

Methodology article

Open Access

## An EM algorithm for mapping segregation distortion loci

Chengsong Zhu and Yuan-Ming Zhang\*

Address: Section on Statistical Genomics, State Key Laboratory of Crop Genetics and Germplasm Enhancement/National Center for Soybean Improvement, Nanjing Agricultural University, Nanjing 210095, China

Email: Chengsong Zhu - [cszhu@ksu.edu](mailto:cszhu@ksu.edu); Yuan-Ming Zhang\* - [soy Zhang@njau.edu.cn](mailto:soy Zhang@njau.edu.cn)

\* Corresponding author

Published: 29 November 2007

Received: 12 May 2007

*BMC Genetics* 2007, **8**:82 doi:10.1186/1471-2156-8-82

Accepted: 29 November 2007

This article is available from: <http://www.biomedcentral.com/1471-2156/8/82>

© 2007 Zhu and Zhang; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** Chromosomal region that causes distorted segregation ratios is referred to as segregation distortion locus (SDL). The distortion is 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, observable phenotypes are distorted for marker loci in the chromosomal region close to the SDL. Under the quantitative genetics model for viability selection by proposing a continuous liability controlling the viability of individual, a simplex algorithm has been used to search for the solution in SDL mapping. However, they did not consider the effects of SDL on the construction of linkage maps.

**Results:** We proposed a multipoint maximum-likelihood method to estimate the position and the effects of SDL under the liability model together with both selection coefficients of marker genotypes and recombination fractions. The method was implemented via an expectation and maximization (EM) algorithm. The superiority of the method proposed under the liability model over the previous methods was verified by a series of Monte Carlo simulation experiments, together with a working example derived from the MAPMAKER/QTL software.

**Conclusion:** Our results suggested that the new method can serve as a powerful alternative to existing methods for SDL mapping. Under the liability model, the new method can simultaneously estimate the position and the effects of SDL as well as the recombinant fractions between adjacent markers, and also be used to probe into the genetic mechanism for the bias of uncorrected map distance and to elucidate the relationship between the viability selection and genetic linkage.

### Background

In a segregation population derived from a cross between two inbred lines, some molecular markers often show distorted segregation ratios from Mendelian expectations [1-3]. The distortion is frequently related to gamete gene, sterile gene and chromosome translocation [4]. So the detection of the gene or locus, known as segregation distortion locus (SDL) mapping, is warranted. However, the challenge encountered in SDL mapping is mainly caused

by the unavailability of phenotypic data for the underlying trait. In fact, molecular markers linked to the SDL frequently show segregation distortion and the degree of distortion depends on the size and the position of SDL. Therefore, it is possible to detect SDL by means of the distortion.

Mapping SDL is usually studied at the population level by examining the change of gene (or genotypic) frequencies

of markers [5]. In the past a single marker was often used to detect the linkage between the marker and SDL [6,7]. Its shortcomings are very similar to those of single-marker approaches in quantitative trait loci (QTL) mapping [8]. Since the introduction of interval mapping of QTL [9], Hedrick and Muona [10] developed a flanking-marker analysis to estimate the fitness parameters for a viability locus. The model of Hedrick and Muona [10] is actually a complete recessive model. Mitchell-Olds [11] detected one putative viability locus at a time and then scanned the entire genome for every putative position to provide a test statistic profile for the detection of SDL. However, his model only test and estimate the degree of dominance. Luo and Xu [12] extended the maximum-likelihood (ML) method to estimate degree of dominance and selection coefficients using an outbred full-sib family as an example. Wang et al. [13] developed a multipoint ML method to estimate the position and the genotypic frequencies of SDL in an  $F_2$  population. However, the efficacies of the methods mentioned above have been seldom addressed in simulation studies. Recently, Luo et al. [14] developed a quantitative genetics model for viability selection. This approach makes it possible to carry out simulation studies, to partition the selection into additive and dominant effects and to remove the effects of non-genetic cofactors from the analysis [14,15]. However, this approach raises two issues. Firstly, they assumed that segregation distortion didn't affect the construction of genetic linkage map. In fact, marker segregation distortion is known to affect the estimates for both recombination fractions in pairwise analysis of markers and the order of the markers on a linkage group [16-18]. As for the genetic parameters, then, Luo et al [14] adopted the Simplex algorithm [19] to search for the solutions at the cost of computational consuming. Under the liability model proposed by Luo et al [14], therefore, in this paper it is necessary to extend the multipoint approach by combining the estimations of the genetic parameters of SDL with the reconstruction of genetic linkage maps. The new method for SDL mapping was implemented via an expectation and maximization (EM) algorithm rather than Simplex procedure. The genetic factors that might affect the estimates of recombination fractions between adjacent markers would be discussed in detail. A series of Monte Carlo simulation experiments together with a working example from the Mapmaker/QTL software were carried out to verify our approach.

**Methods**

**Genetic model**

Considering an SDL in an  $F_2$  population derived from a cross between two inbred lines, we assumed three genotypes at this locus, *AA*, *Aa* and *aa*, to have genotypic values  $\sqrt{2} a - d$ ,  $d$  and  $-\sqrt{2} a - d$ , respectively, with  $a$  and  $d$  indi-

cating additive and dominant effects, and an imaginary trait, liability, invisible to the investigators but visible to nature, controlled the viabilities of individuals. It should be noted that the genetic variance in an  $F_2$  population was  $a^2 + d^2$  rather than  $\frac{1}{2} a^2 + \frac{1}{4} d^2$  as usual. The phenotypic value of the  $j$ th individual was described by the following linear model,

$$z_j = g_j + e_j \tag{1}$$

where  $g_j$  was the genotypic value for the  $j$ th individual, and  $e_j$  a normally distributed residual variable with mean zero and standard deviation 1.0, which accounted for polygenes that were linked to the markers and for environmental variation [14,18]. Provided that the liability was subject to natural selection, an individual would survive if  $z_j \geq 0$  and would be eliminated from the population if  $z_j < 0$ . Since all the sampled individuals had survived from the viability selection, the liability of each genotype followed a truncated distribution with a cumulative probability,  $G_j = h$  ( $h = 1, 2, 3$ ), with

$$f_h = \Pr(z_j \geq 0 | G_j = h) = \Phi[(2 - h)\sqrt{2}a + (-1)^h d] \tag{2}$$

where  $h$  indexed the genotypes of the SDL, and  $f_h$  was referred to as the relative fitness of the  $h$ th genotype [14]. The expected frequencies of the three genotypes were

$$p_{AA} = \frac{0.25f_1}{0.25f_1 + 0.5f_2 + 0.25f_3} = \frac{f_1}{f_1 + 2f_2 + f_3}$$

similarly

$$p_{Aa} = \frac{2f_2}{f_1 + 2f_2 + f_3} \quad p_{aa} = \frac{f_3}{f_1 + 2f_2 + f_3} \tag{3}$$

**Mapping SDL under a liability model**

We assumed that there was no crossing-over interference among the markers on the linkage group considered, an SDL caused segregation distortion of some or all markers linked to the SDL, and three genotypes for each marker had different viability coefficients. Let the order of the  $m$  markers on a same chromosome be  $M_1, M_2, \dots, M_m$ ;  $x_k$  be a dummy variable defined as  $x_k = 1, 0, -1$  for a homozygote of  $P_1$ , a heterozygote and a homozygote of  $P_2$  at the  $k$ th marker, respectively;  $z_k$  be indicator for phenotype of the  $k$ th marker ( $M_k$ );  $r_k$  (or  $r_{k,k+1}$ ) be the recombination fraction between the  $k$ th and  $(k+1)$ th markers; and  $s_{k,1}$  and  $s_{k,2}$  ( $0 \leq s_{k,1} < +\infty$  and  $0 \leq s_{k,2} < +\infty$  for  $k = 1, 2, \dots, m$ ) be the viability coefficients of  $M_k m_k$  and  $m_k m_k$  relative to  $M_k M_k$  at the  $k$ th marker.

Now let an SDL locate between the  $k$ th and  $(k+1)$ th markers, and  $\phi_{jh}$  be the indicator function, taking the value of 1, if the  $j$ th individual belonged to the  $h$ th possible genotype in the  $F_2$  population, otherwise taking the value of zero. The parameters were  $\Omega = (p_{AA}, p_{Aa}, p_{aa}, \delta)$  or  $\Omega = (a, d, \delta)$ , with  $\delta$  indicating the SDL location. The distribution of  $\phi_{jh}$  was described as

$$\Pr(\phi_{jh} | \Omega) = (p_{AA})^{\phi_{j1}} (p_{Aa})^{\phi_{j2}} (p_{aa})^{\phi_{j3}} = \frac{f_1^{\phi_{j1}} (2f_2)^{\phi_{j2}} f_3^{\phi_{j3}}}{f_1 + 2f_2 + f_3} \quad (j = 1, \dots, n) \quad (4)$$

where  $n$  was sample size. The likelihood defined with matrix notation was

$$L(\Omega) = \prod_{j=1}^n \left\{ H_j'(r_{k,k'}) \prod_{o=k-1}^1 H_j'(r_{o,o+1}) q_1 Y [H_j(r_{k',k+1}) \prod_{o=k+1}^{m-1} H_j(r_{o,o+1}) c] \frac{f_1^{\phi_{j1}} (2f_2)^{\phi_{j2}} f_3^{\phi_{j3}}}{f_1 + 2f_2 + f_3} \right\} \quad (5)$$

where  $q_1' = [\Pr(x_1 = 1), \Pr(x_1 = 0), \Pr(x_1 = -1)]$ ,  $c' = [1, 1, 1]$ , ' denoted transpose of a matrix or vector, and the transition probability matrix  $H_j(r_{k,k+1})$  from marker  $M_k$  to  $M_{k+1}$  for the  $j$ th individual was

$$\begin{bmatrix} \frac{(1-r_k)^2}{(1-r_k)^2 + 2s_{k+1,1}r_k(1-r_k) + s_{k+1,2}r_k^2} & \frac{2s_{k+1,1}r_k(1-r_k)}{(1-r_k)^2 + 2s_{k+1,1}r_k(1-r_k) + s_{k+1,2}r_k^2} & \frac{s_{k+1,2}r_k^2}{(1-r_k)^2 + 2s_{k+1,1}r_k(1-r_k) + s_{k+1,2}r_k^2} \\ \frac{r_k(1-r_k)}{(1+s_{k+1,2})r_k(1-r_k) + s_{k+1,1}(1-2r_k+2r_k^2)} & \frac{s_{k+1,1}(1-2r_k+2r_k^2)}{(1+s_{k+1,2})r_k(1-r_k) + s_{k+1,1}(1-2r_k+2r_k^2)} & \frac{s_{k+1,2}r_k(1-r_k)}{(1+s_{k+1,2})r_k(1-r_k) + s_{k+1,1}(1-2r_k+2r_k^2)} \\ \frac{r_k^2}{r_k^2 + 2s_{k+1,1}r_k(1-r_k) + s_{k+1,2}(1-r_k)^2} & \frac{2s_{k+1,1}r_k(1-r_k)}{r_k^2 + 2s_{k+1,1}r_k(1-r_k) + s_{k+1,2}(1-r_k)^2} & \frac{s_{k+1,2}(1-r_k)^2}{r_k^2 + 2s_{k+1,1}r_k(1-r_k) + s_{k+1,2}(1-r_k)^2} \end{bmatrix}$$

There were several ways to find the ML estimates (MLEs) of model parameters. We here adopted an EM algorithm [20] and treated  $\phi_{jh}$  as missing data. We regarded  $\delta$  as constant for the moment, now the parameter set was  $\theta = (a, d)'$ . For the EM algorithm, we needed to obtain the expectation of the complete data log-likelihood function,

$$L = C + \sum_{j=1}^n [p(\phi_{j1} = 1) \ln f_1 + p(\phi_{j2} = 1) \ln(2f_2) + p(\phi_{j3} = 1) \ln f_3 - \ln(f_1 + 2f_2 + f_3)] \quad (6)$$

where the constant  $C$  didn't depend on the parameters of interest, and but did depend on the viability coefficients and map distance between adjacent markers, which could be determined by Zhu et al [18]. The EM algorithm was described as follows.

**E-step**

Provided the initial values for the model parameters, i.e.,  $a^{(0)} = 0.0$  and  $d^{(0)} = 0.0$ . The posterior probabilities of  $\phi_{jh} = 1$  were

$$p(\phi_{jh} = 1) = \frac{\Pr(\phi_{jh} = 1 | z_{j1}, \dots, z_{jM}) p_{jh}^{(0)}}{\sum_{o=1}^3 \Pr(\phi_{jo} = 1 | z_{j1}, \dots, z_{jM}) p_{jo}^{(0)}} \quad (7)$$

where  $p_{jh}^{(0)}$  ( $h = 1, 2, 3$ ) was calculated from equation (3), and  $\Pr(\phi_{jh} = 1 | z_{j1}, \dots, z_{jM})$  ( $j = 1, \dots, n; h = 1, 2, 3$ ) the prior probability of the  $h$ th genotype of SDL for the  $j$ th individual conditional on marker information  $(z_{j1}, \dots, z_{jM})$  by means of the multipoint method [21].

**M-step**

The MLEs of parameters were obtained by the Fisher-scoring algorithm as it was impossible to get their explicit solutions [22]. The  $\theta$  could be updated by

$$\theta^{(1)} = \theta^{(0)} + I^{-1}S(\theta^{(0)}) \quad (8)$$

where  $S(\theta^{(0)})$  was the score function, and  $I$  was the Fisher information matrix (more details were given in Appendix). And  $\theta^{(1)}$  would replace  $\theta^{(0)}$  in all subsequent estimating steps, and the procedure was iterated until the convergence occurred. The converged  $\theta^{(1)}$  was the MLEs of  $\theta$  in this M-step.

The  $E$  and  $M$  steps were iterated until the convergence occurred.

The MLE for the SDL position could be obtained by examining the likelihood-ratio profile along the chromosome as was commonly done in interval mapping of QTL [9].

Following parameter estimation, we tested an overall null hypothesis that was no effect of SDL at the locus of interest ( $\delta$ ). The null hypothesis was formulated as  $H_0: a = d = 0.0$ , which was tested using the likelihood-ratio ( $LR$ ) test statistic:

$$LR = -2[\ln L(0, 0, \delta) - \ln L(a, d, \delta)]$$

Under the null hypothesis, the statistic  $LR$  approximately followed chi-square distribution with two degrees of freedom.

The critical value for power calculation was determined by computing 1,000 permutations [23], the experiment-wise type I error was set at 5%, and the confidence interval of an SDL location was determined by the bootstrapping method [24].

**Simulation model**

We simulated one chromosome of 100 cM (or 50 cM) long covered by  $m$  evenly spaced codominant markers ( $m = 6, 11$  or  $21$ ) and put a single SDL at position 25 cM (another SDL was put at position 75 cM if necessary). The dominance ratio of the SDL was denoted by  $dr = d/a$ . Given the broad heritability ( $h^2$ ) and  $dr$ , the additive and dominant effects could be obtained using numerical algorithm [25]. Based on the method described in Luo *et al.*[14], all genotypes of both distorted markers and SDL for each individual in an  $F_2$  population were simulated. All simulations were replicated 100 or 1000 times depending on the purpose of the analyses. Empirical power was calculated by counting the number of runs in which test statistics were greater than the critical values [26].

**Results**

**Effects of various factors on SDL mapping**

In this simulated experiments, the effects of sample size, SDL heritability and marker interval length on SDL mapping were studied, respectively. The performance of the proposed method was evaluated by statistical power, average and standard deviation of estimates with 100 replicates. All parameters and results were listed in Table 1. The results showed the general behavior of QTL mapping, i.e.,

the estimate for each parameter was very close to its corresponding true value, the power and the precision for SDL mapping increased with the increase in sample size and SDL heritability, respectively. However, marker interval length had slight effect on the power under the three levels studied.

**Mapping multiple SDL**

Similar to the interval mapping procedure of Lander and Botstein [9], the single-locus model for SDL mapping was used to search for multiple loci. Eleven markers were evenly placed on a simulated chromosome of length 100 cM. Two SDL each with a 0.5 dominance-ratio and a 0.15 heritability were respectively located at positions 25 cM and 75 cM on the simulated chromosome. One hundred independent simulation runs were performed for a sample size of 200. The results were listed in Table 2. Both loci were identified at almost 100% power. The results from simulation experiments demonstrated that the new method based on single-SDL model may be considered as an approximate approach to search for multiple loci if the SDL are sufficiently separated by markers.

**A working example**

As a demonstration of the proposed method in this paper, we re-analyzed a sample dataset (the source filename:

**Table 1: Results of segregation distortion locus (SDL) mapping under the fitness and liability models (100 replications)**

Sample size	Interval length (cM)	Broad heritability	Dominance ratio	Power (%)	Positions (cM)	SDL effects		Frequencies of genotypes		
						Additive	Dominant	$\hat{p}_{AA}$	$\hat{p}_{Aa}$	$\hat{p}_{aa}$
100	10	0.05	0.50	24	True value	0.2124	0.1062	0.2880	0.5412	0.1708
					25.00	0.2011	0.1243	0.2809	0.5485	0.1706
					(30.55)	(0.1094)	(0.1101)	(0.0716)	(0.0807)	(0.0551)
200	10	0.05	0.50	55	True value	0.2143	0.1134	0.2871	0.5440	0.1689
					30.06	(0.0988)	(0.0994)	(0.0499)	(0.0573)	(0.0386)
					(19.33)	(0.2097)	(0.1068)	(0.2851)	(0.5454)	(0.1695)
300	10	0.05	0.50	63	True value	(0.0791)	(0.0804)	(0.0345)	(0.0372)	(0.0262)
					(15.68)	0.1679	0.1679	0.2634	0.5656	0.1710
					29.2	0.1664	0.1701	0.2672	0.5649	0.1679
200	10	0.05	1.00	53	True value	(0.0991)	(0.1007)	(0.0434)	(0.0557)	(0.0386)
					(21.58)	0.2476	0.2476	0.2689	0.5944	0.1367
					25.00	0.2584	0.2681	0.2676	0.6017	0.1307
200	10	0.10	1.00	90	True value	(0.0916)	(0.0922)	(0.0372)	(0.0405)	(0.0251)
					(7.75)	0.3162	0.3162	0.2729	0.6170	0.1101
					25.00	0.3066	0.3111	0.2711	0.6181	0.1108
200	10	0.15	1.00	99	True value	(0.0861)	(0.0873)	(0.0349)	(0.0400)	(0.0274)
					(6.66)	0.3132	0.1566	0.3046	0.5590	0.1364
					25.00	0.3270	0.1634	0.3007	0.5723	0.1270
200	20	0.15	0.50	92	True value	(0.0984)	(0.1023)	(0.0393)	(0.0432)	(0.0260)
					(13.32)	0.3248	0.1642	0.3086	0.5615	0.1300
					26.85	(0.0961)	(0.0987)	(0.0364)	(0.0442)	(0.0285)
200	10	0.15	0.50	91	True value	0.3127	0.1603	0.3037	0.5674	0.1289
					25.81	(0.0923)	(0.0911)	(0.0375)	(0.0450)	(0.0263)
					(10.62)					

The standard deviations are in parentheses. The same is true for the later Tables.

**Table 2: Results of two segregation distortion loci (SDL) mapping under the fitness and liability models (100 replicates and 200 individuals)**

	SDL	Power (%)	Position (cM)	$\hat{a}$	$\hat{d}$	$\hat{p}_{AA}$	$\hat{p}_{Aa}$	$\hat{p}_{aa}$
1	True value	/	25.00	0.3996	0.1998	0.3176	0.5726	0.1098
	Estimate	96	29.13(8.96)	0.4014(0.1097)	0.2018(0.1184)	0.3289(0.0376)	0.5653(0.0425)	0.1058(0.0268)
2	True value	/	75.00	0.3996	0.1998	0.3176	0.5726	0.1098
	Estimate	94	66.75(11.30)	0.4037(0.1064)	0.2053(0.1031)	0.3275(0.0411)	0.5544(0.0450)	0.1181(0.0267)

sample.raw) in the MAPMAKER/QTL software [27]. It consisted of 333 F<sub>2</sub> individuals from a cross between two inbred lines in tomato. Each plant was genotyped for 12 marker loci that were divided into two linkage groups. Single-marker chi-square test showed that 5 and 2 markers on the first and second linkage groups deviated from Mendelian segregation ratios, respectively (data not shown). Given the reconstructed linkage maps using the method of Zhu *et al.* [18], 1000 simulated datasets without segregation distortion were simulated and used to determine the critical value [23]. The confidence interval of a SDL location was determined by the Bootstrap method [24].

The map distances between consecutive markers were calculated twice with and without considering SDL. The former was corrected map distance obtained from the method of Zhu *et al.* (2007) [18]; and the latter was uncorrected one using the Mapmaker/EXE 3.0 software [27]. The results were listed in Table 3. The results showed that the corrected map distances differed from the uncorrected ones when there were distorted markers. The genetic reason of these inconsistencies would be discussed in the following section. Using the proposed method here, a total of four SDL were mapped (Table 4, Fig 1). Two SDL were on the first linkage group and the others on the second one. The genetic parameters for the four SDL were listed in Table 4. The results showed that the distortion was stronger for the first linkage group than for the second one (Fig 1). It resulted in a maximum difference between the corrected and uncorrected map distances for the first marker interval on the first linkage group. Moreover, two linked SDL on the second linkage group also gave rise to two big differences (Table 3). As compared to a single

SDL, therefore, linked SDL had a larger effect on the estimate of map distance.

**Effect of genetic model of SDL on the estimation of map distance**

In this section, our purpose was to make clear the genetic reason for the inconsistencies between corrected and uncorrected map distances when there were distorted markers. Six evenly spaced codominant markers were simulated on a single-chromosome segment of length 50 cM. Two linked SDL with locations at positions 10 and 20 cM (exactly the 2nd and 3rd marker loci) were simulated on the simulated chromosome. One hundred simulation runs were performed for a sample size of 300. Each of datasets was analyzed twice by the method of Zhu *et al.* (2007) [18] and the Mapmaker/EXE 3.0 software [27]. The former was corrected map distance and the latter uncorrected one. For an additive-dominant model, all genetic parameters and the results were listed in Table 5. Results showed that uncorrected map distance was underestimated for most cases, overestimated for opposite dominant effects, and unbiased for all negative additive effects. The results from the real dataset analysis above partly confirmed the result that opposite dominant effects of the two linked SDL on the second linkage group (Table 4) gave rise to the overestimation (Table 3). For an epistatic model, all genetic parameters and the results were listed in Table 6. Results showed that uncorrected map distance was underestimated for most situations, overestimated for negative additive-by-additive or negative dominant-by-dominant effects, and unbiased for additive-by-dominant effect. As we expected, corrected genetic distances were unbiased when considering SDL (Table 5 and 6). Hence, corrected linkage maps were recommended to

**Table 3: The uncorrected and corrected map distances in the real data analysis**

Map distance	Linkage group 1				Linkage group 2					
	1	2	3	4	1	2	3	4	5	6
Corrected	5.29	14.93	11.60	12.77	14.58	6.24	18.88	24.04	17.13	27.69
Uncorrected	4.18	15.01	11.88	12.19	14.76	6.38	18.88	24.04	18.12	28.69

Uncorrected and corrected map distances were calculated by Mapmaker/EXE 3.0 software (Lander et al 1987) and DistortedMap software (Zhu et al 2007), respectively. The same is true for Tables 5 and 6.

**Table 4: Results of segregation distortion loci (SDL) mapping in a real data analysis**

Linkage group	Position (cM)	Confidence interval (95%)	Nearest marker to SDL	$\hat{a}$	$\hat{d}$	Dominance ratio	$\hat{p}_{AA}$	$\hat{p}_{Aa}$	$\hat{p}_{aa}$	Selection types
1	2	1~11	T175	1.4628	-0.1669	-0.11	0.5249	0.4598	0.0153	Zygotic
	41	27~44	T508	0.6792	0.0874	0.13	0.3992	0.8280	0.0728	Zygotic
2	84	57~90	T209	-0.0164	0.3133	-19.10	0.1842	0.6229	0.1929	Zygotic
	109	85~127	T17	0.4969	-0.3184	-0.64	0.4347	0.3853	0.1800	Zygotic

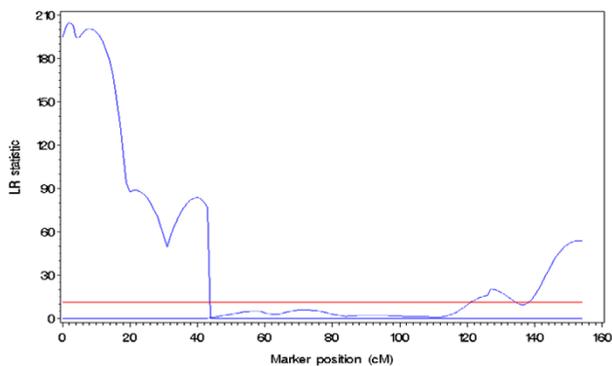
be used for further QTL or SDL analysis unless there was strong evidence to believe that all markers presented typical Mendelian segregation.

**Discussion**

For SDL mapping, most researchers concentrate their attention upon detecting and testing either the selection coefficients or the degree of dominance under the fitness model [7,10,11]. Luo *et al.* [14] pioneered in the development of SDL mapping under a liability model. Zhu *et al.* [18] proposed a new method for the reconstruction of linkage maps with distorted, dominant and missing markers. Under the liability model, we developed a method to simultaneously estimate the position and the effects of SDL as well as the recombination fractions between adjacent markers. This approach remains the merits of Luo *et al.* [14] but differs from others in several aspects. Firstly, it combines the detection of SDL with the reconstruction of marker linkage map. The position and the effect of SDL can be estimated along with the selection coefficient and the degree of dominance. Then, the proposed method may be used to elucidate the relationship between the viability selection and genetic linkage. Thirdly, the likelihood function is involved in the distribution of genotypes of

SDL rather than that of marker genotypes in the previous studies [11,28]. Finally, we adopted an EM algorithm rather than the Simplex procedure to estimate the genetic parameters. Of course, we should notice one common assumption of the mentioned-above approaches that marker segregation distortion is caused by some genetic or viability reasons. For genetic reason, there are two different mechanisms for segregation distortion, one at the gametic level and the other at the zygotic level. In both cases, observable phenotypes are distorted for marker loci in the chromosomal region close to the SDL. Thus the two mechanisms are included in our proposed method. Although we have no way to distinguish them in SDL mapping, the results from the genotype and allele tests [29] for the marker closest to the SDL can be used to infer the presence of zygotic or gametic viability selection in an F<sub>2</sub> population but not in backcross, double haploid and recombinant inbred line populations. Moreover, it should be noted that genetic linkage between distorted markers has been carefully discussed in Wu *et al.* (2007) [30].

There are two primary routes by which selection can affect the extent of linkage disequilibrium [31]. The first is a hitchhiking effect, in which an entire haplotype that flanks a favored variant can be rapidly swept to high frequency or even fixation [32]. The second way in which selection can affect linkage disequilibrium is through epistatic selection for combination of alleles at two or more loci on the same chromosome [33]. This selection form leads to the association of the particular alleles at different loci. The major difficulty in linkage disequilibrium-based mapping is to quantify the relationship between recombination fraction and linkage disequilibrium measurement. Our analyses are confined to exclude the factors that influence linkage disequilibrium except linkage and selection. We first combine the viability selection with quantitative genetics model, and then explore the relationship between genetic modes of the viability genes and the estimates of the recombination fraction. The simulation studies indicated that most of the genetic modes of the viability genes at the two linked SDL may result in underestimation of genetic distance. We hope that the tentative attempt will make for elucidating the genetic relationship between viability selection and genetic linkage.



**Figure 1**  
**The likelihood-ratio (LR) score profile for segregation distortion loci mapping against the tomato genome.** The tomato genome derived from Mapmaker 3.0 software (Lander *et al.* 1987) was composed of two linkage groups.

**Table 5: Effect of genetic modes of two linked SDL on the estimates of map distances under the additive-dominant model**

$a_1$	$d_1$	$a_2$	$d_2$	Method	Estimates of map distances (cM)				
					The 1st interval	The 2nd interval	The 3rd interval	The 4th interval	The 5th interval
0.5	0.0	0.5	0.0	Corrected	10.13(1.25)	9.88(1.02)	10.22(1.51)	10.06(1.50)	10.27(1.53)
				Uncorrected	9.99(1.43)	7.27(1.38)	10.00(1.52)	9.91(1.48)	10.16(1.52)
0.5	0.0	-0.5	0.0	Corrected	9.92(1.41)	9.89(1.21)	9.92(1.52)	10.08(1.51)	10.01(1.59)
				Uncorrected	10.01(1.43)	7.20(1.22)	10.01(1.53)	10.17(1.51)	10.07(1.59)
-0.5	0.0	-0.5	0.0	Corrected	10.32(1.58)	10.48(1.53)	9.96(1.33)	10.01(1.75)	10.09(1.59)
				Uncorrected	10.41(1.57)	10.65(1.55)	10.11(1.35)	10.07(1.75)	10.10(1.59)
0.0	0.5	0.0	0.5	Corrected	9.99(1.60)	10.18(1.12)	10.04(1.42)	9.97(1.51)	9.98(1.51)
				Uncorrected	10.03(1.60)	6.28(1.13)	10.08(1.44)	10.02(1.53)	10.01(1.51)
0.0	0.5	0.0	-0.5	Corrected	10.16(1.62)	10.05(1.27)	10.09(1.49)	9.81(1.47)	10.24(1.36)
				Uncorrected	10.16(1.72)	15.64(1.97)	10.09(1.59)	9.81(1.48)	10.24(1.36)
0.0	-0.5	0.0	-0.5	Corrected	10.14(1.33)	9.94(0.95)	10.38(1.50)	10.08(1.38)	10.09(1.63)
				Uncorrected	9.98(1.53)	6.58(1.17)	10.02(1.56)	9.92(1.34)	10.06(1.62)

$a_1$  and  $a_2$  ( $d_1$  and  $d_2$ ) are the additive (dominant) effects of the two SDL.

In addition, it will be interesting and challenging to combine the SDL analysis with QTL mapping to see what the effects of distorted markers has on the results of QTL mapping. While doing this, one may take a risk of detecting false QTL not due to their genetic effects on the quantitative traits but due to violation of the Mendelian segregation law. It will be a great breakthrough in quantitative genetics area if we can develop a method to separate the effects of viability loci from the effects of QTL [14]. By reason of the complexity of the combined analysis, the related investigations will be discussed separately elsewhere.

**Conclusion**

Our results suggested that the proposed method can serve as a powerful alternative to existing methods. Under the

liability model, the new method can simultaneously estimate the position and the effects of SDL as well as the recombination fractions between adjacent markers, and also be used to probe into the genetic mechanism for the bias of uncorrected map distance and to elucidate the relationship between the viability selection and genetic linkage.

**Authors' contributions**

CZ designed and carried out the simulation study, and drafted the manuscript. YMZ conceived of the study, participated in the design, coordinated it and revised the manuscript. All authors read and approved the final manuscript.

**Table 6: Effect of genetic modes of two linked SDL on the estimates of map distances under the epistatic genetic model**

$i_{11}$	$i_{12}$	$i_{21}$	$i_{22}$	Method	Estimates of map distances (cM)				
					The 1st interval	The 2nd interval	The 3rd interval	The 4th interval	The 5th interval
0.5	0.0	0.0	0.0	Corrected	9.91(1.40)	9.94(1.35)	10.10(1.59)	9.90(1.50)	9.98(1.33)
				Uncorrected	9.91(1.40)	9.20(1.35)	10.10(1.59)	9.90(1.50)	9.98(1.33)
-0.5	0.0	0.0	0.0	Corrected	9.74(1.60)	10.40(1.20)	10.18(1.39)	9.48(1.45)	10.10(1.31)
				Uncorrected	9.86(1.62)	14.64(2.02)	10.33(1.50)	9.61(1.46)	10.16(1.32)
0.0	0.5	0.0	0.0	Corrected	10.13(1.39)	9.97(1.37)	10.01(1.66)	10.08(1.57)	10.21(1.69)
				Uncorrected	10.17(1.39)	9.95(1.37)	10.05(1.67)	10.09(1.57)	10.21(1.69)
0.0	-0.5	0.0	0.0	Corrected	10.20(1.56)	10.05(1.52)	10.20(1.52)	10.05(1.58)	10.18(1.59)
				Uncorrected	10.21(1.57)	10.05(1.53)	10.19(1.53)	10.03(1.59)	10.18(1.59)
0.0	0.0	0.0	0.5	Corrected	9.98(1.64)	9.24(1.37)	10.14(1.55)	9.73(1.42)	10.08(1.57)
				Uncorrected	10.00(1.64)	7.25(1.38)	10.15(1.55)	9.73(1.43)	10.08(1.57)
0.0	0.0	0.0	-0.5	Corrected	10.02(1.39)	10.09(1.30)	10.02(1.10)	9.97(1.37)	10.23(1.44)
				Uncorrected	10.01(1.59)	12.89(1.90)	10.01(1.30)	9.97(1.47)	10.23(1.44)

$i_{11}$ ,  $i_{12}$  ( $i_{21}$ ) and  $i_{22}$  are additive × additive, additive × dominant and dominance × dominance epistatic effects of the two SDL, respectively, and the additive and the dominant effects for the two SDL are set up at zero.

**Appendix: Fisher-scoring algorithms for obtaining MLEs of parameters**

The Fisher-scoring algorithm can be used to estimate parameters in the M-step of EM algorithm. Let  $\theta = (a, d)^T$ . The newly estimated  $\theta$  can be expressed by the score-function vector  $S$  and the Fisher information matrix  $I$ ,

$$\theta^{(1)} = \theta^{(0)} + I_{\theta=\theta^{(0)}}^{-1} S_{\theta=\theta^{(0)}}$$

where  $S = \partial \ln L / \partial \theta = (\partial \ln L / \partial a, \partial \ln L / \partial d)^T$  is score function, and

$$I = -E \left( \frac{\partial^2 \ln L}{\partial \theta \partial \eta} \right) = - \begin{pmatrix} E \left( \frac{\partial^2 \ln L}{\partial a^2} \right) & E \left( \frac{\partial^2 \ln L}{\partial a \partial d} \right) \\ E \left( \frac{\partial^2 \ln L}{\partial d \partial a} \right) & E \left( \frac{\partial^2 \ln L}{\partial d^2} \right) \end{pmatrix}$$

is Fisher information matrix.

More specifically, the score function and the Fisher information index of the expected complete data log-likelihood can be derived using

$$\begin{aligned} \frac{\partial \ln L}{\partial \theta} &= \sum_{j=1}^n \left[ w(\phi_{j1}=1) \frac{\partial \ln(f_1)}{\partial \theta} + w(\phi_{j2}=1) \frac{\partial \ln(2f_2)}{\partial \theta} + w(\phi_{j3}=1) \frac{\partial \ln(f_3)}{\partial \theta} - \frac{\partial \ln(f_1+2f_2+f_3)}{\partial \theta} \right] \\ &= \sum_{j=1}^n \left[ \frac{w(\phi_{j1}=1)}{f_1} \frac{\partial(f_1)}{\partial \theta} + \frac{w(\phi_{j2}=1)}{2f_2} \frac{\partial(2f_2)}{\partial \theta} + \frac{w(\phi_{j3}=1)}{f_3} \frac{\partial(f_3)}{\partial \theta} - \frac{\partial(f_1+2f_2+f_3)/\partial \theta}{f_1+2f_2+f_3} \right] \\ &= \sum_{j=1}^n \left[ \sum_{h=1}^3 \frac{w(\phi_{jh}=1)}{f_j} \frac{\partial f_j}{\partial \theta} - \frac{\partial(f_1+2f_2+f_3)/\partial \theta}{f_1+2f_2+f_3} \right] \end{aligned}$$

Let  $\mu_h = \sqrt{2} (2 - h)a + (-1)^h d$  for  $h = 1, 2, 3$

$$\begin{aligned} \frac{\partial f_h}{\partial \theta} &= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 \frac{\partial e^{-\frac{(z_{jh}-\mu_h)^2}{2}}}{\partial \theta} dz_{jh} \\ &= -\frac{1}{\sqrt{2\pi}} \left\{ \int_{-\infty}^0 e^{-\frac{(z_{jh}-\mu_h)^2}{2}} d \left[ -\frac{(z_{jh}-\mu_h)^2}{2} \right] \right\} \frac{\partial \mu_h}{\partial \theta} \\ &= \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{(0-\mu_h)^2}{2} \right] \frac{\partial \mu_h}{\partial \theta} \\ &= \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{\mu_h^2}{2} \right] \frac{\partial \mu_h}{\partial \theta} \end{aligned}$$

with  $\frac{\partial \mu_h}{\partial \theta} = \sqrt{2}(2-h)$  or  $(-1)^h$  when  $\theta = a$  or  $d$  correspondingly. Hence

$$\begin{aligned} \frac{\partial \ln L}{\partial a} &= \left[ \sum_{j=1}^n \sum_{h=1}^3 \frac{w(\phi_{jh}=1)}{f_h} - \frac{n}{f_1+2f_2+f_3} \right] \sum_{h=1}^3 \frac{(2-h)}{\sqrt{\pi}} e^{-\frac{(-\sqrt{2}(2-h)a-(-1)^h d)^2}{2}} \\ \frac{\partial \ln L}{\partial d} &= \sum_{j=1}^n \sum_{h=1}^3 \frac{w(\phi_{jh}=1)(-1)^h e^{-\frac{(-\sqrt{2}(2-h)a-(-1)^h d)^2}{2}}}{\sqrt{2\pi} f_h} - \frac{n(-1)^h e^{-\frac{(-\sqrt{2}a+d)^2}{2} + 2e^{-\frac{(-d)^2}{2}} + e^{-\frac{(\sqrt{2}a+d)^2}{2}}}}{\sqrt{2\pi}(f_1+2f_2+f_3)} \end{aligned}$$

The second partial derivatives are more messy but a general form was found as

$$\begin{aligned} \frac{\partial^2 \ln L}{\partial \theta \partial \eta} &= \sum_{j=1}^n \left\{ \sum_{h=1}^3 w(\phi_{jh}=1) \left[ \frac{1}{f_h} \frac{\partial^2 f_h}{\partial \theta \partial \eta} - \frac{1}{f_h^2} \frac{\partial f_h}{\partial \theta} \frac{\partial f_h}{\partial \eta} \right] - \frac{1}{f_1+2f_2+f_3} \frac{\partial^2 (f_1+2f_2+f_3)}{\partial \theta \partial \eta} \right. \\ &\quad \left. + \frac{1}{(f_1+2f_2+f_3)^2} \frac{\partial(f_1+2f_2+f_3)}{\partial \theta} \frac{\partial(f_1+2f_2+f_3)}{\partial \eta} \right\} \end{aligned}$$

where

$$\begin{aligned} \frac{\partial^2 f_h}{\partial \theta \partial \eta} &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu_h^2}{2}\right) (-\mu_h) \frac{\partial \mu_h}{\partial \theta} \frac{\partial \mu_h}{\partial \eta} + \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu_h^2}{2}\right) \frac{\partial^2 \mu_h}{\partial \theta \partial \eta} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu_h^2}{2}\right) (-\mu_h) \frac{\partial \mu_h}{\partial \theta} \frac{\partial \mu_h}{\partial \eta} \end{aligned}$$

**Acknowledgements**

We are grateful to the Associate Editor and two anonymous reviewers for their constructive comments and suggestions that significantly improved the presentation of the manuscript. The research was supported in part by 973 program (2006CB101708), the National Natural Science Foundation of China (No.30470998; No.30671333), NCET (NCET-05-0489), Specialized Research Fund for the Doctoral Program of Higher Education (20060307008), the Talent Foundation of Nanjing Agricultural University to YMZ; China (No.2005038246) and Jiangsu province (No.0502012C) Post-doctoral Science Foundation to CSZ; and the Program for Changjiang Scholars and Innovative Research Team in University, the Ministry of Education (IRT0432).

**References**

1. Lyttle TVW: **Segregation distortion.** *Annual Review of Genetics* 1991, **25**:511-557.
2. Carr DE, Dudash MR: **Recent approaches into the genetic basis of inbreeding depression in plants.** *Philos Trans R Soc London B* 2003, **358**:1071-1084.
3. Falconer DS, Mackay TFC: *Introduction to Quantitative Genetics* Fourth edition. London: Longman; 1996.
4. Harushima Y, Nakagahra M, Yano M, Sasaki N: **Diverse variation of reproductive barriers in three intraspecific rice crosses.** *Genetics* 2002, **160**:313-322.
5. Hartl DL, Clark AG: *Principles of population genetics* 3rd edition. Sunderland (MA): Sinauer Associates; 1997.
6. Xu Y, Zhu L, Xiao J, Huang N, McCouch SR: **Chromosomal regions associated with segregation distortion of molecular markers in F<sub>2</sub>, backcross, doubled haploid, and recombinant inbred populations in rice (*Oryza sativa* L.).** *Molecular General Genetics* 1997, **253**:535-545.
7. Fu YB, Ritland K: **Evidence for the partial dominance of viability genes contributing to inbreeding depression in *Mimulus guttatus*.** *Genetics* 1994, **136**:323-331.
8. Ritland K: **Inferring the genetic basis of inbreeding depression in plants.** *Genome* 1996, **39**:1-8.
9. Lander E, Botstein D: **Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps.** *Genetics* 1989, **121**:185-199.
10. Hedrick PW, Muona O: **Linkage of viability genes to marker loci in selfing organisms.** *Heredity* 1990, **64**:67-72.

11. Mitchell-Olds T: **Interval mapping of viability loci causing heterosis in Arabidopsis.** *Genetics* 1995, **140**(3):1105-1109.
12. Luo L, Xu SZ: **Mapping viability loci using molecular markers.** *Heredity* 2003, **90**:459-467.
13. Wang CM, Zhu CS, Zhai HQ, Wan JM: **Mapping segregation distortion loci (SDL) and quantitative trait loci (QTL) for spikelet sterility in rice (Oryza sativa L.).** *Genet Res* 2005, **86**:97-106.
14. Luo L, Zhang YM, Xu SZ: **A quantitative genetics model for viability selection.** *Heredity* 2005, **94**:347-355.
15. Nichols RA: **Quantitative genetics focus issue.** *Heredity* 2005, **94**:273-274.
16. Lorieux MB, Perrier GX, Gonzalez de Leon , Lanaud C: **Maximum likelihood models for mapping genetic markers showing segregation distortion. 1. Backcross population.** *Theor Appl Genet* 1995, **90**:73-80.
17. Lorieux M, Perrier X, Goffinet B, Lanaud C, Gonzalez de Leon D: **Maximum likelihood models for mapping genetic markers showing segregation distortion. 2. F<sub>2</sub> population.** *Theor Appl Genet* 1995, **90**:81-89.
18. Zhu C, Wang C, Zhang YM: **Modeling segregation distortion for viability selection I. Reconstruction of linkage maps with distorted markers.** *Theor Appl Genet* 2007, **114**:295-305.
19. Nelder JA, Mead R: **A simplex method for function minimization.** *The Computational Journal* 1965, **7**:308-313.
20. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via EM algorithm.** *J Royal Stat Soc B* 1977, **39**:1-38.
21. Rao SQ, Xu SZ: **Mapping quantitative trait loci for ordered categorical traits in four-way crosses.** *Heredity* 1998, **81**:214-224.
22. Bailey NTJ: *Introduction to the mathematical theory of genetic linkage* Great Britain: Oxford University Press; 1961.
23. Churchill GA, Doerge RK: **Empirical threshold values for quantitative trait mapping.** *Genetics* 1994, **138**:963-971.
24. Visscher PM, Thompson P, Haley CS: **Confidence intervals in QTL mapping by bootstrapping.** *Genetics* 1996, **143**:1013-1020.
25. Press WH, Flanner BP, Teukolsky SA, Vetterling WT: *Numerical Recipes in C++: The Art of Scientific Computing. 2nd version* Cambridge University Press, New York; 2001.
26. Carbonell EA, Gerig TME, Balansard E, Asins MJ: **Interval mapping in the analysis of non-additive quantitative trait loci.** *Biometrics* 1992, **48**:305-315.
27. Lander E, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L: **MAPMAKER: An interactive computer package for construction primary genetic linkage maps of experimental and natural populations.** *Genomics* 1987, **1**:174-181.
28. Huang H, Richardson TE, Carson SD, Bongarten BC: **Genetic analysis of inbreeding depression in plus tree 850.55 of Pinus radiata D. Don. II Genetics of viability genes.** *Theor Appl Genet* 1999, **99**:140-146.
29. Pham JL, Glaszmann JC, Sano R, Barbier P, Ghesquiere A, Second G: **Isozyme markers in rice: genetic analysis and linkage relationships.** *Genome* 1990, **33**:348-359.
30. Wu R, Ma C, Casella G: *Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL* Springer, New York; 2007:123-134.
31. Ardlie KG, Kruglyak L, Seielstad M: **Patterns of linkage disequilibrium in the human genome.** *Nat Rev Genet* 2002, **3**:299-309.
32. Lewontin RC: **The interaction of selection and linkage. I. General considerations: heterotic models.** *Genetics* 1964, **49**:49-67.
33. Cannon GB: **The effects of heterozygosity and recombination on the relative fitness of experimental populations of Drosophila melanogaster.** *Genetics* 1963, **48**:919-942.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

