

A quantitative genetics model for viability selection

L Luo, Y-M Zhang and S Xu

Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

Viability selection will change gene frequencies of loci controlling fitness. Consequently, the frequencies of marker loci linked to the viability loci will also change. In genetic mapping, the change of marker allelic frequencies is reflected by the departure from Mendelian segregation ratio. The non-Mendelian segregation of markers has been used to map viability loci along the genome. However, current methods have not been able to detect the amount of selection (s) and the degree of dominance (h) simultaneously. We developed a method to detect both s and h using an F2 mating design under the classical fitness model.

We also developed a quantitative genetics model for viability selection by proposing a continuous liability controlling the viability of individuals. With the liability model, mapping viability loci has been formulated as mapping quantitative trait loci. As a result, nongenetic systematic environmental effects can be easily incorporated into the model and subsequently separated from the genetic effects of the viability loci. The quantitative genetic model has been verified with a series of Monte Carlo simulation experiments. *Heredity* (2005) 94, 347–355. doi:10.1038/sj.hdy.6800615
Published online 10 November 2004

Keywords: EM algorithm; liability model; maximum likelihood; segregation distortion; viability locus

Introduction

Natural selection directly acts on the fitness of individuals in a population. The population constantly evolves by responding to the natural selection because the variance of fitness, to some degree, is controlled by genes. There are many fitness components, viability being one of the major components. An individual is viable if it can survive to the adult stage. Therefore, the viability of an individual is simply defined as a binary variable indicating whether or not the individual has survived. To the parent of the individual or the genotype carried by the individual, the viability of the individual and its siblings determines the fecundity of the parent. Fecundity is another major fitness component of the parent. It is defined as the number of viable progeny of a parent. Fecundity may simply be treated as a quantitative trait and analyzed using quantitative genetics theory and technology. Analysis of viability selection, however, requires an entirely different technology.

Viability selection is usually studied at the population level by examining the change of gene frequencies (Hartl and Clark, 1997). Since the introduction of interval mapping for quantitative trait loci (QTL) using molecular markers (Lander and Botstein, 1989), people have attempted to map viability loci using molecular markers. Earlier works may be traced back to Hedrick and Muona (1990), who developed a maximum likelihood (ML) method to estimate the selection coefficient of a viability locus and the recombination fraction between the viability locus and a molecular marker. Following the idea of interval mapping, Hedrick and Muona (1990) developed a flanking marker analysis to estimate the

fitness parameters of a viability locus. The model examined by Hedrick and Muona (1990) is actually a complete recessive model. Fu and Ritland (1994a) showed that the parameter estimates tend to be biased if partial dominance is present. Fu and Ritland (1994a) further proposed a test for the deviation from Mendelian ratio (a form of segregation distortion as they defined it) and showed that the power has substantially increased compared to Hedrick and Muona (1990) whose test was in fact for recessiveness. Because both groups of investigators tried to estimate the recombination fraction and the selection coefficient simultaneously, there are not enough degrees of freedom to estimate the degree of dominance. Fu and Ritland (1994b) later proposed a graphical approach to investigate the degree of dominance. The position of the genotypic frequency array of the investigated population in the graph represents a different degree of dominance. Again, the method was not intended to estimate the degree of dominance, selection coefficient and recombination fraction simultaneously because of the lack of sufficient degrees of freedom to perform such estimation.

Rather than using single markers or flanking markers to estimate the fitness parameters and recombination fraction, Mitchell-Olds (1995) adopted the idea of interval mapping by examining one putative viability locus at a time and then scanning the entire genome for every putative position to provide a visual presentation of the LOD test statistic profile for identification of the viability locus. Because Mitchell-Olds was more interested in heterosis or inbreeding depression, he only fit a dominance model assuming that the fitness values of the two homozygotes are identical. Therefore, only the degree of dominance was estimated and tested. In fact, the interval mapping approach can simultaneously estimate the amount of selection and the degree of dominance. More recently, Vogl and Xu (2000) took a Bayesian approach to mapping multiple viability loci

using a backcross mating design as an example. Luo and Xu (2003) extended the ML method of Mitchell-Olds (1995) to estimate the degree of dominance and viability selection at the allelic level (a kind of additive effect). The type of population examined by Luo and Xu (2003) was a four-way cross family, which mimics an outbred full-sib family.

Other works related to viability mapping can be found in Cheng *et al* (1998), who developed an EM algorithm for estimating selection coefficient and recombination fraction in backcross and double haploid populations. Some theoretical investigation of viability selection and segregation distortion has been conducted by Feldman and Otto (1991), Jin *et al* (1994) and Asmussen *et al* (1998). A comprehensive review can be found in Carr and Dudash (2003).

Some attempt has been made to associate truncated artificial selection for a quantitative trait to viability selection for fitness (Falconer and Mackay, 1996). This has led to a way to express the selection coefficient for the recessive genotype of a viability locus as a function of the selection intensity and the genetic effect of the viability locus that has been interpreted as a QTL. This treatment requires an assumption that the quantitative trait under investigation is the target trait selected. Natural selection, however, does not act on a single trait; it affects the survivorship of individuals based on the overall performance of all fitness components or all quantitative traits. It is more legitimate to think that there is an underlying variable called the liability for each individual. The liability is continuously distributed, just like a quantitative trait, but it is hidden from us. The liability may be determined by a function of all quantitative traits. For example, we may simply imagine that the liability is a kind of Smith-Hazel 'selection index' (Hazel, 1943), but this index can only be seen by nature. If the index value of an individual is greater than a threshold, say zero, nature decides that the individual should survive to the adult stage. Otherwise, nature will eliminate this individual from the population. Such a liability model will allow us to study viability selection using typical quantitative genetics theory. Mapping viability loci can then be formulated as a problem of mapping QTL.

In this study, we first combine the complete recessive model of Hedrick and Muona (1990) and the dominance model of Mitchell-Olds (1995) to formulate a consensus model that allows simultaneous estimation and test of the selection coefficient and the degree of dominance. Since we directly estimate and test the fitness parameters, we call it the fitness model. We then develop a quantitative genetics model for viability selection by proposing an underlying liability that is targeted by natural selection. The models and methods are subsequently tested through a series of Monte Carlo simulation experiments.

Theory and methods

Mapping viability loci under the fitness model

Estimation of the selection coefficients and degree of dominance is difficult in natural populations because these parameters are confounded with gene frequencies. In very limited situations where isozyme markers are

available, the selection coefficients of these isozyme marker genotypes may be estimated and tested using ML method by treating the isozyme markers as candidate viability loci. We now have ample marker data for the purpose of mapping QTL. People often found that some regions of the chromosomes frequently show deviation from Mendelian ratio. We hypothesize that there are some viability loci distributed along these regions that cause the observed departure from Mendelian segregation ratio. Rather than throwing these markers away, we may use them to map the locations of the viability loci. In QTL mapping, we normally select two inbred lines and make a cross to generate genetically uniform F1 individuals. These F1 are selfed (in some plants) or intercrossed (in animals or some plants) to generate a segregating F2 population. In viability mapping, the purpose of the crossing experiment is to generate a population with known gene frequencies (under the hypothesis of no viability selection) so that we can exclusively estimate the fitness parameters. In an F2 population, the gene frequencies for alleles *A* and *a* are $p=1/2$ and $q=1/2$, respectively. When there is no viability selection, we expect the three genotypes to have frequencies of $P(AA)=1/4$, $P(Aa)=1/2$ and $P(aa)=1/4$, respectively. Any significant deviation from this Mendelian ratio will indicate existence of viability selection.

Let $n(AA)$, $n(Aa)$ and $n(aa)$ be the numbers of the three genotypes occurring in the population and $n(AA) + n(Aa) + n(aa) = n$, where n is the sample size of the F2 population. As usual, the relative fitnesses of the three genotypes are defined as $w(AA)=1$, $w(Aa)=1-hs$ and $w(aa)=1-s$, respectively, where s is the selection coefficient and h is the degree of dominance (Hartl and Clark, 1997). Let us define the mean fitness by

$$\begin{aligned}\bar{w} &= P(AA)w(AA) + P(Aa)w(Aa) + P(aa)w(aa) \\ &= 1/4 + (1-hs)/2 + (1-s)/4\end{aligned}\quad (1)$$

The frequencies for the three genotypes among the surviving individuals after selection will be

$$\begin{aligned}\pi(AA) &= \frac{P(AA)w(AA)}{\bar{w}} = \frac{1}{1 + 2(1-hs) + (1-s)} \\ \pi(Aa) &= \frac{P(Aa)w(Aa)}{\bar{w}} = \frac{2(1-hs)}{1 + 2(1-hs) + (1-s)} \\ \pi(aa) &= \frac{P(aa)w(aa)}{\bar{w}} = \frac{1-s}{1 + 2(1-hs) + (1-s)}\end{aligned}\quad (2)$$

In genetic mapping, the genotype of the viability locus is not observable, and thus the count data are actually missing. The log likelihood function under the assumption that these counts are observed is called the complete data likelihood, which is

$$\begin{aligned}L(s, h) &= n(AA) \ln[\pi(AA)] \\ &\quad + n(Aa) \ln[\pi(Aa)] + n(aa) \ln[\pi(aa)]\end{aligned}\quad (3)$$

This leads to the ML estimates of the genotypic frequencies in the progeny,

$$\begin{aligned}\hat{\pi}(AA) &= \frac{n(AA)}{n}, \quad \hat{\pi}(Aa) = \frac{n(Aa)}{n} \\ \hat{\pi}(aa) &= \frac{n(aa)}{n}\end{aligned}\quad (4)$$

Solving for s and h using equation (2), we get,

$$\hat{s} = \frac{\hat{\pi}(AA) - \hat{\pi}(aa)}{\hat{\pi}(AA)} \quad \text{and} \quad \hat{h} = \frac{\hat{\pi}(AA) - \hat{\pi}(Aa)/2}{\hat{\pi}(AA) - \hat{\pi}(aa)} \quad (5)$$

Since the count data are not observable, we need to substitute them by their expectations, which in turn is a function of the parameters. Therefore, we invoke an EM algorithm to obtain the solution. The E-step is to calculate

$$\begin{aligned} E[n(AA)] &= \sum_{j=1}^n P_j^*(AA) = \sum_{j=1}^n \frac{P_j(AA)\pi(AA)}{\bar{\pi}_j} \\ E[n(Aa)] &= \frac{1}{2} \sum_{j=1}^n P_j^*(Aa) = \frac{1}{2} \sum_{j=1}^n \frac{P_j(Aa)\pi(Aa)}{\bar{\pi}_j} \\ E[n(aa)] &= \sum_{j=1}^n P_j^*(aa) = \sum_{j=1}^n \frac{P_j(aa)\pi(aa)}{\bar{\pi}_j} \end{aligned} \quad (6)$$

where

$$\bar{\pi}_j = P_j(AA)\pi(AA) + \frac{1}{2}P_j(Aa)\pi(Aa) + P_j(aa)\pi(aa) \quad (7)$$

An explanation for the 1/2 that appears in front of the heterozygote is given in Appendix A. The probabilities $P_j(AA)$, $P_j(Aa)$ and $P_j(aa)$ are the probabilities of the three genotypes for individual j conditional on marker information. They are the prior probabilities before viability selection has been taken into account. Methods for calculating these three genotypes can be found in Haley and Knott (1992) under the interval mapping framework or Jiang and Zeng (1997) under the multipoint framework. The probabilities $P_j^*(AA)$, $P_j^*(Aa)$ and $P_j^*(aa)$ are the so-called posterior probabilities, which have incorporated the parameters of viability selection. The M-step is to update the parameters by

$$\begin{aligned} \hat{\pi}(AA) &= \frac{E[n(AA)]}{n}, \quad \hat{\pi}(Aa) = \frac{E[n(Aa)]}{n} \quad \text{and} \\ \hat{\pi}(aa) &= \frac{E[n(aa)]}{n} \end{aligned} \quad (8)$$

We go back and forth through the E-step and M-step until the iteration converges to a satisfactory criterion. We then use equation (5) to convert the estimated π values into s and h .

The next step is to test various hypotheses. There are many hypotheses we can test, for example, complete dominance, partial dominance, overdominance and so on. However, two of them are particularly interesting here, that is, no viability selection and no dominance. To test either hypothesis, we need to evaluate the likelihood value under the full model, that is, when both s and h are included in the analysis. This likelihood value is

$$L_{sh} = \sum_{j=1}^n \ln \left[P_j(AA)\hat{\pi}(AA) + \frac{1}{2} P_j(Aa)\hat{\pi}(Aa) + P_j(aa)\hat{\pi}(aa) \right] \quad (9)$$

Derivation of the likelihood function is given in Appendix A, which particularly explains why the 1/2 should appear in front of the heterozygote.

To test the hypothesis that there is no viability selection, we let $s=0$, which leads to the following

likelihood value under the null model:

$$\begin{aligned} L_0 &= \sum_{j=1}^n \ln \left[\frac{1}{4}P_j(AA) + \frac{1}{4}P_j(Aa) + \frac{1}{4}P_j(aa) \right] \\ &= -n \ln(4) \end{aligned} \quad (10)$$

The likelihood ratio test statistic is

$$\lambda_{sh} = -2(L_0 - L_{sh}) \quad (11)$$

To test the hypothesis that there is no dominance, we let $h=1/2$, which leads to

$$\begin{aligned} \pi(AA) &= \frac{1}{2(2-s)}, \quad \pi(Aa) = \frac{1}{2} \quad \text{and} \\ \pi(aa) &= \frac{1-s}{2(2-s)} \end{aligned} \quad (12)$$

Under this hypothesis, we only have one parameter to estimate, that is, $\pi(AA)$ or $\pi(aa)$, because $\pi(AA) + \pi(aa) = 1/2$. The likelihood value is

$$\begin{aligned} L_{h=1/2} &= \sum_{j=1}^n \ln \left[P_j(AA)\hat{\pi}(AA) + \frac{1}{4}P_j(Aa) \right. \\ &\quad \left. + P_j(aa)\hat{\pi}(aa) \right] \end{aligned} \quad (13)$$

where $\hat{\pi}(AA)$ is the MLE of $\pi(AA)$ under this hypothesis. We use the same EM algorithm as described earlier to solve for $\pi(AA)$ except that we always restrict $\pi(Aa)=1/2$ and $\pi(aa)=1/2-\pi(AA)$. The estimated selection coefficient is converted from the following equation:

$$\hat{s} = \frac{2[\hat{\pi}(AA) - 1/4]}{\hat{\pi}(AA)} \quad (14)$$

The test statistic for dominance is

$$\lambda_h = -2(L_{h=1/2} - L_{sh}) \quad (15)$$

As with the usual interval mapping of QTL, we test each putative position of the genome and use the test statistic profiles to detect and localize the viability loci. Our treatment differs from the works carried out by others (Hedrick and Muona, 1990; Fu and Ritland, 1994a,b; Mitchell-Olds, 1995) in that they only tried to estimate and test either the selection coefficient (s) or the degree of dominance (h) but not both. Work previously carried out in our lab (Vogl and Xu, 2000; Luo and Xu, 2003) dealt with either a backcross (BC) in which the degree of dominance was irrelevant or a four-way cross in which the degree of dominance was formulated differently from that in the biallelic system.

Mapping viability loci under the liability model

Systematic environmental effects may mask the effects of viability loci and cause low power of detection. It is impossible to remove the systematic error from the analysis using the classical fitness model described above. However, the liability model proposed here provides an extremely convenient way to remove such systematic errors.

Let y_j be the liability of individual j in the F2 population under study. It may be described by the following linear model:

$$y_j = \mathbf{X}_j\mathbf{b} + Z_ja + W_jd + \varepsilon_j \quad (16)$$

where \mathbf{X}_j is an incidence matrix, \mathbf{b} is a vector for systematic environmental effects, a is the additive effect, d is the dominance effect, $\varepsilon_j \sim N(0, \sigma^2)$ is a normally distributed residual error, and Z_j and W_j are defined as

$$Z_j = \begin{cases} \sqrt{2} & \text{if } j \text{ takes genotype } AA \\ 0 & \text{if } j \text{ takes genotype } Aa \\ -\sqrt{2} & \text{if } j \text{ takes genotype } aa \end{cases} \quad (17)$$

and

$$W_j = \begin{cases} -1 & \text{if } j \text{ takes genotype } AA \\ 1 & \text{if } j \text{ takes genotype } Aa \\ -1 & \text{if } j \text{ takes genotype } aa \end{cases}$$

Since the liability is simply a hypothetical variable, the residual variance can be arbitrarily defined without affecting the conclusion. For convenience, we set $\sigma^2 = 1$, and thus only a and d are unknown parameters of interest. We hypothesize that the liability is subject to natural selection. An individual will survive if $y_j \geq 0$ and will be eliminated from the population if $y_j < 0$. Since all the sampled individuals have survived from the viability selection, the liability of each observed F2 individual will follow a truncated normal distribution with a cumulative probability

$$\Pr(y_j > 0) = \Phi(\mathbf{X}_j \mathbf{b} + Z_j a + W_j d) \quad (18)$$

This may be considered as the relative fitness for individual j and thus denoted by

$$w_j = \Phi(\mathbf{X}_j \mathbf{b} + Z_j a + W_j d) \quad (19)$$

Since there are three possible genotypes for each individual, we may define

$$\begin{aligned} w_j(AA) &= \Phi(\mathbf{X}_j \mathbf{b} + \sqrt{2}a - d) \\ w_j(Aa) &= \Phi(\mathbf{X}_j \mathbf{b} + d) \\ w_j(aa) &= \Phi(\mathbf{X}_j \mathbf{b} - \sqrt{2}a - d) \end{aligned} \quad (20)$$

Define the mean of w_j by

$$\begin{aligned} \bar{w}_j &= \frac{1}{4}\Phi(\mathbf{X}_j \mathbf{b} + \sqrt{2}a - d) + \frac{1}{2}\Phi(\mathbf{X}_j \mathbf{b} + d) \\ &\quad + \frac{1}{4}\Phi(\mathbf{X}_j \mathbf{b} - \sqrt{2}a - d) \end{aligned} \quad (21)$$

We now have the following individual specific survivorship

$$\begin{aligned} \pi_j(AA) &= \frac{\frac{1}{4}w_j(AA)}{\bar{w}_j} \\ &= \frac{\Phi(\mathbf{X}_j \mathbf{b} + \sqrt{2}a - d)}{\Phi(\mathbf{X}_j \mathbf{b} + \sqrt{2}a - d) + 2\Phi(\mathbf{X}_j \mathbf{b} + d) + \Phi(\mathbf{X}_j \mathbf{b} - \sqrt{2}a - d)} \\ \pi_j(Aa) &= \frac{\frac{1}{2}w_j(Aa)}{\bar{w}_j} \\ &= \frac{2\Phi(\mathbf{X}_j \mathbf{b} + d)}{\Phi(\mathbf{X}_j \mathbf{b} + \sqrt{2}a - d) + 2\Phi(\mathbf{X}_j \mathbf{b} + d) + \Phi(\mathbf{X}_j \mathbf{b} - \sqrt{2}a - d)} \\ \pi_j(aa) &= \frac{\frac{1}{4}w_j(aa)}{\bar{w}_j} \\ &= \frac{\Phi(\mathbf{X}_j \mathbf{b} - \sqrt{2}a - d)}{\Phi(\mathbf{X}_j \mathbf{b} + \sqrt{2}a - d) + 2\Phi(\mathbf{X}_j \mathbf{b} + d) + \Phi(\mathbf{X}_j \mathbf{b} - \sqrt{2}a - d)} \end{aligned} \quad (22)$$

from which a likelihood function can be constructed,

$$\begin{aligned} L(a, d) &= \sum_{j=1}^n \ln[P_j(AA)\pi_j(AA) \\ &\quad + \frac{1}{2}P_j(Aa)\pi_j(Aa) + P_j(aa)\pi_j(aa)] \end{aligned} \quad (23)$$

Unfortunately, we do not enjoy the luxury of using EM to provide the solution. Instead, we use the simplex algorithm of Nelder and Mead (1965) to search for the solution. Hypotheses tests under the liability model follow exactly the same methods as used in QTL mapping (Lander and Botstein, 1989). The flexibility of the liability model is also reflected by the easy way of quantifying the relative importance or the genetic determination of the viability locus. As a quantitative trait, the trait variance contributed by the viability locus in the liability scale is determined by

$$h_G^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma^2} = \frac{a^2 + d^2}{a^2 + d^2 + 1} \quad (24)$$

where

$$\sigma_G^2 = \text{Var}(Z)a^2 + \text{Var}(W)d^2 = a^2 + d^2 \text{ and } \sigma^2 = 1$$

Variables Z and W are independent and both have mean 0 and variance 1 because of the special way they are defined in equation (17). With the liability model, we have unified viability locus analysis with QTL analysis. However, we gain the flexibility at the cost of computational complexity.

In the absence of systematic environmental effects, that is, $\mathbf{b} = \mathbf{0}$, the liability model is identical to the fitness model. This has been theoretically demonstrated in a previous section. If \mathbf{b} is significantly large, ignoring \mathbf{b} may inflate the residual variance to $\mathbf{b}^T \Sigma \mathbf{b} + 1$, where $\Sigma = \text{Var}(\mathbf{X}_j)$. This inflated variance will decrease the QTL effect by a factor of $(\mathbf{b}^T \Sigma \mathbf{b} + 1)^{-1/2}$. In other words, if the systematic environmental effects are ignored, the true additive effect a will be reduced to $(\mathbf{b}^T \Sigma \mathbf{b} + 1)^{-1/2}a$.

Simulation study

Statistical power of the liability model

When there are no systematic environmental effects, the fitness model and liability model are identical because one is simply a reparameterization of the other. As a result, we can focus only on the liability model and further evaluate the performance of the method. We simulated one chromosome of 100 cM long covered by 11 evenly spaced codominant markers. We put a single viability locus at position 25 cM (between markers 3 and 4). We now concentrate on the intensity of viability selection, the mode of viability selection and the sample size. The liability model allows us to use the proportion of the liability variance contributed by the QTL, denoted by h_G^2 , as a convenient measure of the selection intensity. This is due to the fact that if there is no selection, both a and d will be zero, leading to a zero h_G^2 . Three levels of h_G^2 were set up: 0.05, 0.15 and 0.25. Three modes of the viability selection were investigated: additive only (A), dominance only (D) and both additive and dominance (A and D). When both the additive and dominance effects are present, they contributed equally to the total liability variance. The sample size (n) was investigated in

three levels: 100, 200 and 300. Each parameter combination (scenario) was simulated 200 times. The performance of the method was evaluated by the statistical powers, the average estimates of the genetic parameters and the standard deviations of the estimates. The critical values of the test statistics for declaring significance were calculated using the approximate method of Piepho (2001).

Instead of evaluating all the 27 possible cases of the parameter combinations (three factors each with three levels), we first investigated the effect of sample size on the performance of the method with the selection intensity (measured by the heritability of the fitness in the scale of liability) fixed at $h_G^2 = 0.15$ and the gene action fixed at the A and D (both additive and dominance) mode. The results of 200 replicated simulations are summarized in Table 1. The means of the estimated parameters are close to the true values with the standard deviations among the replicates changing in the correct direction, that is, as sample size increases, the standard deviation decreases. The empirical statistical power also changes in the correct direction. Note that the power is quite low when the sample size was 100. A dramatic

increase in the power has been observed when the sample size increased from 100 to 200, but only a slight increase in the power has been observed when the sample size increased from 200 to 300.

When we investigated the effect of the mode of gene action on the performance of the method, we fixed the sample size at 200 and fixed the size of the viability locus at $h_G^2 = 0.15$. Table 2 shows that the estimated parameters are close to the true parametric values under all three modes of gene action. The A and D mode of gene action, however, has a substantially higher power than either mode of additive or dominance alone.

Finally, we investigated the effect of the size of the viability locus (measured by the heritability of fitness in the liability scale) on the performance of the method under the A and D mode of gene action with sample size fixed at 200. The results are summarized in Table 3. First, the mean estimates of all parameters (except the position of the viability locus) are all close to the true values. Second, the standard deviations of the estimates, in general, show a trend of increase as h_G^2 decreases. Third, the estimated position of the viability locus shows a phenomenon normally observed in QTL mapping, that

Table 1 Effect of sample size on the parameter estimation under the additive and dominance (A and D) mode of gene action when the viability locus contributes 15% of the liability variance ($h_G^2 = 0.15$)

Sample size		Power (%)	Position (cM)	a	d	$\pi(AA)$	$\pi(Aa)$	$\pi(aa)$
	True		25.00	0.2970	0.2970	0.2719	0.6109	0.1172
100	Estimate	41	25.99 (12.54)	0.3213 (0.1596)	0.3194 (0.1848)	0.2720 (0.0592)	0.6157 (0.0666)	0.1123 (0.0414)
200	Estimate	93	24.63 (5.82)	0.3087 (0.0924)	0.3217 (0.1086)	0.2694 (0.0356)	0.6185 (0.0392)	0.1122 (0.0226)
300	Estimate	95	25.26 (5.00)	0.3092 (0.0780)	0.2855 (0.0809)	0.2770 (0.0259)	0.6057 (0.0290)	0.1174 (0.0215)

The standard deviations among 200 replicated simulations are given in parentheses after the mean estimate.

Table 2 Effect of the mode of gene action on the parameter estimation when the viability locus contributes 15% of the liability variance ($h_G^2 = 0.15$) and the sample size is 200

Gene action		Power (%)	Position (cM)	a	d	$\pi(AA)$	$\pi(Aa)$	$\pi(aa)$
A and D	True		25.00	0.2970	0.2970	0.2719	0.6109	0.1172
	Estimate	93	24.63 (5.82)	0.3087 (0.0924)	0.3217 (0.1086)	0.2694 (0.0356)	0.6185 (0.0392)	0.1122 (0.0226)
Additive	True		25.00	0.4201	0.0000	0.3619	0.5	0.1381
	Estimate	80	27.08 (9.92)	0.4725 (0.1107)	-0.0025 (0.1268)	0.3725 (0.0376)	0.4993 (0.0453)	0.1282 (0.0280)
Dominance	True		25.00	0.0000	0.4201	0.1686	0.6628	0.1686
	Estimate	80	25.63 (7.47)	-0.0049 (0.1145)	0.4171 (0.0935)	0.1688 (0.0323)	0.6598 (0.0340)	0.1714 (0.0357)

The standard deviations among 200 replicated simulations are given in parentheses after the mean estimate.

Table 3 Effect of the variance contributed by the viability locus (selection intensity) on the parameter estimation under the additive and dominance (A and D) mode of gene action when the sample size is 200

h_G^2		Power (%)	Position (cM)	a	d	$\pi(AA)$	$\pi(Aa)$	$\pi(aa)$
0.25	True	—	25.00	0.4082	0.4082	0.2772	0.6436	0.0792
	Estimate	100	25.70 (3.31)	0.4182 (0.1013)	0.4091 (0.1000)	0.2786 (0.0335)	0.6422 (0.0333)	0.0792 (0.0194)
0.15	True	—	25.00	0.2970	0.2970	0.2719	0.6109	0.1172
	Estimate	93	24.63 (5.82)	0.3087 (0.0924)	0.3217 (0.1086)	0.2694 (0.0356)	0.6185 (0.0392)	0.1122 (0.0226)
0.05	True	—	25.00	0.1622	0.1622	0.2630	0.5635	0.1735
	Estimate	20	34.64 (22.61)	0.1631 (0.1134)	0.1804 (0.1481)	0.2591 (0.0438)	0.5696 (0.0572)	0.1713 (0.0395)

The standard deviations among 200 replicated simulations are given in parentheses after the mean estimate.

is, when h_C^2 is small, the estimate tends to be biased toward the center of the chromosome with a large standard deviation. Finally, the empirical statistical power shows a trend of decrease as h_C^2 decreases. When $h_C^2 = 0.05$, the power is down to 20%.

Removal of systematic environmental errors

We now add a systematic environmental error to the liability and try to remove this error using the liability model. For the same marker map and QTL position, we simulated a single QTL with $h_C^2 = 0.15$ and $n = 200$ under the additive mode of gene action. Each individual was randomly assigned one of two locations. If an individual was assigned to location one, its liability was increased by an effect b , otherwise, its liability was decreased by b . Therefore, the liability is described by the following model

$$y_j = X_j b + Z_j a + W_j d + \varepsilon_j \quad (25)$$

where $X_j = 1$ if j is in location one and $X_j = -1$ otherwise. The effect of the systematic environmental effect was examined at the following levels: $b = 0.5$ and 1.0 . The X variable defined this way has an expectation of 0 and variance of $\Sigma = 1$. If the systematic error is ignored, the residual error variance will be inflated to

$$b^T \Sigma b + 1 = b^2 + 1 \quad (26)$$

which is 1.25 when $b = 0.5$ and 2.0 when $b = 1.0$. Therefore, the QTL effect will be deflated by a factor of $1/\sqrt{1.25}$ and $1/\sqrt{2}$, respectively, for the two levels of b . We analyzed the same data set with both the correct model where the systematic error has been fully taken into account and the wrong model where the systematic error has been completely ignored. The average esti-

mated parameters and their standard deviations among 100 replicated simulations are given in Table 4. When the systematic error is large ($b = 1.0$), ignoring this error has significantly decreased the power and the accuracy of parameter estimation. However, ignoring a less important systematic error ($b = 0.5$) does not seem to cause any major problem.

For the same model, we changed the discrete X variable into a continuous one, for example, age. Without loss of generality, we simulated the continuous X from an $N(0,1)$ distribution. Again, the same levels of b were examined. The results are listed in Table 5. Virtually no differences were found between the continuous X (Table 4) and discrete X (Table 5). Our conclusion was that if there is a good reason to believe that a large systematic error exists, one should include this error in the model and try to remove it from the analysis. This kind of error, however, cannot be removed using the classical fitness model.

Multiple viability loci

Our model was developed only for a single viability locus. However, like the original interval mapping procedure (Lander and Botstein, 1989), the single locus model may be used to search for multiple loci, which are implied by multiple peaks on the test statistic profiles. We simulated two viability loci on the same chromosome, one at position 25 cM and the other at position 75 cM. Each locus contributed 0.15 of the liability variance and each with an additive mode of gene action. The simulated sample size was $n = 200$. The simulation was repeated 100 times. The average test statistic profile is depicted in Figure 1, which does show two peaks approximately at the corresponding positions where the

Table 4 Effect of systematic error on the parameter estimation when the incidence variable for the systematic error is discrete under the additive mode of gene action with $h_C^2 = 0.15$ and $n = 300$

Systematic error	Power (%)	Position	b	a	d	$\pi(AA)$	$\pi(Aa)$	$\pi(aa)$
$b = 1.0$ True		25.00	1.0	0.4201	0.0000	0.3619	0.5000	0.1381
Estimate (1)	93	25.82 (10.23)	1.1529 (0.5768)	0.4567 (0.1521)	0.0113 (0.1424)	0.3654 (0.0385)	0.5026 (0.0473)	0.1320 (0.0404)
Estimate (2)	56	26.70 (18.20)	—	0.2768 (0.0936)	-0.0030 (0.1018)	0.3259 (0.0320)	0.4988 (0.0393)	0.1753 (0.0307)
$b = 0.5$ True		25.00	0.5	0.4201	0.0000	0.3619	0.5000	0.1381
Estimate (1)	98	24.61 (6.24)	0.5102 (0.3249)	0.4293 (0.0876)	0.0124 (0.1083)	0.3605 (0.0302)	0.5042 (0.0390)	0.1352 (0.0261)
Estimate (2)	93	24.57 (6.73)	—	0.3810 (0.0758)	0.0067 (0.0983)	0.3504 (0.0286)	0.5024 (0.0364)	0.1473 (0.0233)

Estimate (1): estimate from the correct model where the systematic error has been taken care of.

Estimate (2): estimate from the wrong model where the systematic error has been ignored.

Table 5 Effect of systematic error on the parameter estimation when the incidence variable for the systematic error is continuous under the additive mode of gene action with $h_C^2 = 0.15$ and $n = 300$

Systematic error	Power (%)	Position	b	a	d	$\pi(AA)$	$\pi(Aa)$	$\pi(aa)$
$b = 1.0$ True		25.00	1.0	0.4201	0.0000	0.3619	0.5000	0.1381
Estimate (1)	93	26.57 (8.63)	1.1146 (0.5519)	0.4346 (0.1044)	0.0123 (0.1207)	0.3614 (0.0337)	0.5041 (0.0434)	0.1345 (0.0310)
Estimate (2)	70	27.53 (10.55)	—	0.3126 (0.0803)	0.0108 (0.0940)	0.3325 (0.0297)	0.5041 (0.0357)	0.1634 (0.0239)
$b = 0.5$ True		25.00	0.5	0.4201	0.0000	0.3619	0.5000	0.1381
Estimate (1)	99	25.80 (8.91)	0.5410 (0.3707)	0.4327 (0.0815)	-0.0040 (0.0934)	0.3646 (0.0278)	0.4985 (0.0343)	0.1369 (0.0234)
Estimate (2)	95	25.95 (9.56)	—	0.3908 (0.0760)	-0.0094 (0.0859)	0.3559 (0.0258)	0.4965 (0.0320)	0.1476 (0.0226)

Estimate (1): estimate from the correct model where the systematic error has been taken care of.

Estimate (2): estimate from the wrong model where the systematic error has been ignored.

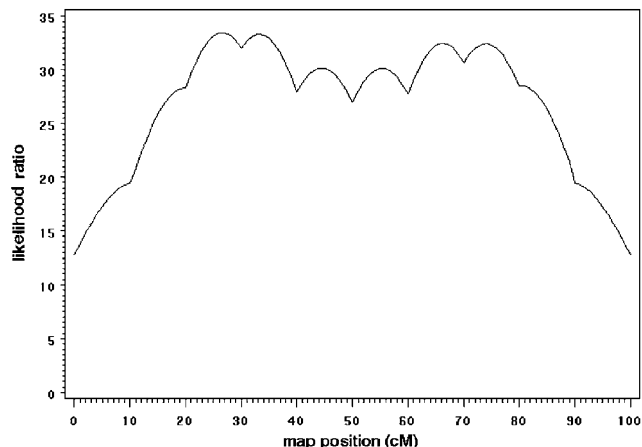


Figure 1 The average test statistic profile among 100 replicated simulation experiments for two viability loci under the liability model. The true positions of the two viability loci are at 25 and 75 cM. Each viability locus explains 0.15 of the variance of the underlying liability.

true loci were simulated. The estimated parameters of the two viability loci are given in Table 6. Both loci were identified at almost 100% power. The estimated positions for the two loci are biased toward the center of the chromosome. The estimated genetic effects are also biased upward. This simulation experiment demonstrates that the single locus model may be considered as an approximate approach to searching for multiple loci if the loci are sufficiently separated by markers.

Discussion

We developed a method to map viability loci using the classical fitness model by simultaneously estimating and testing the amount of selection and the degree of dominance, which is in contrast to previous studies where either the selection coefficient or the degree of dominance is tested but not both (Hedrick and Muona, 1990; Fu and Ritland, 1994a; Mitchell-Olds, 1995). We also developed a quantitative genetics model under the framework of liability. With a proper reparameterization, we showed that the two models are equivalent. The liability model, however, is more general than the fitness model in the sense that we can partition the total amount of selection into selection due to additive effect and selection due to dominance effect. The real motivation of developing the liability model is to unify the method of viability mapping and that of QTL mapping. We have

formulated viability mapping into a problem of QTL mapping, where the theory and methodology have been well developed. QTL mapping is so flexible that it can include nongenetic cofactors in the model so that their effects do not interfere with the result of QTL mapping. We have demonstrated this advantage in the simulation studies where inclusion or exclusion of a large systematic error does make a difference. For discrete systematic variation with a few levels, the fitness model may still be applied simply by analyzing the data separately within each environment. As the number of different environments (levels of the systematic variable) increases, the number of parameters will increase dramatically, leading to low power and inconclusive results. If the systematic environmental variable is continuous, as demonstrated by the second example, the fitness model cannot be used. However, this can be handled easily with the liability model. Any type of controllable environmental effects can be handled with the liability model. If there is a good reason to believe the existence of QTL by environment interaction, simple modification may be conducted to take care of it, but only with the liability model.

The liability model gains its flexibility at the cost of computational complexity. The fitness model parameters can be estimated using a simple EM algorithm (Dempster *et al.*, 1977). Such an EM algorithm does not exist for the liability model. We simply adopted the simplex method (Nelder and Mead, 1965) in SAS (SAS Institute, 1999) to search for the solution. The study emphasizes the quantitative genetics model, rather than the computational algorithm. Researchers interested in the computational algorithm may want to develop a more efficient algorithm, for example, Newton–Raphson ridge algorithm or Fisher-scoring algorithm, for the ML solution. Interested researchers may also want to develop a special algorithm to obtain the estimation errors of the parameters.

As demonstrated by the last simulation experiment, the interval mapping approach can be used to search for multiple viability loci. The parameter estimates, however, are biased when two loci are not far away. This observation is consistent with that found by Fu and Ritland (1994a). To reduce the bias, a multiple locus model may be required. It has proved to be difficult to implement such a multiple locus model under the ML framework, even in standard QTL mapping. With the liability model, the composite interval mapping approach (Zeng, 1994) may be adopted in which markers outside the current interval may be included in the model to control the background effects. The exact multiple locus model may be developed using the

Table 6 Results of mapping multiple viability loci using a genome scan approach

QTL		Power (%)	Position (cM)	a	d	$\pi(AA)$	$\pi(Aa)$	$\pi(aa)$
1	True		25.00	0.4201	0.0000	0.3619	0.5	0.1381
	Estimate	100	32.26 (9.88)	0.5982 (0.1147)	−0.0017 (0.1338)	0.3988 (0.0365)	0.4991 (0.0449)	0.1021 (0.0261)
2	True		75.00	0.4201	0.0000	0.3619	0.5	0.1381
	Estimate	99	66.83 (10.14)	0.5896 (0.1154)	−0.0038 (0.1479)	0.39723 (0.0418)	0.4988 (0.0499)	0.1039 (0.0251)

The estimate values were obtained by the averages of 100 replicated simulations under the additive model for two viability loci each contributing 0.15 of the liability variance. The sample size of each simulated sample was 200. The standard deviations of the estimated parameters among the 100 replicates are given in parentheses.

Bayesian method where even the number of loci can be treated as a parameter (Sillanpaa and Arjas, 1998). However, a multiplicative model must be assumed if the fitness model is to be taken, whereas, the liability model handles multiple loci simply by assuming additivity among the loci. Theoretically, interaction effects among multiple viability loci can be modeled as epistatic effects among multiple loci in the scale of liability, which has been well developed in QTL mapping (Yi and Xu, 2002; Yi *et al*, 2003).

From an evolutionary point of view, identifying viability loci is interesting in its own right. From a quantitative genetics point of view, identifying viability loci may improve the efficiency of QTL mapping. Viability loci may cause segregation distortion in markers. Using distorted markers for QTL mapping is risky because the basic assumption of Mendelian segregation is violated. Most quantitative geneticists interested in QTL mapping do not want to use distorted markers for QTL mapping. These markers are typically removed from the marker map. Unfortunately, one cannot rule out the possibility that some important QTL may reside nearby a viability locus. When the distorted markers are removed, these linked QTL will be removed as well along with the distorted markers. Therefore, too many distorted markers will cause tremendous information loss in QTL mapping. The method of viability locus mapping takes advantage of all markers, whether they follow Mendelian ratio or are distorted. Therefore, the same data set generated by experimentalists may be used by both quantitative geneticists for QTL mapping using subset of markers and naturalists for mapping viability loci using all markers. This is a 'one-stone-kills-two-birds' approach. It will be interesting to include the distorted markers also in QTL mapping to see what the effect of the distorted markers has on the result of QTL mapping. While doing this, one may take a risk of detecting false QTL not due to their genetic effects on the quantitative trait but due to violation of the Mendelian segregation law. It will be a tremendous breakthrough in the genetic mapping area if we can develop a method to separate the effects of viability loci from the effects of QTL. Because of the complexity of the combined analysis, it will be investigated separately in a future project.

Acknowledgements

We are grateful to three anonymous reviewers and the associated editor for their constructive comments on the manuscript and suggestions on the improvement of the presentation. This research was supported by the National Institutes of Health Grant GM55321 and the USDA National Research Initiative Competitive Grants Program 00-35300-9245 to SX.

References

Asmussen MA, Gilliland LU, Meagher RB (1998). Detection of deleterious genotypes in multigenerational studies. II. Theoretical and experimental dynamics with selfing and selection. *Genetics* **149**: 717–737.

- Carr DE, Dudash MR (2003). Recent approaches into the genetic basis of inbreeding depression in plants. *Philos Trans R Soc Lond B* **358**: 1071–1084.
- Cheng R, Kleinhofs A, Ukai Y (1998). Method for mapping a partial lethal-factor locus on a molecular-marker linkage map of a backcross and doubled-haploid population. *Theor Appl Genet* **97**: 293–298.
- Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* **39**: 1–38.
- Falconer DS, Mackay TFC (1996). *Introduction to Quantitative Genetics*. Longman: London.
- Feldman MW, Otto SP (1991). A comparative approach to the population-genetics theory of segregation distortion. *Am Nat* **137**: 443–456.
- Fu Y-B, Ritland K (1994a). Evidence for the partial dominance of viability genes contributing to inbreeding depression in *Mimulus gittatus*. *Genetics* **136**: 323–331.
- Fu Y-B, Ritland K (1994b). On estimating the linkage of marker genes to viability genes controlling inbreeding depression. *Theor Appl Genet* **88**: 925–932.
- Haley CS, Knott SA (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**: 315–324.
- Hartl DL, Clark AG (1997). *Principles of Population Genetics*. Sinauer Associates Inc.: Sunderland, MA.
- Hazel LN (1943). The genetic basis for constructing selection indexes. *Genetics* **28**: 476–490.
- Hedrick PW, Muona O (1990). Linkage of viability genes to marker loci in selfing organisms. *Heredity* **64**: 67–72.
- Jiang C, Zeng Z-B (1997). Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica* **101**: 47–58.
- Jin K, Speed TP, Klitz W, Thomson G (1994). Testing for segregation distortion in the HLA complex. *Biometrics* **50**: 1189–1198.
- Lander ES, Botstein D (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185–199.
- Luo L, Xu S (2003). Mapping viability loci using molecular markers. *Heredity* **90**: 459–467.
- Mitchell-Olds T (1995). Interval mapping of viability loci causing heterosis in *Arabidopsis*. *Genetics* **140**: 1105–1109.
- Nelder JA, Mead R (1965). A simplex method for function minimization. *Comput J* **7**: 308–313.
- Piepho HP (2001). A quick method for computing approximately thresholds for quantitative trait loci detection. *Genetics* **157**: 425–432.
- SAS Institute (1999). *SAS/IML User's Guide*, version 8. SAS Institute Inc., Cary.
- Sillanpaa MJ, Arjas E (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**: 1373–1388.
- Vogl C, Xu S (2000). Multipoint mapping of viability and segregation distorting loci using molecular markers. *Genetics* **155**: 1439–1447.
- Yi N, Xu S (2002). Mapping quantitative trait loci with epistatic effects. *Genet Res* **79**: 185–198.
- Yi N, Xu S, Allison DB (2003). Bayesian model choice and search strategies for mapping interacting quantitative trait loci. *Genetics* **165**: 867–883.
- Zeng Z-B (1994). Precision mapping of quantitative trait loci. *Genetics* **136**: 1457–1468.

Appendix A

Derivation of the likelihood function

This appendix provides the derivation of the likelihood function and explains why it appears to be different from the likelihood function commonly seen in F2 mapping.

The survivorships of individual j conditional on the genotype are defined as

$$\begin{aligned} \pi_j(AA) &= \frac{\Phi(\mathbf{X}_j\mathbf{b} + \sqrt{2}b_1 - b_2)}{\Phi(\mathbf{X}_j\mathbf{b} + \sqrt{2}b_1 - b_2) + 2\Phi(\mathbf{X}_j\mathbf{b} + b_2) + \Phi(\mathbf{X}_j\mathbf{b} - \sqrt{2}b_1 - b_2)} \\ \pi_j(Aa) &= \frac{2\Phi(\mathbf{X}_j\mathbf{b} + b_2)}{\Phi(\mathbf{X}_j\mathbf{b} + \sqrt{2}b_1 - b_2) + 2\Phi(\mathbf{X}_j\mathbf{b} + b_2) + \Phi(\mathbf{X}_j\mathbf{b} - \sqrt{2}b_1 - b_2)} \\ \pi_j(aa) &= \frac{\Phi(\mathbf{X}_j\mathbf{b} - \sqrt{2}b_1 - b_2)}{\Phi(\mathbf{X}_j\mathbf{b} + \sqrt{2}b_1 - b_2) + 2\Phi(\mathbf{X}_j\mathbf{b} + b_2) + \Phi(\mathbf{X}_j\mathbf{b} - \sqrt{2}b_1 - b_2)} \end{aligned} \quad (\text{A.1})$$

However, the survivorship of the heterozygote is the sum of the two phase specific survivorships, that is,

$$\pi_j(Aa) = \pi_j\{Aa\} + \pi_j\{aA\} \quad (\text{A.2})$$

where

$$\pi_j\{Aa\} = \pi_j\{aA\} = \frac{1}{2}\pi_j(Aa) \quad (\text{A.3})$$

Similarly, the probability of heterozygote conditional on markers should also be seen as the sum of two phase specific probabilities, that is,

$$P_j(Aa) = P_j\{Aa\} + P_j\{aA\} \quad (\text{A.4})$$

where

$$P_j\{Aa\} = P_j\{aA\} = \frac{1}{2}P_j(Aa) \quad (\text{A.5})$$

The log likelihood function is actually constructed by taking into account all the four possible genotypes for each individual:

$$\begin{aligned} L &= \sum_{j=1}^n \ln[P_j(AA)\pi_j(AA) \\ &\quad + P_j\{Aa\}\pi_j\{Aa\} + P_j\{aA\}\pi_j\{aA\} + P_j(aa)\pi_j(aa)] \\ &= \sum_{j=1}^n \ln\left[P_j(AA)\pi_j(AA) + \frac{1}{4}P_j(Aa)\pi_j(Aa)\right] \end{aligned}$$

$$\begin{aligned} &+ \frac{1}{4}P_j(Aa)\pi_j(Aa) + P_j(aa)\pi_j(aa) \Big] \\ &= \sum_{j=1}^n \ln\left[P_j(AA)\pi_j(AA) + \frac{1}{2}P_j(Aa)\pi_j(Aa) + P_j(aa)\pi_j(aa)\right] \end{aligned} \quad (\text{A.6})$$

This is contradictory to the following wrong likelihood function that appears to be logical:

$$\begin{aligned} L &= \sum_{j=1}^n \ln[P_j(AA)\pi_j(AA) \\ &\quad + P_j(Aa)\pi_j(Aa) + P_j(aa)\pi_j(aa)] \end{aligned} \quad (\text{A.7})$$

For the same reason, the posterior probabilities of the genotypes conditional on markers and the QTL parameters are

$$\begin{aligned} P_j^*(AA) &= \frac{P_j(AA)\pi_j(AA)}{P_j(AA)\pi_j(AA) + \frac{1}{2}P_j(Aa)\pi_j(Aa) + P_j(aa)\pi_j(aa)}, \end{aligned} \quad (\text{A.8})$$

$$\begin{aligned} P_j^*(Aa) &= P_j^*\{Aa\} + P_j^*\{aA\} \\ &= \frac{P_j\{Aa\}\pi_j\{Aa\} + P_j\{aA\}\pi_j\{aA\}}{P_j(AA)\pi_j(AA) + P_j\{Aa\}\pi_j\{Aa\} + P_j\{aA\}\pi_j\{aA\} + P_j(aa)\pi_j(aa)} \\ &= \frac{\frac{1}{2}P_j(Aa)\pi_j(Aa)}{P_j(AA)\pi_j(AA) + \frac{1}{2}P_j(Aa)\pi_j(Aa) + P_j(aa)\pi_j(aa)} \end{aligned} \quad (\text{A.9})$$

and

$$\begin{aligned} P_j^*(aa) &= \frac{P_j(aa)\pi_j(aa)}{P_j(AA)\pi_j(AA) + \frac{1}{2}P_j(Aa)\pi_j(Aa) + P_j(aa)\pi_j(aa)} \end{aligned} \quad (\text{A.10})$$