Multipoint Mapping of Viability and Segregation Distorting Loci Using Molecular Markers

Claus Vogl*,† and Shizhong Xu†

**Department of Biology, University of Oulu, FIN-90401 Oulu, Finland and †Department of Botany and Plant Sciences, University of California, Riverside, California 92521*

Manuscript received July 13, 1998
Accepted for publication April 3, 2000

ABSTRACT

In line-crossing experiments, deviations from Mendelian segregation ratios are usually observed for some markers. We hypothesize that these deviations are caused by one or more segregation-distorting loci (SDL) linked to the markers. We develop both a maximum-likelihood (ML) method and a Bayesian method to map SDL using molecular markers. The ML mapping is implemented via an EM algorithm and the Bayesian method is performed via the Markov chain Monte Carlo (MCMC). The Bayesian mapping is computationally more intensive than the ML mapping but can handle more complicated models such as multiple SDL and variable number of SDL. Both methods are applied to a set of simulated data and real data from a cross of two Scots pine trees.

CHROMOSOMAL regions that cause distorted segregation ratios in early life stages may be referred to as segregation-distorting loci (SDL). These distortions are caused either by differential representation of SDL genotypes in gametes before fertilization or by viability differences of SDL genotypes after fertilization but before genotype scoring. In both cases, the observable phenotype is a distortion of marker locus genotypes in chromosomal regions close to the SDL. Hence, regardless of the timing of action of the SDL, mapping of locations and estimation of effects of SDL follow the same statistical treatment.

Let us first discuss mechanisms that cause deviated segregation ratios by altering the gametic proportions. With meiotic drive, gametic proportions become distorted during meiosis because one chromosome type may preferentially end up in the egg nucleus (meiotic drive). Meiotic drive is known, *e.g.*, for the maize chromosome 10 where a variant carrying a heterochromatic knob is preferentially transmitted (reviewed in Grant 1975). Gametes carrying a certain allele act to render gametes carrying the homologous chromosome, *e.g.*, the segregation distorter (SD) and sex ratio (SR) loci of Drosophila and the *t*-alleles of mice (*e.g.*, Hartl and Clark 1997, p. 244ff). Meiotic drive can be a powerful selective force. The *t*-alleles are maintained in the population, even though they are homozygous lethals, due to their 0.95 probability of being passed to the next generation in heterozygotes. In many species hybridizations, outbreeding depression and segregation distortion have been observed in the F_2 generation. These are often caused by structural differences between chromosomes (Whitkus 1998), *i.e.*, by events before fertilization.

Haploid life stages can be exposed to selection, especially in plants. In the life cycle of mosses, the haploid life stage (the gametophyte) is dominant over the diploid life stage (the sporophyte). In vascular plants, maize gametophytic mutations indicate that pollen tube growth rates are determined in part by the genotypes of the microgametophytes (reviewed in Grant 1975).

Viability selection after fertilization may be more important than gametic selection. Viability selection is common in consanguinous matings where inbreeding depression reduces the survival of homozygotes compared to heterozygotes (Charlesworth and Charlesworth 1987). Viability selection gives rise to segregation ratios distorted from 1:2:1 at linked loci. Inbreeding depression is often expressed in very early life stages (Husband and Schemske 1996). In Scots pine, only ~15% of self-fertilized embryos develop into mature seeds, whereas ~75% do so in wind-pollinated seeds (Kärkkäinen *et al.* 1996). Some aspects of the genetic basis of inbreeding depression require further investigation, *e.g.*, number and effects of loci and degree of dominance. Yet these factors have major consequences for mating system evolution (Charlesworth and Charlesworth 1998), conservation genetics (Hedrick 1994), and plant breeding (*e.g.*, Williams and Savolainen 1996). A biased segregation ratio due to viability differences of genotypes also occurs in the F_2 generation of wide crosses. This is generally thought to be caused by epistatic interactions.

Often events before fertilization cannot be distinguished from events after fertilization. McColdrick and Hedgecock (1997) reported that crosses of *Crassos-*

Corresponding author: Claus Vogl, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521.
E-mail: claus@genetics.ucr.edu

trea gigas, the Pacific oyster, produced biased segregation ratios when tested as adults. Later Launey and Hedgecock (1999) showed that, for many loci, the ratios were Mendelian when 6-hr-old larvae were assessed, but the ratios deviated from the Mendelian ratios when the animals were 2 to 3 mo old in the same crosses. Hence, the differences are due to post-fertilization viability selection.

Quantitative trait loci (QTL) are usually mapped in agronomically important plants and animals. To increase differences of parental types, and thus to increase the power of mapping, crosses are often conducted between inbred lines or between distantly related cultivars or even between species. As discussed above, these conditions promote segregation distortion.

For molecular characterization of the genetic causes of distorted segregation ratios, mapping of the location and effects of SDL would be desirable. As the phenotype in SDL mapping is different from that of QTL mapping (data in SDL mapping usually consist of frequencies of genotypes among survivors), QTL methods cannot be used for SDL mapping. Development of advanced methods for estimation of locations and effects of SDL has been lagging behind that for QTL mapping. In the past, often a single marker was considered at a time, where only the linkage between one fully informative marker and a single SDL was tested (Sorensen 1967; Servitová and Cetl 1984; Hedrick and Muona 1990; Fu and Ritland 1994a; Kärkkäinen *et al.* 1999). In a single-marker test, the number of distinguishable genotypic configurations of the marker is at best equal to the number of genotypic configurations of a linked SDL, but the genotypic frequencies of the marker are affected by the recombination fraction in addition to the frequencies of the SDL's genotypic configurations. Hence, for a single-marker test, estimations of the position and effect are confounded.

Errors in marker genotyping may also cause systematic deviations from the expected segregation ratio. Randomly amplified polymorphic DNA (RAPD) markers are often misscored as a faint band and may be interpreted as absent. This may lead to misscoring of only a single marker. In contrast, if segregation distortion is caused by SDL, all markers in the vicinity of the SDL will be affected.

Fu and Ritland (1994b), Mitchell-Olds (1995), and Cheng *et al.* (1996) have developed maximum-likelihood methods for mapping one SDL using flanking markers, *i.e.*, an interval mapping strategy (Lander and Botstein 1989). Given a map of fully informative markers, no missing data, no interference between recombinations, and no more than one SDL per chromosome, this theory can be used to scan the genome for SDL. Under these assumptions, loci outside the interval flanking the SDL contribute no information to the segregation of the SDL. But more than one SDL per chromosome may be present and markers may be only partially informative. Furthermore, due to the effects of SDL, estimation of map distances of markers might become biased (Lorieux *et al.* 1995a,b; Liu 1998). This might cause the interval mapping method to become inefficient and biased.

The SDL analysis is based on binomial (or multinomial) distributions instead of normal distributions, and hence multiple regression is not readily available and cannot be combined with conventional interval mapping as in the composite interval mapping (CIM; Zeng 1994) or the multiple QTL mapping (MQM) scheme (Jansen and Stam 1994). Therefore, multiple SDL on a single chromosome pose an unsolved theoretical problem. On the other hand, if maps are inferred correctly and if SDL on different chromosomes do not interact epistatically, *i.e.*, SDL effects combine multiplicatively, linkage to an SDL is solely responsible for the phenotype. SDL analysis of one chromosome is therefore usually independent from other chromosomes.

We present a multipoint method for mapping multiple SDL using a backcross design. The multipoint method is developed under both the maximum-likelihood and the Bayesian frameworks.

THEORY

Model: We develop and present the model under a backcross design only, although the method can be applied to other controlled mating designs as well. We assume that the parents that initiate the cross are pure inbred lines. The F_1 of the cross is backcrossed to one of the parents and a total of N individuals are generated in the backcross (BC) family for mapping. We are interested in mapping loci responsible for segregation distortion using multiple markers that are already mapped on the genome. The data here are the observed marker genotypes (configurations). The parameters, however, are the number of SDL, the locations, and effects of these loci. We assume that all markers are neutral in the sense that their segregations would be Mendelian if there were no linked SDL on the same chromosome. The observed segregation distortions on these neutral markers, however, are caused by one or more SDL near the markers.

Note that the flow of causality is from the SDL to the genotypic configurations of the SDL, then from the genotypic configurations of the SDL to the genotypic configurations of the marker loci, and finally from the genotypic configurations of the marker loci to the observed marker information. We first consider a single SDL. The genotype of the F_1 is heterozygous and that of a BC individual (generated from F_1 backcrossed to the first inbred parent) is either heterozygous or homozygous for the allele of the first parent with an unequal probability. The degree of asymmetry in the probability is determined by the effect (size) of the SDL. Define

$$\varphi_i = \begin{cases} 0 & \text{if } i \text{ is homozygous} \\ 1 & \text{if } i \text{ is heterozygous} \end{cases}$$

for $i = 1, \ldots, N$. This indicator variable, φ_i, is also called the "inheritance digit" because it indicates which of the two alleles carried by the F_1 has been inherited to the ith progeny. Parameters of interest are the *effect*, denoted by π, and *location*, denoted by λ of the segregation distorting locus. The distribution of φ_i is Bernoulli with

$$\Pr(\varphi_i|\pi,\lambda) = \Pr(\varphi_i|\pi) = \pi^{1-\varphi_i}(1-\pi)^{\varphi_i} \quad (1)$$

for $i = 1, \ldots, N$, with

$$\pi = \Pr(\varphi_i = 0). \quad (2)$$

Note that in the SDL case the distribution of the inheritance digit of the SDL given π is independent of the location. Another parameter of interest is the *location* of the SDL on the chromosome, denoted by λ, which will be dealt with later. In the absence of segregation distortion, we have $\pi = \frac{1}{2}$. Therefore, the deviation of π from $\frac{1}{2}$ is the effect or size of the SDL. If φ_i were observable, we could directly estimate and test π. The maximum-likelihood estimate would be

$$\hat{\pi} = \frac{1}{N}\sum_{i=1}^{N}(1-\varphi_i) \quad (3)$$

if we could maximize the following log-likelihood function:

$$\begin{aligned} l(\pi|\varphi) &= \sum_{i=1}^{N} \ln \Pr(\varphi_i|\pi) \\ &= \sum_{i=1}^{N}[(1-\varphi_i)\log\pi + \varphi_i \log(1-\pi)]. \end{aligned} \quad (4)$$

But φ_i is not observable; only the inheritance digits of marker alleles can be observed. Therefore, an entirely different approach is required to estimate π. Consider M markers with known map positions on the chromosome of interest. Define the inheritance digits of the ith individual at the jth marker locus as

$$\phi_{ij} = \begin{cases} 0 & \text{if } i \text{ is homozygous for marker } j \\ 1 & \text{if } i \text{ is heterozygous for marker } j \end{cases}$$

for $i = 1, \ldots, N$. Without genotyping errors, there are just three possibilities of marker information I_{ij} of the ith individual at the jth marker locus. The first two cases are mutually exclusive events: either one or the other marker inheritance digit is observed. In the third case of a missing observation, we define the marker information as the union of the former two cases. Thus, $\Pr(I_{ij}|\phi_{ij}) = 1$ if the marker information is compatible with ϕ_{ij} and $\Pr(I_{ij}|\phi_{ij}) = 0$ otherwise. In the latter case, $\Pr(I_{ij}|\phi_{ij}) = 1$ is equal to 1 independent of the inheritance digit. If there are genotyping errors $\Pr(I_{ij}|\phi_{ij})$ will assume values intermediate between 0 and 1. Note that $\Pr(I_{i1}, \ldots, I_{iM}|\phi_{i1}, \ldots, \phi_{iM}) = \Pi_{j=1}^{M} \Pr(I_{ij}|\phi_{ij})$ because conditional on the jth inheritance digit the jth marker information is independent from all other variables.

Given the position (λ) of the SDL on the chromosome, the joint distribution for φ_i and $\phi_{i1}, \ldots, \phi_{iM}$ is

$$\Pr(\phi_{i1}, \ldots, \phi_{iM}, \varphi_i|\pi,\lambda) = \Pr(\phi_{i1}, \ldots, \phi_{iM}|\varphi_i,\lambda)\Pr(\varphi_i|\pi), \quad (5)$$

where $\Pr(\phi_{i1}, \ldots, \phi_{iM}|\varphi_i, \lambda)$ can be found using the property of a two-state Markov chain (Lander and Green 1987; Jiang and Zeng 1997). We assume that there is no interference between two consecutive crossovers so that Haldane's mapping function applies. Under this assumption, the sequence

$$\{\phi_{i1}, \ldots, \phi_{ik}, \varphi_i, \phi_{i(k+1)}, \ldots, \phi_{iM}\}$$

forms a Markov chain with two discrete states, where the markers are ordered according to their positions on the chromosome and the SDL is located between markers k and $k+1$. We, thus, have

$$\begin{aligned} \Pr(\phi_{i1}, \ldots, \phi_{iM}|\varphi_i,\lambda) = &\left[\Pr(\phi_{ik}|\varphi_i,\lambda)\prod_{j=1}^{k-1}\Pr(\phi_{ij}|\phi_{i(j+1)})\right] \\ &\times\left[\Pr(\phi_{i(k+1)}|\varphi_i,\lambda)\prod_{j=k+1}^{M-1}\Pr(\phi_{i(j+1)}|\phi_{ij})\right], \end{aligned} \quad (6)$$

where

$$\Pr(\phi_{ij}|\phi_{i[j+1)}) = \begin{cases} 1-r_{j(j+1)} & \text{if } \phi_{ij} = \phi_{i(j+1)} \\ r_{j(j+1)} & \text{if } \phi_{ij} \neq \phi_{i(j+1)} \end{cases}$$

is the transition probability between two consecutive loci and $r_{j(j+1)}$ is the recombination fraction between loci j and $j+1$. The transition probability between the SDL and the nearby marker k is

$$\Pr(\phi_{ik}|\varphi_i,\lambda) = \begin{cases} 1-r_{kl} & \text{if } \phi_{ik} = \phi_i \\ r_{kl} & \text{if } \phi_{ik} \neq \phi_i, \end{cases}$$

where r_{kl} is the recombination fraction between the kth marker and the SDL identified as locus l. The transition probability between the SDL and the $(k+1)$th locus is obtained similarly.

Let $I_i = [I_{i1}, \ldots, I_{iM}]$. Combining formula (6) with the marker information and "summing out" the marker inheritance digits, we get

$$\Pr(I_i|\varphi_i,\lambda) = \sum_{\phi_{i1}}\cdots\sum_{\phi_{iM}}\left(\Pr(\phi_{i1}, \ldots, \phi_{iM}|\varphi_i,\lambda)\prod_{j=1}^{M}\Pr(I_{ij}|\phi_{ij})\right),$$

where we have made use of the independence from other markers of the jth marker information conditional on the jth marker inheritance digit. Combining the previous formula with formula (5) results in the following equation:

$$\begin{aligned} \Pr(I_i,\varphi_i|\pi,\lambda) &= \Pr(I_i|\varphi_i,\lambda)\Pr(\varphi_i|\pi) \\ &= \Pr(I_i|\varphi_i,\lambda)\pi^{(1-\varphi_i)}(1-\pi)^{\varphi_i}. \end{aligned} \quad (7)$$

Maximum likelihood: Having formulated the proba-

bility model, we now introduce a maximum-likelihood method to estimate and test the SDL. There are several ways to find the maximum-likelihood estimate of π; we adopt an expectation maximization (EM) algorithm and treat φ_i as missing data. We treat λ as a known constant for the moment. Let $I = [I_1, \ldots, I_N]$ and $\varphi = [\varphi_1, \ldots, \varphi_N]$. For the EM algorithm we need to determine the logarithm of $\Pr(I, \varphi|\pi, \lambda)$, *i.e.*,

$$\log \Pr(I,\varphi|\pi,\lambda) = \sum_{i=1}^{N} \log[\Pr(I_i|\varphi_i,\lambda)\pi^{(1-\varphi_i)}(1-\pi)^{\varphi_i}] = \text{const} + \sum_{i=1}^{N}[(1-\varphi_i)\log(\pi) + \varphi_i \log(1-\pi)]. \tag{8}$$

The constant does not depend on the parameter of interest, π.

Conditional on the data, the position, and the initial value of the parameter, $\pi^{(0)}$, the posterior probabilities of $\varphi_i = 0$ and $\varphi_i = 1$ are, respectively,

$$\Pr(\varphi_i = 0|I_i, \pi^{(0)}, \lambda) = \frac{\Pr(I_i|\varphi = 1, \lambda)(1-\pi^{(0)})}{\Pr(I_i|\varphi = 0, \lambda)\pi^{(0)} + \Pr(I_i|\varphi = 1, \lambda)(1-\pi^{(0)})} \tag{9a}$$

and

$$\Pr(\varphi_i = 1|I_i, \pi^{(0)}, \lambda) = \frac{\Pr(I_i|\varphi = 0, \lambda)\pi^{(0)}}{\Pr(I_i|\varphi = 0, \lambda)\pi^{(0)} + \Pr(I_i|\varphi = 1, \lambda)(1-\pi^{(0)})}. \tag{9b}$$

Because $\Pr(\varphi_i \mid I_i, \pi^{(0)}, \lambda)$ follows a Bernoulli distribution, the probability in (9a) is equivalent to the expectation $E[\varphi_i|I_i, \pi^{(0)}, \lambda] = \hat{\varphi}_i^{(0)}$. Taking the expectation of (8) with respect to φ and substituting φ_i into the resulting formula, we have completed the expectation step in the EM-algorithm. The M-step consists of maximizing the resulting equation to obtain

$$\pi^{(1)} = \frac{\sum_i^N (1-\hat{\varphi}^{(0)})_i}{N}. \tag{10}$$

Equations 9 and 10 are iterated until convergence.

We can now test the null hypothesis that there is no segregation distortion for the particular location λ. The null hypothesis is formulated as H_0: $\pi = \frac{1}{2}$, which can be tested using the likelihood-ratio test statistic $\Lambda = -2(l(\frac{1}{2}, \lambda) - l(\hat{\pi}, \lambda))$, where $l(\hat{\pi}, \lambda)$ is the log likelihood

$$\log \Pr(I|\pi,\lambda) = \sum_{i=1}^{N} \log\left[\sum_{\varphi_i} \Pr(I_i,\varphi_i|\pi,\lambda)\right] \tag{11}$$

evaluated at the maximum-likelihood estimate $\hat{\pi}$, and $l(\frac{1}{2}, \lambda) = N \log(\frac{1}{2})$ is the log-likelihood value under Mendelian segregation. Under the null model, Λ is approximately distributed as a chi-square variable with 1 d.f.

The maximum-likelihood estimate of the position of the SDL, λ, can be obtained by examining the likelihood-ratio profile along the chromosome, as is commonly done in interval mapping of QTL.

Bayesian analysis: We now introduce the Bayesian analysis of SDL implemented via the Markov chain Monte Carlo (MCMC). We first classify variables into observables and unobservables. The observables are the data, denoted by *I*. The unobservables include parameters and missing information. The parameters here include π and λ, and the missing information consists of the inheritance digits ϕ and ϕ in the current situation. We always sum over all the missing information, such that inheritance digits will only appear in intermediate steps. The joint posterior distribution of the parameters is

$$\Pr(\pi, \lambda|I) \propto \Pr(\pi)\Pr(\lambda)\prod_{i=1}^{N}\Pr(I_i|\pi, \lambda) = \Pr(\pi)\Pr(\lambda)\prod_{i=1}^{N}\sum_{\varphi_i}\Pr(I_i|\varphi_i, \lambda)\Pr(\varphi_i|\pi), \tag{12}$$

where $\Pr(\pi)$ and $\Pr(\lambda)$ are the prior distributions for the parameters of interest; beta with Beta(1, 1) for the former and uniform for the latter. Samples are simulated from the joint posterior distribution via the MCMC. In the MCMC analysis, instead of sampling all the unobservables simultaneously, we sample one unobservable at a time with others taking values simulated in the previous cycle. When all the unobservables are updated, we have completed one cycle of the Markov chain. When the chain reaches a stationary stage, subsequent samples are considered to be drawn from the joint posterior distribution.

Starting with an initial value for each parameter, $\{\pi^{(0)}, \lambda^{(0)}\}$, we sample π using the Metropolis-Hastings algorithm (*e.g.*, Gelman *et al.* 1995). A new proposal, π^*, is sampled from a beta proposal distribution $J(\pi^*|\pi^{(0)}) = \text{Beta}(\pi^{(0)}N + 2, (1-\pi^{(0)})N + 2)$. The proposal π^* is accepted with probability $\min\{1, a(\pi^*, \pi^{(0)})\}$, where

$$a(\pi^*, \pi^{(0)}) = \frac{\Pr(\pi^*, \lambda^{(0)}|I)}{\Pr(\pi^{(0)}, \lambda^{(0)}|I)} \frac{J(\pi^{(0)}|\pi^*)}{J(\pi^*|\pi^{(0)})}. \tag{13}$$

Note that the first term is the ratio of posterior probabilities of the parameters and the second term is the ratio of proposal probabilities. If π^* is accepted, we take $\pi^{(1)} = \pi^*$; otherwise we do not update the effect of the SDL and simply take $\pi^{(1)} = \pi^{(0)}$. The beta proposal distribution assures that $0 \le \pi \le 1$. The simulated value of π, denoted by $\pi^{(1)}$, is then used to generate λ. We use the Metropolis algorithm (*e.g.*, Gelman *et al.* 1995). First, a new value of λ is proposed by a small perturbation from $\lambda^{(0)}$, *i.e.*,

$$\lambda^* = \lambda^{(0)} \pm x,$$

where x is a uniform variable sampled from $U(0, d)$ and d is a small positive number, *e.g.*, 0.1 times the length

of the linkage group. We accept this proposal with probability $\min\{1, a(\lambda^*, \lambda^{(0)})\}$, where

$$a(\lambda^*, \lambda^{(0)}) = \frac{\Pr(\lambda^*, \pi^{(1)}|I)}{\Pr(\lambda^{(0)}, \pi^{(1)}|I)}. \quad (14)$$

If λ^* is accepted, we take $\lambda^{(1)} = \lambda^*$; otherwise $\lambda^{(1)} = \lambda^{(0)}$.

Multiple-SDL model: Consider the joint action of L SDL located on the chromosome of interest. Define the locations of these SDL by $\boldsymbol{\lambda} = \{\lambda_l\}$ for $l = 1, \ldots, L$, in contrast to the single-SDL model where λ is a scalar. Also define the marginal effects of the SDL by $\boldsymbol{\pi} = \{\pi_l\}$ for $l = 1, \ldots, L$. Assume that these SDL act multiplicatively then the joint effect of all the SDL can be formulated as a product of these marginal effects. Define $\varphi_i = [\varphi_{i1}, \ldots, \varphi_{iL}]$ and $\phi_i = [\phi_{i1}, \ldots, \varphi_{iM}]$ as vectors of inheritance digits of all SDL and marker loci, respectively, for the ith individual. Using Bayes' theorem, the joint posterior distribution of φ_i can be formulated as

$$\Pr(\varphi_i|\boldsymbol{\pi}, \boldsymbol{\lambda}) = \frac{(\prod_{l=1}^{L-1} \Pr(\varphi_{i(l+1)}|\varphi_{il}, \lambda))\prod_{l=1}^{L}\Pr(\varphi_{il}|\pi_l)}{\sum_{\phi_i}(\prod_{l=1}^{L-1}\Pr(\varphi_{i(l+1)}|\varphi_{il}, \lambda))\sum_{l=1}^{L}\Pr(\varphi_{il}|\pi_l)}. \quad (15)$$

The joint posterior distribution of the parameters is

$$\Pr(\boldsymbol{\pi}, \boldsymbol{\lambda}|I) \propto \Pr(\boldsymbol{\pi})\Pr(\boldsymbol{\lambda})\prod_{i=1}^{N}\sum_{\varphi i}(\Pr(I_i|\varphi_i, \lambda)\Pr(\phi_i|\boldsymbol{\pi}, \boldsymbol{\lambda})), \quad (16)$$

where $\Pr(\boldsymbol{\pi}) = \Pi_{l=1}^{L}\Pr(\pi_l)$, $\Pr(\boldsymbol{\lambda}) = \Pi_{l=1}^{L}\Pr(\lambda_l)$, and

$$\Pr(I_i|\varphi_i, \boldsymbol{\lambda}) = \frac{\Pr(I_i, \varphi_i|\boldsymbol{\lambda})}{\Pr(\varphi_i|\boldsymbol{\lambda})} = \frac{\sum_{\varphi i}(\Pr(\varphi_i, \varphi_i|\boldsymbol{\lambda})\prod_j \Pr(I_{ij}|\phi_{ij}))}{\Pr(\varphi_i|\boldsymbol{\lambda})}. \quad (17)$$

Under the multiple-SDL model, formulation of an EM algorithm seems impossible. On the other hand, the Bayesian method requires little modification: instead of updating the effect and location of a single locus at a time, $\boldsymbol{\lambda}$ and $\boldsymbol{\pi}$ are updated iteratively for all loci.

With the Bayesian approach, the number of SDL (L) can be treated as an unknown variable. This involves a change in the dimension of the model. Reversible jump MCMC (Green 1995; Satagopan and Yandell 1996; Heath 1997; Richardson and Green 1997; Sillanpää and Arjas 1998; Stephens and Fisch 1998) is an extension to the Metropolis-Hastings sampler, permitting moves to be made between models with different dimensions. The joint posterior distribution of the parameters is

$$\Pr(\boldsymbol{\pi}, \boldsymbol{\lambda}, L|I) \propto \Pr(\pi|L)\Pr(\boldsymbol{\lambda}|L)\Pr(L) \times \prod_{i=1}^{N}\sum_{\varphi i}(\Pr(I_i|\varphi_i, \boldsymbol{\lambda}, L)\Pr(\varphi_i|\boldsymbol{\pi}, \boldsymbol{\lambda}, L)), \quad (18)$$

where $\Pr(L)$ is the prior probability of the number of SDL. We chose a Poisson prior (with mean $\mu = 1$) for $\Pr(L)$ truncated at $L_{\max}$. After each existing SDL has been updated, we propose two types of move to update L, adding a locus if $L < L_{\max}$ (with probability p_a) and deleting a locus if $L > 0$ (with probability p_d).

For adding an SDL, a new location $\boldsymbol{\lambda}_{L+1}$ and effect $\boldsymbol{\pi}_{L+1}$ are sampled from their uniform priors for the new SDL. The new sets of parameters are $\boldsymbol{\pi}^* = (\boldsymbol{\pi}^{(0)}, \boldsymbol{\pi}_{L+1})$ and $\boldsymbol{\lambda}^* = (\boldsymbol{\lambda}^{(0)}, \boldsymbol{\lambda}_{L+1})$. We then accept this new SDL with probability $\min\{1, a(L + 1, L)\}$, where

$$a(L + 1, L) = \frac{\Pr(I|\boldsymbol{\pi}^*, \boldsymbol{\lambda}^*, L + 1)}{\Pr(I|\boldsymbol{\lambda}^{(0)}, L)} \frac{1}{L + 1} \frac{p_d}{p_a}. \quad (19)$$

If the new SDL is accepted, its location and effect are accepted simultaneously; otherwise, the number of SDL remains the same. In the deleting step, a random SDL is proposed to be deleted. Then the SDL are renumbered such that the candidate SDL is the last SDL, *i.e.*, the Lth SDL. The new parameter sets will be $\boldsymbol{\pi}^* = (\pi_1^{(0)}, \ldots, \pi_{L-1}^{(0)})$ and $\boldsymbol{\lambda}^* = (\lambda_1^{(0)}, \ldots, \lambda_{L-1}^{(0)})$. The proposal is accepted with probability $\min\{1, a(L - 1, L)\}$, where

$$a(L + 1, L) = \frac{\Pr(I|\boldsymbol{\pi}^*, \boldsymbol{\lambda}^*, L - 1)}{\Pr(I|\boldsymbol{\pi}^{(0)}, \boldsymbol{\lambda}^{(0)}, L)} \frac{L}{1} \frac{p_a}{p_d}. \quad (20)$$

Note that we handle SDL within the same marker interval in exactly the same way as SDL in different intervals and that (20) is just the inverse of (19). Our interpretation of the terms $(L + 1)^{-1}$ and L in (19) and (20), respectively, differs from the usual. Usually, these terms are included to account for a perceived imbalance in the number of loci selected for a delete step *vs.* that selected for an addition step if the order of loci is not fixed. We believe that the balance is one to one in both the addition and deletion steps and no balancing is necessary; we include these terms because of the Poisson prior. The difference to the usual algorithm, however, is just a minor modification of the prior distribution and thus irrelevant in most biological applications.

APPLICATIONS

To illustrate the method, a simulation study and an analysis of a data set from one cross of two Scots pine (*Pinus sylvatica*) trees are presented. The simulation study conforms to an inbred line BC situation. In the pine data analysis, we concentrate on the maternal part of the progeny of a single tree, *i.e.*, a pseudobackcross design. In a backcross it is not possible to distinguish between gametic selection and viability selection after fertilization.

Simulations: In the simulation study, first, a single viability locus that eliminates 50% of the progeny of the heterozygous genotype, *i.e.*, $\pi = \frac{2}{3}$, was placed in the middle of a chromosome of length 1 M; six markers were spaced at regular intervals of 0.2 M along the chromosome; no missing data were considered. In the second simulation, two SDL with the same effects as in

the single-SDL situation were placed at locations 0.33 M and 0.67 M, respectively. In both cases, simulations with sample sizes of 500 were repeated five times and results were compared; additionally, simulations with sample sizes of 100 and less were also performed. Compared to empirical reports of distortions of marker loci from Mendelian ratios, the simulated effect is high but not unrealistic. The marker map is rather dense and fully informative.

The outcomes of the analyses of the five simulated data sets were almost identical such that we present only one of them. In the maximum-likelihood (ML) analysis, the number of SDL was fixed to one. The inferred effect, the likelihood-ratio statistic Λ, is reported at each location. We also performed an MCMC analysis of the same data. From Figure 1A, we see that the position and effect of the SDL are estimated quite accurately. For the other four simulations, the inferred positions were also mostly between the two middle markers and the estimated effects were close to the true value. Reducing sample sizes did not appreciably change the estimate of location or effect. The likelihood-ratio statistic, however, dropped considerably (results not shown). We do not present the ML results with two SDL, because the model is not appropriate.

With the Bayesian MCMC analysis, the Poisson prior mean was set to $\mu = 1$ and the maximum number of SDL was set to three. The chain length was 10^5. The chain was thinned by storing only after every 10th cycle. No burn-in period was discarded because the chain reached approximate stationarity very quickly. The posterior probability of the simulated number of SDL (*i.e.*, one or two, respectively) was always between 0.6 and 0.9. In the one-SDL case, frequencies are higher at the center, *i.e.*, close to the simulated position (Figure 1B). Effects are very similar to those estimated with the ML method. In the two-SDL case, posterior distributions of both the locations as well as the effects are about correct (Figure 1C). It can be easily discerned from the posterior distribution of frequencies that there are actually two SDL present. When the number of individuals was reduced, the posterior probability of the different numbers of SDL approached that of the prior distribution rapidly (data not shown). This corresponds to the decrease in the likelihood-ratio statistic with decreasing sample size.

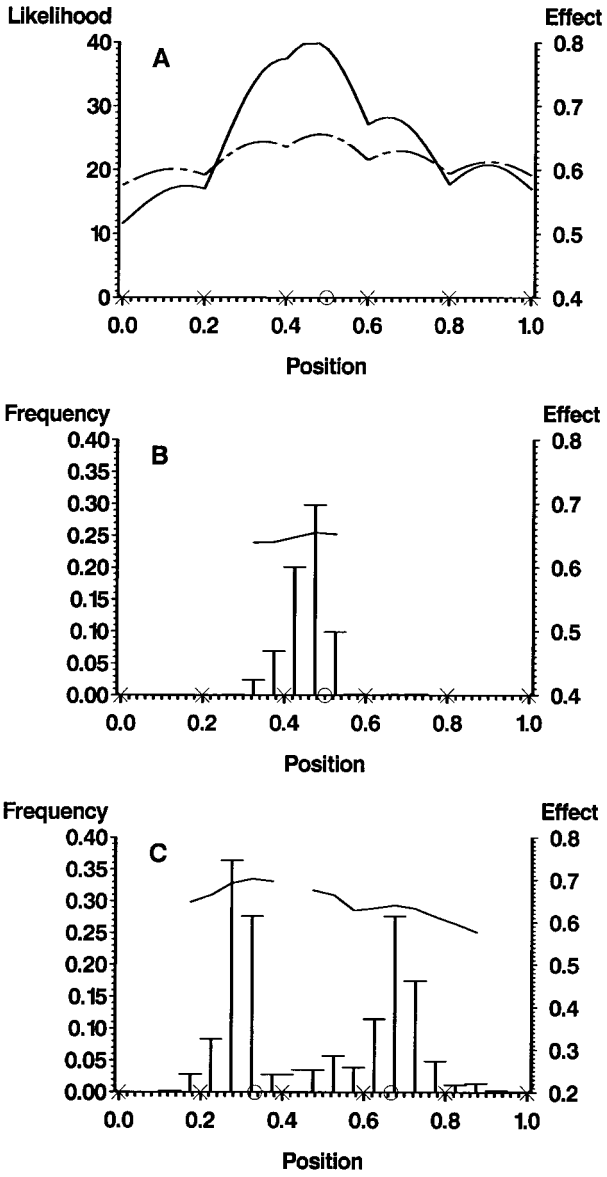

Figure 1.—Simulated data. (A and B) A simulation with one SDL; (C) a simulation with two SDL. The scale on the *x*-axis is 1 M, the positions of the markers are indicated with an "×," while the positions of the SDL are indicated with a circle. "Likelihood" refers to the broken line and to twice the log-likelihood ratio; "frequency" to the posterior probability of an SDL in an interval of 0.04 the length of the linkage group; and "effect" to the solid line and to the probability of finding the homozygote genotype in the BC.

Pine data: In the second application, data consisted of the megagametophytes of open-pollinated offspring of a single Scots pine *P. sylvestris* tree, P304 (Hurme and Savolainen 1999). Megagametophytes are haploid tissues consisting of the maternal part of the seedling's genome and can be scored at the seedling stage without damaging the seedling. We treated the progeny of this tree as a pseudobackcross family. Map distances and linkage phases were determined with Mapmaker as described in Hurme and Savolainen (1999). Five RAPD markers from linkage group 2 were used in this family: C02-680, G13-750, K09-750, E09-250, and AC15-270 at positions 0.038 M, 0.115 M, 0.287 M, 0.461 M, and 0.478 M, respectively. As determined from other crosses, the map length of the whole linkage group was ~0.85 M. The sample size was 73 individuals, and in many individuals some markers were scored as missing.

With the ML analysis, the log-likelihood ratio statistic was appreciable only close to the marker G13-750 (Figure 2A). At this location the inferred effect was an excess

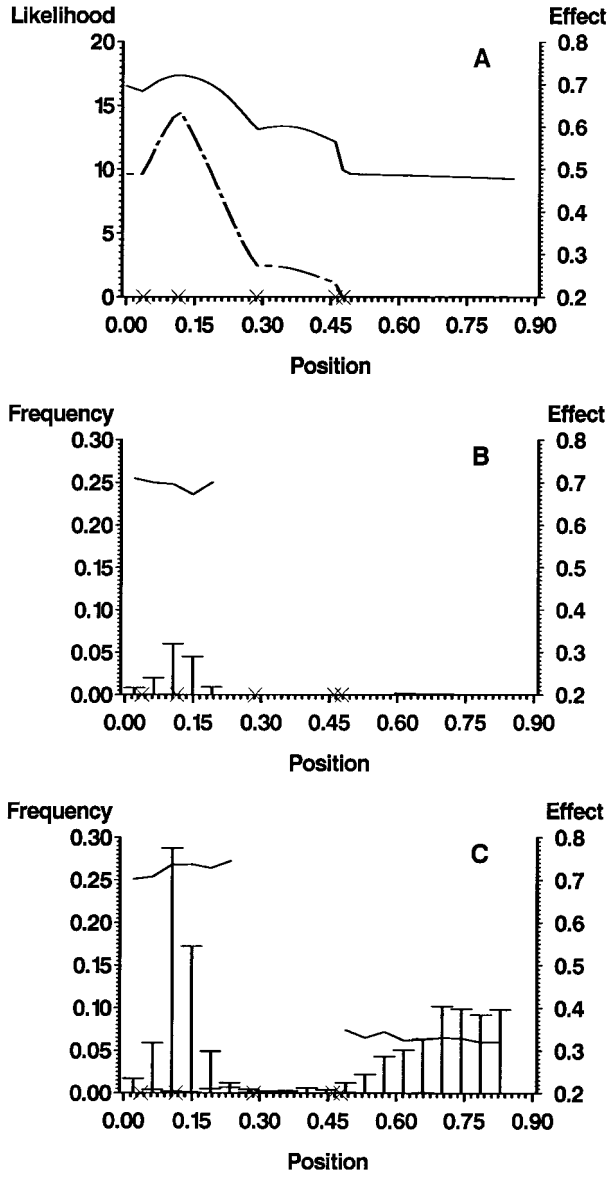

Figure 2.—Pine data. The notation is the same as in Figure 1. The ML result is presented in A, and the posterior distribution of the single-SDL case is in B and of the two-SDL cases in C. The marker loci are (from left to right) C02-680, G13-750, K09-750, E09-250, and AC15-270.

of the heterozygous genotype of ~0.2 over the Mendelian value of 0.5. For the Bayesian MCMC analysis, the prior distribution was the same as for the simulation study. The posterior probabilities of zero, one, two, and three SDL were 0.01, 0.15, 0.61, and 0.23, respectively. This result is, however, quite sensitive to the prior distribution of SDL number. We report the posterior distributions of both one and two inferred SDL. If a single SDL was inferred, it was most often placed close to marker C02-680 (the beginning of the marker region), and the inferred effect was a considerable increase in the second genotype, as in the ML analysis (Figure 2B). If two SDL were inferred, most often location and effect of one of the SDL was similar to the single-SDL case, while the other counteracted its effect at the other end of the linkage group (Figure 2C).

DISCUSSION

Herein, a method for mapping SDL in a backcross is presented. The method makes efficient use of a map of partially or fully informative marker loci by using the multipoint method (Lander and Green 1987; Jiang and Zeng 1997). A maximum-likelihood analysis via an EM algorithm as well as a Markov chain Monte Carlo Bayesian analysis using a reversible jump algorithm for varying the number of loci is presented in detail. Given a dense marker map, the method can be used for precision analysis of positions and effects of the SDL. The best previously available methods (Fu and Ritland 1994b; Mitchell-Olds 1995; Cheng *et al.* 1996) rely on fully informative markers flanking the putative SDL and assume just one SDL per chromosome.

With our approach, it is possible to efficiently analyze the number, positions, and effects of SDL in organisms, for which a high-resolution marker map has been developed and where inbred line crosses can be performed easily. Analysis can be extended easily to a general full-sib family or to the selfing of an outcrossing individual: the dimension changes from two to four, binomial distributions change to multinomial distributions, and the transition probabilities between adjacent loci change. Marker information now contributes to the full or partial identification of four combinations of genotypic configurations. As with the BC case, partial marker information can be defined as the union of compatible cases. All the above changes are rather trivial consequences of the change in dimension but complicate presentation substantially. Additionally, the missing phase information needs to be considered. Furthermore, the multipointing algorithm becomes more important for the full-sib design.

Presently, our method for the backcross can only be used to analyze the SDL *currently* segregating in the two lines, not those that have been segregating in the *ancestral* population from which the inbred lines derived. Segregation distortion might have already affected the inbreeding process for creation of the lines. Extrapolation from the current to the ancestral situation is therefore problematic. This problem is even more pressing for recombinant inbred lines, where overrepresention of chromosomal fragments of one or the other parent is commonly observed (*e.g.*, Lister and Dean 1993) and requires a more elaborate approach.

A distinction needs to be made between segregation distortion before and after fertilization. An SDL acting before fertilization can only alter gametic proportions. Thus genotypic proportions will only be altered indirectly through the combination of gametic proportions,

which restricts the achievable combinations of genotypic proportions. On the other hand, SDL acting after fertilization may alter genotypic proportions directly. Thus, many more combinations of genotypic proportions are possible for SDL acting after fertilization. In experimental crosses more complex than the backcross design, inferred genotypic proportions of an SDL may thus render unlikely prefertilization mechanisms of segregation distortion. Two or more SDL acting before fertilization may, however, mimic the effect of SDL acting after fertilization because of the increase in combinatorial possibilities.

In hybrids of species or subspecies, segregation distortion commonly occurs (see, *e.g.*, Whitkus 1998 and references therein). This may be caused by structural rearrangements, *e.g.*, inversions, which constitute a prefertilization mechanism. Alternatively, the segregation distortion may be caused by postfertilization differences in viability between genotypic configurations, most probably caused by epistatic interactions. Our method can be used to detect chromosomal areas that are causing these distortions. But because of the presumed epistasis, relaxation of the assumption of a multiplicative effect of different SDL may be necessary.

Our method may also be used to map loci influencing early viability. This would enhance our understanding of the nature of early inbreeding depression. The method provides another approach for estimating the number and effects of loci causing inbreeding depression. Traditionally, such information has been derived mainly from biometric analysis of crosses (*e.g.*, Dudash and Carr 1998). But as inbreeding depression can be expressed in embryonic life stages not amenable to biometric analysis, application of this method is limited. To gain insight on these early life stages, sparse maps and single-marker methods have been used to infer the effect of a viability locus influencing inbreeding depression (Sorensen 1967; Servitova and Cetl 1984; Hedrick and Muona 1990; Fu and Ritland 1994a; Kärkkäinen *et al.* 1999). With single-marker analysis, estimation of position and effect of the SDL is, however, confounded and multiple SDL on a single linkage group cannot be handled at all. Interval methods (Fu and Ritland 1994b; Mitchell-Olds 1995; Cheng *et al.* 1996) rely on fully informative markers flanking the putative SDL and assume just one SDL per chromosome. Dense linkage maps of fully informative markers may be hard to obtain in closely related individuals that need to be considered in the analysis of inbreeding depression. Like the interval methods, our method requires a dense linkage map of polymorphic markers but is not restricted to fully informative markers; instead it can make efficient use of, *e.g.*, dominant markers.

Only rarely have data sets been gathered for mapping segregation distortion or viability selection (see, however, Harushima *et al.* 1996 and Kuang *et al.* 1998). But often in QTL experiments, wide crosses are used to increase differences between parents and thus the power of mapping. Probably for this reason, markers with segregation ratio distortions are commonly observed in data sets used for QTL mapping resulting from wide crosses (*e.g.*, van Ooijen *et al.* 1994). Segregation ratio distortion is also commonly observed in doubled haploid lines (*e.g.*, Fulton *et al.* 1997).

Usually generation of a linkage map of marker loci precedes QTL analysis. If a dense map of informative markers is inferred correctly, the bias introduced by segregation distortion into QTL analysis will be negligible. But if recombination fractions or, worse, order of marker loci are inferred incorrectly, basic assumptions of QTL analysis do not hold and results will be imprecise at best. Hence, aside from being interesting in themselves, SDL cause practical problems in QTL projects as observed, *e.g.*, by Sandbrink *et al.* (1995). Thus, segregation distortion should be accounted for in mapping projects.

Segregation distortion is known to bias estimation of recombination fractions in two-point inference of recombination distances between markers (Lorieux *et al.* 1995a,b; Liu 1998). If markers are fully informative, estimation of the recombination fraction of only the markers flanking the SDL will be affected. Only in the unlikely case of coincidence of SDL and marker location will no bias be observed. If less than fully informative markers are used, the effects of the distortion are spread out to the smallest interval of fully informative markers flanking the distorted region. As a remedy, markers that show obvious segregation distortion are often excluded from the map. But that reduces coverage of the genome and qualitative or quantitative trait loci might be missed.

Our method can be extended to allow for detection of SDL concurrently with estimation of a linkage. Cheng *et al.* (1996) have already developed an EM algorithm to infer positions of two fully informative markers in the presence of a single SDL (an interval method) in a backcross or doubled haploid lines. This could be extended to a multipoint inference of a marker map in the presence of SDL by augmenting the EM or MCMC schemes presented herein by allowing the markers to change their positions relative to each other.

The source code for a C++ program and executables for a Sun workstation, with which the above calculations can be performed, are available from Claus Vogl (claus@genetics.ucr.edu).

We thank Päivi Hurme and Outi Savolainen for the data set and Elja Arjas, Anita de Haan, Mikko Sillanpää, and Nengjun Yi for discussion of this and related issues. Outi Savolainen, Elja Arjas, and Lori Weingartner have commented on earlier versions of this manuscript. We thank Zhao-Bang Zeng and two anonymous reviewers for their patient work, which helped to improve this article a lot. This work was supported by grants from the Environment and Natural Resources Research Council and the Medical Research Council to Outi Savolainen and by the National Institutes of Health Grant GM-55321 and the U.S. Department of Agriculture National Research Initiative Competitive Grants Program 97-35205-5075 to S.X.

LITERATURE CITED

Charlesworth, B., and D. Charlesworth, 1987 Inbreeding depression and its evolutionary consequences. Annu. Rev. Ecol. Syst. **18:** 237–268.

Charlesworth, B., and D. Charlesworth, 1998 Some evolutionary consequences of deleterious mutations. Genetica **102/103:** 3–19.

Cheng, R., A. Saito and Y. Ukai, 1996 Estimation of the position and effect of a lethal factor locus on a molecular marker linkage map. Theor. Appl. Genet. **93:** 494–502.

Dudash, M. W., and D. E. Carr, 1998 Genetics underlying inbreeding depression in *Mimulus* with contrasting mating systems. Nature **393:** 682–684.

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

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

Fulton, T.-M., J. C. Nelson and S. D. Tanksley, 1997 Introgression and DNA marker analysis of *Lycopersicum peruvianum,* a wild relative of the cultivated tomato, into Lycopersicum esculentum, followed through three successive backcross generations. Theor. Appl. Genet. **95:** 895–902.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 1995 *Bayesian Data Analysis.* Chapman and Hall, London.

Grant, V., 1975 *Genetics of Flowering Plants.* Columbia University Press, New York.

Green, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711–732.

Hartl, D. L., and A. G. Clark, 1997 *Principles of Population Genetics,* Ed. 3. Sinauer, Sunderland, MA.

Harushima, Y., N. Kurata, M. Yano, Y. Nagamura, T. Sasaki *et al.*, 1996 Detection of segregation distortions in an indica-japonica rice cross using a high-resolution molecular map. Theor. Appl. Genet. **92:** 145–150.

Heath, S. C., 1997 Markov-chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. **61:** 748–760.

Hedrick, P. W., 1994 Purging inbreeding depression and the probability of extinction: full-sib families. Heredity **73:** 363–372.

Hedrick, P. W., and O. Muona, 1990 Linkage of viability genes to marker loci in selfing organisms. Heredity **64:** 67–72.

Hurme, P., and O. Savolainen, 1999 Comparison of homology and linkage of RAPD markers between individual trees of Scots pine (*Pinus sylvestris* L.). Mol. Ecol. **8:** 15–22.

Husband, B. C., and D. W. Schemske, 1996 Evolution of magnitude and timing of inbreeding depression in plants. Evolution **50:** 554–570.

Jansen, R. C., and P. Stam, 1994 High resolution of quantitative traits into multiple loci via interval mapping. Genetics **136:** 1447–1455.

Jiang, J., and Z.-B. Zeng, 1997 Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica **101:** 47–58.

Kärkkäinen, K., V. Koski and O. Savolainen, 1996 Geographical variation in inbreeding depression in Scots pine. Evolution **50:** 111–119.

Kärkkäinen, K., H. Kuittinen, R. van Treuren, C. Vogl and O. Savolainen, 1999 Genetic basis of inbreeding depression in *Arabis petrea.* Evolution **53:** 1354–1365.

Kuang, H., T. E. Richardson, S. D. Carson and B. C. Bongarten, 1998 An allele responsible for seedling death in *Pinus radiata* D. Don. Theor. Appl. Genet. **96:** 640–644.

Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

Lander, E. S., and P. Green, 1987 Construction of multilocus genetic maps in humans. Proc. Natl. Acad. Sci. USA **84:** 2363–2367.

Launey, S., and D. Hedgecock, 1999 Genetic load causes segregation ratio distortion in oysters: mapping at 6 hours. Plant and Animal Genome VII, abstracts W14, p. 33.

Lister, C., and C. Dean, 1993 Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana.* Plant J. **4:** 745–750.

Liu, B. H., 1998 *Statistical Genomics: Linkage, Mapping, and QTL Analysis.* CRC Press, Boca Raton, FL.

Lorieux, M., B. Goffinet, X. Perrier, D. González de León and C. Lanaud, 1995a Maximum likelihood models for mapping genetic markers showing segregegation distortion. 1. Backcross populations. Theor. Appl. Genet. **90:** 73–80.

Lorieux, M., X. Perrier, B. Goffinet, C. Lanaud and D. González de León, 1995b Maximum likelihood models for mapping genetic markers showing segregegation distortion. 2. F2-populations. Theor. Appl. Genet. **90:** 81–89.

McColdrick, D. J., and D. Hedgecock, 1997 Fixation, segregation and linkage of allozyme loci in inbred families of the Pacific oyster *Crassostrea giga* (Thunberg): implications for the causes of inbreeding depression. Genetics **146:** 321–334.

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

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

Sandbrink, J. M., J. W. van Oijen, C. C. Purimahua, M. Vrielink, R. Verkerk *et al.*, 1995 Localization of genes for bacterial resistance in *Lycopersicon peruvianum* using RFLPs. Theor. Appl. Genet. **90:** 444–450.

Satagopan, R. J., and B. S. Yandell, 1996 Estimating the number of quantitative trait loci via Bayesian model determination. Special Contributed Paper Session on Genetic Analysis of Quantitative Traits and Complex Diseases. Biometric Section, Statistical Meeting. Chicago, IL.

Servitová, J., and I. Cetl, 1984 The use of recessive lethal chlorophyll mutants for linkage mapping of *Arabidopsis thaliana* (L.) Heynh. Arabidopsis Inf. Serv. **21:** 59–64.

Sillanpää, M., and E. Arjas, 1998 Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics **148:** 1373–1388.

Sorensen, F. C., 1967 Linkage between marker genes and embryonic lethal factors may cause distrubed segregation rations. Silvae Genet. **16:** 132–134.

Stephens, D. A., and R. D. Fisch, 1998 Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. Biometrics **54:** 1334–1347.

van Ooijen, J. W., J. M. Sandbrink, M. Vrielink, R. Verkerk, P. Zabel *et al.*, 1994 An RFLP linkage map of *Lycopersicum peruvianum.* Theor. Appl. Genet. **89:** 1007–1013.

Whitkus, R., 1998 Genetics of adaptive radiation in Hawaiian and Cook Island species of Tetramolopium (Asteraceae). II. Genetic linkage map and its implications for interspecific breeding barriers. Genetics **150:** 1209–1216.

Williams, C. G., and O. Savolainen, 1996 Inbreeding depression in conifers implications for breeding strategy. For. Sci. **42:** 102–117.

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

Communicating editor: Z-B. Zeng