0.996	0.991	0.984	0.999	0.999	0.998	1.000	0.999	0.999
	p_{uni}	0.975	0.885	0.829	0.999	0.991	0.979	0.999	0.999	0.998
$(M_1M_2\|M_1m_2)$	p_{inc}	0.997	0.989	0.986	0.999	0.999	0.998	1.000	1.000	1.000
	p_{uni}	0.977	0.890	0.821	0.999	0.992	0.981	1.000	0.999	0.999
$(M_1M_2\|m_1m_2)$	p_{inc}	0.993	0.987	0.976	0.999	0.999	0.992	1.000	1.000	0.996
	p_{uni}	0.916	0.829	0.620	0.992	0.987	0.888	0.998	0.999	0.998
$(m_1M_2\|M_1m_2)$	p_{inc}	0.998	0.990	0.976	1.000	0.999	0.992	1.000	1.000	0.999
	p_{uni}	0.989	0.903	0.623	0.999	0.992	0.901	1.000	0.999	0.998

Table 2. The estimated probability of selecting the correct configuration in the candidate configuration (model) space, p_{inc}, and the estimated probability of selecting a unique configuration, p_{uni}, for a tetraploid with two markers, *10,000* simulated data sets, and a cutoff of *0.90* for both the binomial and multinomial model reduction methods. The sample size is denoted as *n*, and *d* stands for the marker genetic distance with unit Morgan (M).

Parental Config	Ratio	Selected Config	Freq	$E(d(M_1, Q))$	$\mathrm{std}(d(M_1, Q))$	$E(\hat{a})$	$E(\hat{b})$	$E(\hat{\sigma})$
$M_1\,Q\,M_2$ $M_1\,q\,M_2$	1:9	$M_1\,Q\,M_2$ $M_1\,q\,M_2$	99	0.0464	0.0344	1.989	10.004	1.003
	1:9	$M_1\,Q\,M_2$ $M_1\,Q\,M_2$	1	0.0001	—	0.997	10.081	1.270
	3:7	$M_1\,Q\,M_2$ $M_1\,q\,M_2$	95	0.1442	0.0438	1.999	10.000	1.000
	3:7	$M_1\,Q\,M_2$ $M_1\,Q\,M_2$	5	0.1661	0.0321	1.162	9.8820	1.132
	5:5	$M_1\,Q\,M_2$ $M_1\,q\,M_2$	90	0.2475	0.0409	2.006	9.9970	0.999
	5:5	$M_1\,Q\,M_2$ $M_1\,Q\,M_2$	10	0.2551	0.0268	1.198	9.8300	1.109
$M_1\,Q\,M_2$ $M_1\,Q\,M_2$	1:9	$M_1\,Q\,M_2$ $M_1\,Q\,M_2$	100	0.0482	0.0199	1.997	10.004	0.996
	3:7	$M_1\,Q\,M_2$ $M_1\,Q\,M_2$	100	0.1460	0.0236	2.000	10.004	0.992
	5:5	$M_1\,Q\,M_2$ $M_1\,Q\,M_2$	100	0.2482	0.0241	1.996	10.005	0.994

Table 3. Estimated location of the QTL and trait distribution parameter *$\theta = (a, b, \sigma)$* from *100* simulated data sets with sample size *500* under the parental configuration (M_1, Q, M_2 | M_1, Q, M_2). *a* is the additive QTL effect. *b* is the grand mean of trait distribution. *σ* is the stand deviation of the trait distribution. "Ratio" stands for d_1:d_2. "Freq" denotes the frequency of having the largest likelihood ratio test statistic for the corresponding model.