Heredity (2003) 90, 459–467

www.nature.com/hdy

Mapping viability loci using molecular markers

L Luo and S Xu
Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

In genetic mapping experiments, some molecular markers often show distorted segregation ratios. We hypothesize that these markers are linked to some viability loci that cause the observed segregation ratios to deviate from Mendelian expectations. Although statistical methods for mapping viability loci have been developed for line-crossing experiments, methods for viability mapping in outbred populations have not been developed yet. In this study, we develop a method for mapping viability loci in outbred populations using a full-sib family as an example. We develop a maximum likelihood (ML) method that uses the observed marker genotypes as data and the proportions of the genotypes of the viability locus as parameters. The ML solutions are obtained via the expectation–maximization algorithm. Application and efficiencies of the method are demonstrated and tested using a set of simulated data. We conclude that mapping viability loci can be accomplished using similar statistical techniques used in quantitative trait locus mapping for quantitative traits.
Heredity (2003) **90,** 459–467. doi:10.1038/sj.hdy.6800264

Keywords: EM algorithm; four-way cross; maximum likelihood; segregation distortion

Introduction

The genetic consequence of selection is the change in frequencies of the genes affecting fitness. The process of evolution is reflected by the dynamic change of gene frequencies by selection and other evolutionary agents. Fitness is a complicated trait, which can be decomposed into many fitness components (Falconer and Mackay, 1996; Hartl and Clark, 1997). Therefore, the genetic variance of fitness is considered to be controlled by the segregation of multiple genes. Fitness behaves like a quantitative trait. It responds to natural selection with a response equal to the genetic variance of fitness (Fisher, 1958). To study the genetic architecture of fitness, it is important to explore the change of gene frequency of alleles at individual loci. However, only in very limited situations, for example, where allozyme markers are available, can we evaluate natural selection on individual loci. In most situations, we do not know what the genes are and where in the genome the genes are located. With the rapid development of molecular technology, large amounts of molecular data are now available, which provide a great opportunity to estimate the effects and locate the chromosomal positions of loci responsible for complicated traits, for example, quantitative traits. The technology is now called quantitative trait locus (QTL) mapping. Since fitness is just another complicated trait with a polygenic background, a similar technology can be applied to map loci determining variation in fitness.

Although it does not seem easy to map fitness loci, statistical methods of mapping QTL can be adopted (Lander and Botstein, 1989). Fu and Ritland (1994a,b) first utilized a QTL mapping approach to map viability (a fitness component) loci under the maximum likelihood (ML) framework. Mitchell-Olds (1995) also proposed a similar ML method for viability mapping in F_2 families. Recently, Vogl and Xu (2000) investigated a Bayesian method to map viability loci in a backcross family. All the aforementioned existing methods deal with line-crossing experiments that require inbred lines. Inbred lines, however, may not be available for many species, such as humans, large animals and trees (Hedrick and Muona, 1990). Mapping viability loci may be more relevant to natural populations than to line crosses. This is equivalent to the situation where mapping QTLs is more relevant to breeding populations than to designed line crosses. However, it is easier to map QTLs in line-crossing experiments because we can control the genetic background and environments. After QTL are mapped in line crosses, the results may be extended to natural populations or used to find homologous loci in closely related species. Similarly, viability loci may be mapped in line crosses and the inference later extended to natural populations. In this study, we attempt to map viability loci directly in outbred populations. Full-sib families are the simplest outbred populations. Although not necessarily natural populations, they are one step closer to natural populations than are line crosses.

The fitness of a genotype at a locus is the average fitness of all individuals bearing this genotype. If we assign the fitness for the 'best' genotype a value of one, the selection coefficient for an arbitrary genotype is defined as the reduction in fitness from this maximum value. Therefore, we only describe the measurement of fitness (rather than the selection coefficient) in subsequent discussion. Viability is only one of many components of fitness. Fecundity is another important component. In this study, however, we focus only on

Correspondence: S Xu, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA.
E-mail: xu@genetics.ucr.edu
Received 7 June 2002; accepted 17 January 2003

loci responsible for viability selection, assuming that all surviving individuals have an equal fecundity.

We develop a model of viability mapping that uses a full-sib family derived from the mating of two unrelated outbred parents. A full-sib family contains four different alleles at a single locus, rather than two as is usually assumed in inbred line crosses. Mapping in a full-sib family requires the general rule of allelic transmission from parents to children and thus the algorithm can be extended to pedigree analysis. The method can be directly applied to fitness analysis for open-pollinated plants.

Theory and methods

Genetic model of fitness

Consider a single viability locus and a full-sib family. Denote the genotypes of the sire (paternal parent) and dam (maternal parent) by $A_1^sA_2^s$ and $A_1^dA_2^d$, respectively. Mating between the two parents will generate progenies each with one of the four possible genotypes: $\{A_1^sA_1^d, A_1^sA_2^d, A_2^sA_1^d, A_2^sA_2^d\}$. Under the assumption of Mendelian segregation, the four genotypes will have an equal frequency, that is, $\frac{1}{4}$. If this locus is subject to viability selection, we will observe two or more genotypes, which have frequencies different from Mendelian expectations.

To model viability selection, we define the underlying frequencies of the four genotypes in the progeny by a vector $\mathbf{w} = [\,w_{11} \quad w_{12} \quad w_{21} \quad w_{22}\,]$ for $0 \le w_{kl} \le 1$, $\sum_{kl} w_{kl} = 1$ and $k, l = 1, 2$. These frequencies are now defined as the relative fitness of the four genotypes. This is a little different from the usual definition of relative fitness in which the maximum fitness is set to one and the rest expressed as reduced values relative to one. Deviation of $\mathbf{w}$ from the Mendelian vector $\mathbf{w}_0 = \left[\frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{4}\right]$ reflects the intensity of viability selection.

The fitness of a genotype can be decomposed into the product of the fitness of the two alleles that make up the genotype and a deviation reflecting the interaction between the two alleles, called the dominance effect, that is,

$$w_{kl} = w_k^s w_l^d + \delta_{kl} \tag{1}$$

where w_k^s and w_l^d denote the relative fitness of the kth allele of the sire and the lth allele of the dam, respectively, and δ_{kl} is the dominance effect. This partitioning of the fitness is important because we can separate gametic selection from zygotic selection using statistical technology. Note that there are four possible genotypes in the progeny, but after the decomposition we have eight parameters. Therefore, we must impose some restriction to the parameters to make the model estimable. We take the restrictions similar to those used in the four-way cross model (Xu, 1998) and define three new independent parameters:

$$\begin{aligned} w^s &= w_{11} + w_{12} - w_{21} - w_{22} \\ w^d &= w_{11} - w_{12} + w_{21} - w_{22} \\ \delta &= w_{11}w_{22} - w_{12}w_{21} \end{aligned} \tag{2}$$

It is interesting to know that the fitness values of the four genotypes can be expressed as functions of the three independent parameters, as shown below:

$$\begin{aligned} w_{11} &= \tfrac{1}{4}(1 + w^s)(1 + w^d) + \delta \\ w_{12} &= \tfrac{1}{4}(1 + w^s)(1 - w^d) - \delta \\ w_{21} &= \tfrac{1}{4}(1 - w^s)(1 + w^d) - \delta \\ w_{22} &= \tfrac{1}{4}(1 - w^s)(1 - w^d) + \delta \end{aligned} \tag{3}$$

This model is important in hypothesis tests and computer simulations that will be discussed in later sections.

ML estimation

We first assume that the four alleles of the viability locus in the parents are distinguishable and the genotypes are observable. Suppose that we sample n individuals from the full-sib family in question. Let us define

$$\mathbf{y}_j = [\,y_{j(11)} \quad y_{j(12)} \quad y_{j(21)} \quad y_{j(22)}\,] \text{ for } j = 1, \ldots, n$$

where $y_{j(kl)} = 1$ and $y_{j(k'l')} = 0$ for $k' \ne k$ and $l' \ne l$ if individual j takes genotype $A_k^sA_l^d$. We now have the data, $\mathbf{y}$, and the parameter, $\mathbf{w}$, which allow the construction of the log likelihood:

$$L_c(\mathbf{w}) = \sum_{j=1}^{n} \left[\sum_{k=1}^{2} \sum_{l=1}^{2} y_{j(kl)} \ln(w_{kl}) \right] \tag{4}$$

The ML estimate of $\mathbf{w}$ is simply

$$\hat{w}_{kl} = \frac{1}{n} \sum_{j=1}^{n} y_{j(kl)} \tag{5}$$

for $k, l = 1, 2$.

In fact, the genotype of a viability locus cannot be observed and we must use markers to infer the genotype. Unless the viability locus is located exactly at a fully informative marker, inference will be subject to error. The amount of error depends on the distances of the viability locus from marker loci, the level of marker polymorphism and the genotypes of the markers. As a result of the error, we are not certain about the actual genotype of the viability locus for each individual, even though we can observe the marker genotypes. The viability locus can take any one of the four genotypes, but with a different probability for each genotype given the marker information. Define the four conditional probabilities of the given viability locus markers by $\mathbf{p}_j = [\,p_{j(11)} \quad p_{j(12)} \quad p_{j(21)} \quad p_{j(22)}\,]$ for $0 \le p_{j(kl)} \le 1$ and $\sum_{k=1}^{2}\sum_{l=1}^{2} p_{j(kl)} = 1$. This is a typical problem of missing values in statistics where we can use the expectation-maximization (EM) algorithm to solve for the MLE. The actual incomplete-data log likelihood is

$$L(\mathbf{w}) = \sum_{j=1}^{n} \ln\left(\sum_{k=1}^{2} \sum_{l=1}^{2} p_{j(kl)} w_{kl} \right) \tag{6}$$

where the missing data $\mathbf{y}$ have been 'integrated out'. There are several ways to solve the MLE, but we take the EM algorithm (Dempster *et al*, 1977).

First, we choose an initial value $\mathbf{w}^{(0)}$ and calculate the expectation of $y_{j(kl)}$ conditional on $\mathbf{w} = \mathbf{w}^{(0)}$,

$$E[\,y_{j(kl)}] = \hat{y}_{j(kl)} = \frac{p_{j(kl)} w_{kl}^{(0)}}{\sum_{k'=1}^{2}\sum_{l'=1}^{2} p_{j(k'l')} w_{k'l'}^{(0)}} \tag{7}$$

which is also called the posterior probability of $y_{j(kl)}$. We have now completed the expectation step (E-step). The maximization step (M-step) is simply to replace $y_{j(kl)}$ in equation (7) by the conditional expectation,

$$w_{kl}^{(1)} = \frac{1}{n}\sum_{j=1}^{n}\hat{y}_{j(kl)} \tag{8}$$

This concludes the first iteration of the EM algorithm. The iteration continues until convergence at the tth iteration and the MLE takes $\hat{\mathbf{w}} = \mathbf{w}^{(t)}$. According to the invariance property of MLE, we have

$$\begin{aligned}\hat{w}^{\mathrm{s}} &= \hat{w}_{11} + \hat{w}_{12} - \hat{w}_{21} - \hat{w}_{22}\\ \hat{w}^{\mathrm{d}} &= \hat{w}_{11} - \hat{w}_{12} + \hat{w}_{21} - \hat{w}_{22}\\ \hat{\delta} &= \hat{w}_{11}\hat{w}_{22} - \hat{w}_{12}\hat{w}_{21}\end{aligned} \tag{9}$$

The EM algorithm provides a convenient way to solve the MLE, but it does not automatically give the asymptotic variance–covariance matrix of $\hat{\mathbf{w}}$, which must be obtained separately through some additional computation (Louis, 1982). This is the drawback of the EM algorithm compared to Fisher's scoring method, which automatically provides an asymptotic variance–covariance matrix for the MLE. However, Fisher's scoring method requires calculation of the information matrix, which is not easy in the missing value problem. In practice, we can use the bootstrap method (Efron, 1979) to assess the variance–covariance matrix. The bootstrap method is computationally demanding, but the method is executed only once after convergence has been reached and only on the positions that show significant evidence of viability selection.

Hypothesis test

Recall that the conditional probability of the viability locus genotype is calculated from marker information with the assumption that the location of the viability locus relative to the markers is known. Therefore, the hypothesis test on the effects of the viability locus is actually a conditional test given the position of the viability locus. If the test is not significant, we will conclude that the current position of the chromosome being tested does not segregate for a viability locus. To test the overall hypothesis of no viability selection, we need to scan the entire genome (multiple tests). The null hypothesis (no viability selection) will be rejected if none of the locus-specific tests is significant. We will discuss the overall test later and now focus on the test of an individual locus.

The first null hypothesis is H_0: $\mathbf{w} = \mathbf{w}_0$, which tests no segregation distortion for the locus of interest. The test statistic is $\lambda = -2[L(\mathbf{w}_0) - L(\hat{\mathbf{w}})]$, where $L(\mathbf{w}_0) = n\ln(\frac{1}{4}) = -1.3863n$. Under the null hypothesis, λ will approximately follow a χ^2 distribution with three degrees of freedom.

If this null hypothesis is rejected, we can further test the significance of each component. The null hypothesis that the two alleles carried by the sire have identical fitness is formulated by H_{s}: $w^{\mathrm{s}} = 0, w^{\mathrm{d}} \neq 0, \delta \neq 0$. The test statistic for H_{s} is $\lambda_{\mathrm{s}} = -2[L(\hat{\mathbf{w}}_{\mathrm{s}}) - L(\hat{\mathbf{w}})]$ where $L(\hat{\mathbf{w}}_{\mathrm{s}})$ is the log likelihood value obtained by maximizing $L(\mathbf{w})$ under the restriction of $w^{\mathrm{s}} = (w_{11} + w_{12}) - (w_{21} + w_{22}) = 0$, which is achieved by using the Lagrange multiplier. A more intuitive and easier way to enforce the restriction is to make the substitutions, $w_{12} = \frac{1}{2} - w_{11}$ and $w_{22} = \frac{1}{2} - w_{21}$, which reduces the number of parameters to two, w_{11} and w_{21}. The EM solutions of these two parameters are

$$\hat{w}_{11} = \frac{1}{2}\frac{\sum_{j=1}^{n}\hat{y}_{j(11)}}{\sum_{j=1}^{n}\hat{y}_{j(11)} + \sum_{j=1}^{n}\hat{y}_{j(12)}} = \frac{1}{n}\sum_{j=1}^{n}\hat{y}_{j(11)}$$

and

$$\hat{w}_{21} = \frac{1}{2}\frac{\sum_{j=1}^{n}\hat{y}_{j(21)}}{\sum_{j=1}^{n}\hat{y}_{j(21)} + \sum_{j=1}^{n}\hat{y}_{j(22)}} = \frac{1}{n}\sum_{j=1}^{n}\hat{y}_{j(21)}$$

because the denominators equal $\frac{n}{2}$ due to the restrictions. The MLE of the remaining parameters are $\hat{w}_{12} = \frac{1}{2} - \hat{w}_{11}$ and $\hat{w}_{22} = \frac{1}{2} - \hat{w}_{21}$. Under H_{s}, λ_{s} will approximately follow a χ^2 distribution with one degree of freedom.

The null hypothesis that the two alleles carried by the dam have identical fitness is formulated by H_{d}: $w^{\mathrm{d}} = 0, w^{\mathrm{s}} \neq 0, \delta \neq 0$, where the test statistic for H_{d} is $\lambda_{\mathrm{d}} = -2[L(\hat{\mathbf{w}}_{\mathrm{d}}) - L(\hat{\mathbf{w}})]$, with $L(\hat{\mathbf{w}}_{\mathrm{d}})$ being the log likelihood value obtained by maximizing $L(\mathbf{w})$ under the restriction of $w^{\mathrm{d}} = (w_{11} + w_{21}) - (w_{12} + w_{22}) = 0$. The EM solutions of the parameters are

$$\hat{w}_{11} = \frac{1}{2}\frac{\sum_{j=1}^{n}\hat{y}_{j(11)}}{\sum_{j=1}^{n}\hat{y}_{j(11)} + \sum_{j=1}^{n}\hat{y}_{j(21)}} = \frac{1}{n}\sum_{j=1}^{n}\hat{y}_{j(11)}$$

and

$$\hat{w}_{12} = \frac{1}{2}\frac{\sum_{j=1}^{n}\hat{y}_{j(12)}}{\sum_{j=1}^{n}\hat{y}_{j(12)} + \sum_{j=1}^{n}\hat{y}_{j(22)}} = \frac{1}{n}\sum_{j=1}^{n}\hat{y}_{j(12)}$$

The MLE of the remaining parameters are $\hat{w}_{21} = \frac{1}{2} - \hat{w}_{11}$ and $\hat{w}_{22} = \frac{1}{2} - \hat{w}_{12}$.

Again, under H_{d}, λ_{d} will approximately follow a χ^2 distribution with one degree of freedom.

The null hypothesis that the dominance effect is absent is formulated as H_{d}: $\delta = 0, w^{\mathrm{s}} \neq 0, w^{\mathrm{d}} \neq 0$. Let us define

$$a = \frac{1}{n}\left(\sum_{j=1}^{n}\hat{y}_{j(11)} + \sum_{j=1}^{n}\hat{y}_{j(12)}\right) \text{ and}$$

$$b = \frac{1}{n}\left(\sum_{j=1}^{n}\hat{y}_{j(11)} + \sum_{j=1}^{n}\hat{y}_{j(21)}\right)$$

The MLE under this restriction are $\hat{w}_{11} = ab$, $\hat{w}_{12} = a(1-b)$, $\hat{w}_{21} = (1-a)b$ and $\hat{w}_{22} = (1-a)(1-b)$. Again, the test statistic for H_δ is $\lambda_\delta = -2[L(\hat{\mathbf{w}}_\delta) - L(\hat{\mathbf{w}})]$, which follows approximately a χ^2 distribution with one degree of freedom.

Genome scanning

To scan viability loci for the entire genome, we need to move the putative position from one end to the other end of the genome. The genotype of each chromosome position for each individual is inferred from marker information, that is, $p_{j(kl)} = \Pr(y_{j(kl)} = 1|I_{\mathrm{M}})$, where I_{M} stands for marker information. For outbred populations, not all markers are fully informative. Therefore, we adopted the multipoint method developed by Rao and Xu (1998) to infer the probabilities of viability loci. This multipoint method is identical to that of Kruglyak and Lander (1995) when the linkage phases of the parents are

Table 1 Parameter values used in the simulation experiments

Parameters	*Genetic model*						
	Additive (A)			*Dominance (D)*			*Both A and D*
	High	*Medium*	*Low*	*High(−)*	*Low*	*High(+)*	
w^s	0.300	0.200	0.100	0.000	0.000	0.000	0.150
w^d	0.300	0.200	0.100	0.000	0.000	0.000	0.150
δ	0.000	0.000	0.000	−0.150	0.050	0.150	0.100
w_{11}	0.4225	0.3600	0.3025	0.100	0.300	0.400	0.4306
w_{12}	0.2275	0.2400	0.2475	0.400	0.200	0.100	0.1444
w_{21}	0.2275	0.2400	0.2475	0.400	0.200	0.100	0.1444
w_{22}	0.1225	0.1600	0.2025	0.100	0.300	0.400	0.2806
s_{11}	0.0000	0.0000	0.0000	0.750	0.0000	0.000	0.0000
s_{12}	0.4615	0.3333	0.1818	0.000	0.3333	0.750	0.6647
s_{21}	0.4615	0.3333	0.1818	0.000	0.3333	0.750	0.6647
s_{22}	0.7100	0.5555	0.3305	0.750	0.0000	0.000	0.3483

known. In our study, we focus on developing the genetic model of viability mapping rather than the multipoint method. Therefore, we assume that the parental marker linkage phases are known without error. This assumption holds very well when the family size is sufficiently large because the true linkage phases can be easily recovered using marker information of the progeny.

To find the optimal location of the viability locus on the chromosome, we test all putative positions. However, the chromosome is a continuous linear structure, and there are an infinite number of putative positions. As usually done in interval mapping (Lander and Botstein, 1989), we scan the whole chromosome from one end to the other by evaluating a position in every one or two cM. The likelihood ratio test statistic is then plotted against the chromosomal position to form a test statistic profile. The MLE of the position of viability locus takes the one where the peak occurs. The critical value used for declaring at least one viability locus on the entire genome with a type I error rate of $\alpha = 0.05$ is found using the permutation test (Churchill and Doerge, 1994).

Monte Carlo simulation

We simulated one chromosome of length 100 cM with 11 markers evenly spaced. The two alleles of each parent at each locus were randomly assigned from five distinguishable alleles (randomly selecting two out of five). This generates markers with a range of information content. A single viability locus was simulated at position 25 cM, that is, between markers 3 and 4. The following factors were considered in the simulations: the mode of viability selection, the intensity of viability selection and sample size of the mapping population. The purpose of the simulation was not to compare the relative efficiencies of different methods for viability mapping (since there are no other methods to compare), nor to investigate the range of parameter values where the method works best. Instead, we simply attempted to demonstrate that the method works well and the test statistic behaves as expected. From this simulation study, we try to validate our method and program of viability mapping.

The mode of viability selection was investigated under three levels: an additive model, a dominance model and a combination of both additive and dominance. For the additive model, we set $\delta = 0$ and $w^s = w^d = 0.1, 0.2, 0.3$. From these parameters, the fitness values of the four genotypes were generated. Under the dominance model, we set $w^s = w^d = 0$ and $\delta = -0.15, 0.05, 0.15$. We also investigated one model with both the additive and dominance effects, that is $w^s = w^d = 0.15$ and $\delta = 0.1$.

From the three effects of the viability locus, we use equation (3) to calculate the actual fitness values of the four possible genotypes. For example, when $w^s = w^d = 0.15$ and $\delta = 0.1$, the four fitness values are

$$w_{11} = \tfrac{1}{4}(1 + 0.15)(1 + 0.15) + 0.1 = 0.4306$$
$$w_{12} = \tfrac{1}{4}(1 + 0.15)(1 - 0.15) - 0.1 = 0.1444$$
$$w_{21} = \tfrac{1}{4}(1 - 0.15)(1 + 0.15) - 0.1 = 0.1444$$
$$w_{22} = \tfrac{1}{4}(1 - 0.15)(1 - 0.15) + 0.1 = 0.2806$$

Following the conventional notation of natural selection, we calculate the selection coefficient for the fitness of genotype $A_k^s A_l^d$ using $s_{kl} = 1 - w_{kl}/w_{\max}$ (Hartl and Clark, 1997). These selection coefficients were used to determine whether an individual with genotype $A_k^s A_l^d$ should be deleted from the mapping population (Table 1).

Three different sample sizes under each of the above models were investigated, $n = 50, 100, 200$. The estimated location of the putative viability locus under each analysis took the position where the peak of the test statistic profile occurred. The simulation was replicated 100 times under each setting. The means and standard deviations of the 100 replicates were used to evaluate the performance of each parameter combination.

The empirical statistical power for each setting was calculated as the percentage of the replicates (out of 100 simulations) with the highest (overall) test statistic (along the chromosome) greater than the empirical critical value. The expected standard error for the empirical power is $\sqrt{\beta(1-\beta)/N_r}$ where N_r is the number of replicates. For example, if the true power is $1 - \beta = 0.8$, the standard error is 0.04, which is reasonably small. The critical value was obtained by simulating additional 1000

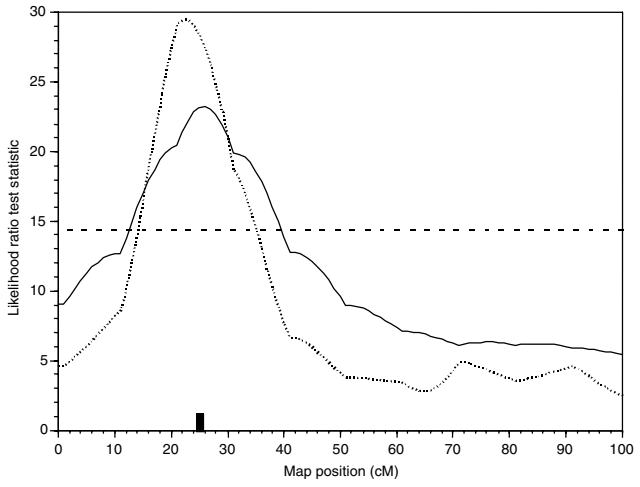

Figure 1 Likelihood ratio test statistic profiles for the combined A/D model $w^s = w^d = 0.15$ and $\delta = 0.1$ with sample size $n = 100$. The simulated position of the viability locus is located at position 25 cM (indicated by the solid bar). The solid line is the average profile of 100 replicates, the dotted line is the profile of a randomly picked single run from the 100 replicates and the dashed horizontal line is the threshold value for the test statistic at $\alpha = 0.05$.

samples under the null hypothesis. The highest test statistics of the 1000 samples were ranked from the lowest to the highest. The empirical critical value took the 95th percentile of the distribution of the null samples.

The test statistic profile of a single replicate for the combined additive and dominance model ($w^s = w^d = 0.15$ and $\delta = 0.1$) with sample size $n = 100$ is demonstrated in Figure 1 (the dotted line). From the total test statistic profile, we can see that the viability locus has been identified at position 23 cM, very close to the true position (25 cM). The estimated effects for this particular run are $\hat{w}_{11} = 0.5902$, $\hat{w}_{12} = 0.03290$, $\hat{w}_{21} = 0.03410$ and $\hat{w}_{22} = 0.3528$. These estimated fitness values were converted into $\hat{w}^s = 0.2361$, $\hat{w}^d = 0.2385$ and $\hat{\delta} = 0.2071$ using equation (2). The estimated effects are larger than the simulated effects, but maintain the same trend. The deviations are not larger than expected considering the sampling errors with $n = 100$. The average test statistic profile of the 100 replicates for this setting is shown in Figure 1 (solid line), clearly showing the expected property of the test statistic profile for QTL mapping.

The empirical critical values appear to be quite independent of the sample size and they are about 14.5 at $\alpha = 0.05$ and about 18.0 at $\alpha = 0.01$. These empirical critical values are clearly larger than $\chi^2_{3,0.95} = 7.815$ and $\chi^2_{3,0.99} = 11.34$. Therefore, we used the empirical critical values to declare significance.

Means and standard deviations of the estimated parameters for various genetic models are given in Table 2 for $n = 50$, Table 3 for $n = 100$ and Table 4 for $n = 200$. The results do follow the expected trends: the viability locus location is more accurately estimated as the sample size and the selection intensity increase. When the sample size is small, the estimated position of the locus is severely biased towards the center of the chromosome. Besides these general trends, we found that the additive models are more sensitive to the intensity of selection. Under different levels of parameters (high, medium and low), the accuracy of both the estimated viability locus location and parameters varies more than the dominance models and both additive and dominance (A/D) model. Overall, the A/D model gives the highest accuracy of estimation.

The empirical statistical powers under various genetic models and sample sizes are given in Table 5. The powers are quite low for small sample size ($n = 50$) and are reasonably high when sample size reaches 200. These observations are the same as those expected in the more usual QTL mapping studies. The results of these

Table 2 Means and standard deviations (in parentheses) of estimated parameter values for the EM algorithm with sample size 50

Parameters	*Genetic model*						
	Additive (A)			*Dominance (D)*			*Both A and D*
	High	*Medium*	*Low*	*High(−)*	*Low*	*High(+)*	
cM_A	31.01 (18.35)	39.38 (26.66)	46.43 (31.52)	27.22 (5.96)	50.8 (34.59)	26.94 (8.63)	30.16 (20.01)
w^s	0.3219 (0.1768)	0.2171 (0.2039)	0.1254 (0.1882)	0.0047 (0.1594)	0.0399 (0.1942)	-0.0327 (0.1702)	0.1841 (0.1567)
w^d	0.3064 (0.1685)	0.2814 (0.1877)	0.0844 (0.2195)	−0.0362 (0.1465)	−0.0034 (0.1829)	0.0076 (0.1874)	0.1902 (0.1526)
δ	0.0042 (0.0474)	−0.0098 (0.0516)	0.0129 (0.0532)	−0.1515 (0.0297)	0.0391 (0.0639)	0.1494 (0.0354)	0.1052 (0.0411)
w_{11}	0.4392 (0.0919)	0.3828 (0.0875)	0.3231 (0.0965)	0.0967 (0.0550)	0.3036 (0.0816)	0.3985 (0.0827)	0.4623 (0.0771)
w_{12}	0.2318 (0.0832)	0.2358 (0.0972)	0.2496 (0.0899)	0.4158 (0.0735)	0.2264 (0.0896)	0.0953 (0.0666)	0.1398 (0.0579)
w_{21}	0.2240 (0.0842)	0.2680 (0.0885)	0.2291 (0.0877)	0.3952 (0.0653)	0.2047 (0.0861)	0.1154 (0.0520)	0.1429 (0.0623)
w_{22}	0.1250 (0.0553)	0.1335 (0.0721)	0.2182 (0.0843)	0.1125 (0.0526)	0.2853 (0.1052)	0.4110 (0.0838)	0.2751 (0.0732)

cM_A: the estimated location of the viability locus.

Table 3 Means and standard deviations (in parentheses) of estimated parameter values for the EM algorithm with sample size 100

Parameters	*Genetic model*						
	Additive (A)			*Dominance (D)*			*Both A and D*
	High	*Medium*	*Low*	*High(−)*	*Low*	*High(+)*	
cM_A	26.52 (11.58)	30.43 (20.09)	45.13 (33.50)	26.05 (3.61)	37.63 (26.76)	25.91 (3.30)	28.79 (12.13)
w^s	0.3135 (0.1098)	0.2241 (0.1094)	0.0945 (0.1325)	0.0001 (0.1081)	−0.0035 (0.1428)	−0.0223 (0.1025)	0.1486 (0.1189)
w^d	0.3036 (0.1235)	0.2105 (0.1322)	0.1129 (0.1345)	0.0103 (0.1003)	−0.0002 (0.1370)	−0.0083 (0.1098)	0.1542 (0.1221)
δ	0.0013 (0.0343)	−0.0056 (0.0317)	−0.0037 (0.0443)	−0.1521 (0.0216)	0.0514 (0.0386)	0.1470 (0.0236)	0.1013 (0.0302)
w_{11}	0.4320 (0.0651)	0.3679 (0.0679)	0.3030 (0.0677)	0.1030 (0.0365)	0.3027 (0.0607)	0.3920 (0.0521)	0.4356 (0.0654)
w_{12}	0.2297 (0.0555)	0.2491 (0.0560)	0.2492 (0.0677)	0.4020 (0.0471)	0.2005 (0.0625)	0.1017 (0.0315)	0.1437 (0.0473)
w_{21}	0.2247 (0.0512)	0.2423 (0.0482)	0.2584 (0.0634)	0.4071 (0.0537)	0.2022 (0.0611)	0.1088 (0.0361)	0.1465 (0.0441)
w_{22}	0.1234 (0.0463)	0.1505 (0.0456)	0.1993 (0.0595)	0.0978 (0.0291)	0.3045 (0.0665)	0.4074 (0.0536)	0.2842 (0.0530)

cM_A: the estimated location of the viability locus.

Table 4 Means and standard deviations (in parentheses) of estimated parameter values for the EM algorithm with sample size 200

Parameters	*Genetic model*						
	Additive (A)			*Dominance (D)*			*Both A and D*
	High	*Medium*	*Low*	*High(−)*	*Low*	*High(+)*	
cM_A	26.80 (6.43)	29.91 (15.26)	35.43 (27.34)	25.98 (1.84)	34.43 (22.34)	25.89 (1.85)	26.53 (4.34)
w^s	0.2986 (0.0643)	0.1862 (0.0837)	0.1125 (0.0907)	0.0021 (0.0765)	−0.0095 (0.0933)	−0.0050 (0.0801)	0.1528 (0.0728)
w^d	0.3024 (0.0681)	0.2177 (0.0898)	0.1040 (0.1009)	0.0081 (0.0651)	0.0068 (0.0894)	−0.0074 (0.0659)	0.1539 (0.0717)
δ	0.0017 (0.0201)	0.0019 (0.0203)	−0.0008 (0.0246)	−0.1479 (0.0169)	0.0498 (0.0276)	0.1517 (0.0148)	0.1004 (0.0167)
w_{11}	0.4255 (0.0366)	0.3639 (0.0385)	0.3073 (0.0432)	0.1058 (0.0229)	0.3004 (0.0472)	0.3998 (0.0372)	0.4339 (0.0340)
w_{12}	0.2262 (0.0337)	0.2316 (0.0413)	0.2514 (0.0412)	0.3977 (0.0365)	0.1972 (0.0382)	0.1000 (0.0195)	0.1449 (0.0262)
w_{21}	0.2281 (0.0315)	0.2473 (0.0359)	0.2471 (0.0444)	0.4007 (0.0348)	0.2054 (0.0426)	0.0988 (0.0236)	0.1454 (0.0301)
w_{22}	0.1250 (0.0237)	0.1619 (0.0315)	0.1990 (0.0390)	0.1007 (0.0239)	0.3018 (0.0402)	0.4061 (0.0347)	0.2806 (0.0309)

cM_A: the estimated location of the viability locus.

simulations have verified the derivations of our methods and the computer programs; more importantly, they have demonstrated that viability locus mapping can be accomplished following the usual approach of QTL mapping.

Discussion

The fitness considered here is a special fitness component, the viability, which relates to the change of gene frequencies in the current generation where the mapping individuals are collected. Another major fitness component is the fecundity, that is, the number of progenies produced by the individual of interest. Fecundity is also related to the change of gene frequencies, but it affects the gene frequencies in the next generation. Fecundity is measured quantitatively and thus mapping fecundity loci can be directly accomplished using standard QTL mapping approaches. Therefore, we only focused on the statistics of mapping viability loci in this study. The ultimate result of viability selection in a population is the change in gene frequencies, but if we concentrate

Table 5 Empirical statistical powers (%) under type I error rates of 0.05 and 0.01

Sample size	*Type I error*	*Genetic model*						
		Additive (A)			*Dominance (D)*			*Both A and D*
		High	*Medium*	*Low*	*High(−)*	*Low*	*High(+)*	
50	0.05	52	36	14	82	12	86	57
	0.01	29	18	2	68	2	64	37
100	0.05	82	47	13	100	19	99	86
	0.01	74	22	2	98	12	98	68
200	0.05	100	84	27	100	40	100	100
	0.01	99	68	10	100	24	100	100

Table 6 The definitions of fitness parameters for an outbred population with F founder alleles

Paternal	*Maternal*				
	w_1	w_2	$\cdots$	w_F	
w_1	$w_{11} = w_{1.}w_{.1} + \delta_{11}$	$w_{12} = w_{1.}w_{.2} + \delta_{12}$	$\cdots$	$w_{1F} = w_{1.}w_{.F} + \delta_{1F}$	$w_{1.} = \sum_{k=1}^{F} w_{1k}$
w_2	$w_{21} = w_{2.}w_{.1} + \delta_{21}$	$w_{22} = w_{2.}w_{.2} + \delta_{22}$	$\cdots$	$w_{2F} = w_{2.}w_{.F} + \delta_{2F}$	$w_{2.} = \sum_{k=1}^{F} w_{2k}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
w_F	$w_{F1} = w_{F.}w_{.1} + \delta_{F1}$	$w_{F2} = w_{F.}w_{.2} + \delta_{F1}$	$\cdots$	$w_{FF} = w_{F.}w_{.F} + \delta_{FF}$	$w_{F.} = \sum_{k=1}^{F} w_{Fk}$
	$w_{.1} = \sum_{k=1}^{F} w_{k1}$	$w_{.2} = \sum_{k=1}^{F} w_{k2}$	$\cdots$	$w_{.F} = \sum_{k=1}^{F} w_{kF}$	$w_{..} = \sum_{kl} w_{kl} = 1$

on one particular family or pedigree, the result of viability selection is the deviation of allelic segregation from the expected Mendelian ratio. The non-Mendelian segregation of a viability locus causes deviation from Mendelian segregation for markers linked to the viability locus. The viability considered in this study is defined in the adult stage (genotype). However, the statistics developed allow us to separate the gametic selection from zygotic selection. The maternal and paternal allelic effects represent the gametic selection and the dominance effect represents the zygotic selection.

The purpose of the simulation studies is to demonstrate that viability mapping can be performed in the same way as QTL mapping. There was no attempt to explore the range of parameter values in which the method works better than for other ranges. That would require extensive simulation studies with exhaustive combinations of parameter values. However, from the results of the limited simulation experiments, we conclude that the sample size should be sufficiently large to be able to detect a locus subject to a weak selection. For the parameter values selected in our simulation experiments, $n \geq 200$ seems to be reasonable.

In the evolutionary literature, the fitness of the best genotype (the maximum fitness) is usually set to unity and the fitness values of all other genotypes are then expressed as lower values than unity (Hartl and Clark, 1997). As a result of the restriction, $w_{\max} = 1$, the fitness defined in this way is called the relative fitness. The fitness values defined in this study are also relative fitness but with a different restriction, $\sum_{kl} w_{kl} = 1$. The difference in the restriction has no effect on the estimation and statistical tests. This has been verified by our simulation studies where we converted the fitness values into selection coefficients by setting $w_{\max} = 1$ and expressed the selection coefficients as $s_{kl} = 1 - w_{kl}/w_{\max}$. The estimated fitness values are very close to the true values simulated. In fact, researchers often convert the relative fitness into selection coefficients as we did in the simulations and investigate the magnitudes of the selection coefficients. In natural populations, people often concentrate on the biallelic situation with only three phenotypes: A_1A_1, A_1A_2 and A_2A_2. Using the selection coefficients, researchers are able to investigate the degree of dominance. If the A_1A_1 is the fittest genotype, the fitness values of the three genotypes are defined as $w_{11} = 1$, $w_{12} = 1 - hs$ and $w_{22} = 1 - s$, respectively, where s represents the 'additive effect' or gametic selection and h represents the 'degree of dominance' or zygotic selection. We simply used a different but more general notation to handle multiple alleles.

Mapping viability loci has only been investigated in line-crossing experiments (Fu and Ritland, 1994a; Mitchell-Olds, 1995; Vogl and Xu, 2000). Results are only rarely inferred to natural populations, which are usually outbred. A full-sib family is the simplest case that can be studied from an outbred population. This research is the first attempt to extend viability mapping to outbred populations. The results can be easily extended to more complicated outbred pedigrees, commonly seen in humans, trees and large animals. In pedigree analysis, we focus on the relative representation of founder alleles. Each founder carries two alternative alleles at any locus. Under Mendelian segregation (no viability selection), the two alleles should be equally represented in the descendents. However, if there is evidence that the two

alleles are not equally represented, the locus may be subject to viability selection. The allele comparisons from different founders can be combined to increase the power of viability locus detection. The multiple allelic model in pedigree analysis may be investigated as follows. Assume that there are $F/2$ founders with a total of F founder alleles. The model parameters may be set up in an $F \times F$ table as shown in Table 6. The fitness of genotype A_kA_l is w_{kl} for $k,l = 1,\ldots,F$, which is partitioned as $w_{kl} = w_{k.}w_{.l} + \delta_{kl}$, where $w_{k.} = w_{.k}$ is the proportion of allele A_k represented in the mapping population and $\delta_{kl} = \delta_{lk}$ is the dominance effect. Notice the symmetry of the definitions. The parameters of the viability locus are $w_{k.}$ and δ_{kl} for $k,l = 1,\ldots,F$. Restrictions are required to make the model estimable and they are $\sum_{k=1}^{F} w_{k.} = 1$ and $\sum_{k=1}^{F} \delta_{kl} = 0$ for all l. To test the hypothesis that there is no gametic selection, we test $w_{k.} - \mu_k = 0$ for all k, where μ_k is the theoretical proportion of the presence of allele A_k in the mapping population and can be calculated based on pedigree information. For example, in a diallele mating design, $\mu_k = 1/F$ for all k. To test the hypothesis of no zygotic selection, we test $\delta_{kl} = 0$ for all k and l. In fact, it is convenient to formulate viability mapping in pedigrees as a random model problem where we are interested in testing

$$\sigma^2_{\mathrm{A}} = \frac{1}{F}\sum_{k=1}^{F}(w_{k.} - \mu_k)^2 = 0 \text{ and}$$

$$\sigma^2_{\mathrm{D}} = \frac{2}{F(F+1)}\sum_{k=1}^{F}\sum_{l=k}^{F}\delta^2_{kl} = 0$$

Of course, the founder alleles cannot be traced without marker information. Inference of the relative contributions of the founder alleles in the descendents is not easy. We need to invoke the recurrent algorithm of Yi and Xu (2001) to trace the allelic origin of each allele in the mapping population. If missing markers are involved, the descent graph algorithm of Sobel and Lange (1996) is also needed. In addition, we will need to adopt the Bayesian method implemented via the Markov chain Monte Carlo (Gelman *et al*, 1995; Green, 1995; Satagopan and Yandell, 1996; Heath, 1997; Richardson and Green, 1997; Sillanpaa and Arjas, 1998; Stephens and Fisch, 1998; Vogl and Xu, 2000). The detailed algorithm has not been worked out and development of such an algorithm is our next project.

Our assumption is that segregation distortion is caused by viability selection. However, it may also be caused by genotyping errors. If genotyping errors happen randomly across loci and genotypes within loci, they may not bias our results but increase the errors of our detection and estimation. This can be compensated by increasing the sample size. However, if genotyping errors happen in a systematic manner, that is, some genotype is more often scored as another genotype, then the result will be confounded with segregation distortion. If we know this *a priori*, we may put a flag on the genotype and treat the genotype as incomplete. For example, if A_1A_2 is often scored as A_1A_1 for some markers, the experimenter should warn us that 'A_1A_1' may have a certain probability of being A_1A_2. This probability may be incorporated into our analysis. This incomplete information is still useful because we are sure that 'A_1A_1' is not A_2A_2.

We have not investigated multiple viability loci in this study, simply because the single locus model can also be applied to the search for multiple loci. This is equivalent to the interval mapping of Lander and Botstein (1989), which is a single QTL mapping procedure but has been used to search for multiple QTLs by most people. When two or more loci are located in a single chromosome but with a large distance between pairs of loci, multiple loci can be detected from multiple separated peaks of the test statistic profiles. Similar to multiple QTL mapping, when two viability loci are close together or there is an interaction (non-multiplicative) effect between the two loci, the model needs to be revised to take into consideration multiple viability loci, as the multiple QTL model of Kao *et al* (1999). The situation of multiple locus viability mapping in pedigrees is extremely complicated. A Bayesian approach should be considered (Sheehan and Thomas, 1993; Lin *et al*, 1993, 1994; Lin, 1995, 1996; Hoeschele *et al*, 1997). The ML analysis for a single locus proposed in this study serves as a necessary first step towards a full solution of viability mapping.

Acknowledgements

We are grateful to two anonymous reviewers and the editor for their thoughtful criticisms, comments and suggestions on an early version of the manuscript. This major revision of the manuscript has been greatly improved both in scientific merit and presentation by incorporating their suggestions. The research was supported by the National Institutes of Health Grant R01-GM55321 and the USDA National Research Initiative Competitive Grants Program 00-35300-9245 to SX.

References

Churchill GA, Doerge RW (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–971.

Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* **39**: 1–38.

Efron B (1979). Bootstrap methods: another look at the jacknife. *Ann Statist* **7**: 1–26.

Falconer DS, Mackay, TFC (1996). *Introduction to Quantitative Genetics*. Longman: Harlow, UK.

Fisher RA (1958). *The Genetical Theorem of Natural Selection*. Dover Publ.: New York.

Fu Y-B, Ritland K (1994a). On estimating the linkage of marker genes to viability genes controlling inbreeding depression. *Theor Appl Genet* **88**: 925–932.

Fu Y-B, Ritland K (1994b). Evidence for the partial dominance of viability genes contributing to inbreeding depression in *Mimulus guttatus*. *Genetics* **136**: 323–331.

Gelman A, Carlin JB, Stern HS, Rubin DB (1995). *Bayesian Data Analysis*. Chapman & Hall: London.

Green PJ (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**: 711–732.

Hartl DL, Clark AG (1997). *Principles of Population Genetics*. Sinauer Associates, Inc.: Sunderland, MA.

Heath SC (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* **61**: 748–760.

Hedrick PW, Muona O (1990). Linkage of viability genes to marker loci in selfing organisms. *Heredity* **64**: 67–72.

Hoeschele I, Uimari P, Grignola FE, Zhang Q, Gage KM (1997). Advances in statistical methods to map quantitative trait loci in outbred populations. *Genetics* **147**: 1445–1457.

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

Kruglyak L, Lander ES (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet* **57**: 439–454.

Lander ES, Botstein D (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185–199.

Lin S (1995). A scheme for constructing an irreducible Markov chain for pedigree data. *Biometrics* **51**: 318–322.

Lin S (1996). Multipoint linkage analysis via metropolis jumping kernels. *Biometrics* **52**: 1417–1427.

Lin S, Thompson EA, Wijsman E (1993). Achieving irreducibility of the Markov chain Monte Carlo method applied to pedigree data. *IMA J Math Appl Med Biol* **10**: 1–17.

Lin S, Thompson EA, Wijsman E (1994). Finding non-communicating sets for Markov chain Monte Carlo estimates on pedigrees. *J Am Hum Genet* **54**: 695–704.

Louis TA (1982). Finding the observed information matrix when using the EM algorithm. *J R Stat Soc B* **44**: 226–233.

Mitchell-Olds T (1995). Interval mapping of viability loci causing heterosis in *Arabidopsis*. *Genetics* **140**: 1105–1109.

Rao S, Xu S (1998). Mapping quantitative trait loci for categorical traits in four-way crosses. *Heredity* **81**: 214–224.

Richardson S, Green PJ (1997). On Bayesian analysis of mixtures with an unknown number of components. *J R Stat Soc B* **59**: 731–792.

Satagopan RJ, Yandell BS (1996). Estimation the number of quantitative trait loci via Bayesian model determination. In Special contributed paper session on Genetic Analysis of Quantitative Traits and Complex Diseases. Biometric Section, Statistical Meeting, Chicago, IL.

Sheehan N, Thomas A (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics* **49**: 163–175.

Sillanpaa MJ, Arjas E (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**: 1373–1388.

Sobel E, Lange K (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* **58**: 1323–1337.

Stephens DA, Fisch RD (1998). Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics* **54**: 1334–1347.

Vogl C, Xu S (2000). Multipoint mapping of viability and segregation distorting loci using molecular markers. *Genetics* **155**: 1439–1447.

Xu S (1998). Iteratively reweighted least squares mapping of quantitative trait loci. *Behav Genet* **28**: 341–355.

Yi N, Xu S (2001). Bayesian mapping of quantitative trait loci under complicated mating designs. *Genetics* **1574**: 1759–1771.