Theor Appl Genet (2007) 114:295–305
DOI 10.1007/s00122-006-0432-x

ORIGINAL PAPER

Modeling segregation distortion for viability selection I. Reconstruction of linkage maps with distorted markers

Chengsong Zhu · Chunming Wang · Yuan-Ming Zhang

Received: 28 December 2005 / Accepted: 14 October 2006 / Published online: 22 November 2006

Abstract Molecular markers have been widely used to map quantitative trait loci (QTL). The QTL mapping partly relies on accurate linkage maps. The non-Mendelian segregation of markers, which affects not only the estimation of genetic distance between two markers but also the order of markers on a same linkage group, is usually observed in QTL analysis. However, these distorted markers are often ignored in the real data analysis of QTL mapping so that some important information may be lost. In this paper, we developed a multipoint approach via Hidden Markov chain model to reconstruct the linkage maps given a specified gene order while simultaneously making use of distorted, dominant and missing markers in an F_2 population. The new method was compared with the methods in the MapManager and Mapmaker programs, respectively, and verified by a series of Monte Carlo simulation experiments along with a working example. Results showed that the adjusted linkage maps can be used for further QTL or segregation distortion locus (SDL) analysis unless there are strong evidences to prove that all markers show normal Mendelian segregation.

Keywords EM algorithm · Genetic linkage map · Hidden Markov chain · Segregation distortion · Viability selection

Communicated by A. Charcosset.

C. Zhu · Y.-M. Zhang (✉)
Section on Statistical Genomics, State Key Laboratory of Crop Genetics and Germplasm Enhancement/National Center for Soybean Improvement, College of Agriculture, Nanjing Agricultural University, 1 Weigang Road, Nanjing 210095, People's Republic of China
e-mail: soyzhang@njau.edu.cn

C. Wang
Molecular Population Genetics Group, Temasek Life Sciences Laboratory, 1 Research Link, National University of Singapore, Singapore 117604, Singapore

Introduction

Segregation distortion loci (SDL), defined as chromosomal regions that cause distorted segregation ratios, are usually detected by means of the non-Mendelian segregation of markers linked to the SDL (Lyttle 1991; Carr and Dudash 2003). The distortion is presumed to be resulted from altered survival among some classes of gametes at an SDL before fertilization or from viability differences of SDL genotypes post-fertilization but before genotype scoring (Falconer and Mackey 1996). The previous studies showed that segregation distortion affects linkage tests and the estimation of genetic distances (Garcia-Dorado and Gallego 1992; Lorieux et al. 1995a, b). However, most statistical methods used for map construction ignore these distorted markers (Lander and Green 1987; Jiang and Zeng 1997). Therefore, it is necessary to study the effect of distortion on the estimation of genetic distance.

In the construction of genetic linkage maps using molecular markers, there are three steps: firstly clustering markers into linkage groups, secondly estimating pair-wise recombination frequencies in each of the linkage groups and finally optimizing the orders of markers in all linkage groups. Lander and Green (1987) developed a multilocus analysis method using Hidden

Markov chain model to construct genetic linkage maps. Jiang and Zeng (1997) extended the statistical method of Lander and Green (1987) to the situation of systematically dealing with dominant and missing markers in experimental populations derived from two inbred lines. However, the above-mentioned studies did not address the case of distorted markers. Lorieux et al. (1995a, b) used two-point method to reestimate the genetic distance between adjacent markers under the two viability genes model, but this approach raises some problems. For the codominant markers in a F_2 population, firstly, the first derivative of the likelihood function with respect to recombination fraction has no analytical solution, thus Newton–Raphson algorithm has to be used (Edwards 1972). Then, the updated value for recombination frequency may occasionally have lower likelihood. Finally, the convergence may be very slow if the initial value of parameter is far from the maximum likelihood estimate (MLE), especially in high-dimensional spaces (Lander and Green 1987). Therefore, it is necessary to extend the multipoint analysis method to more general situations, considering distorted, dominant and missing markers at the same time. In the paper, we focus on estimating the recombination frequency between distorted markers given a specified order of loci on a linkage group.

Several programs offer options to compute the recombination fractions in case of deviation from the Mendelian hypothesis, i.e., G-Mendel, MapManager, Mapdisto et al, but almost all the cited programs can not validate the corrected values. The reason is that the above investigations have been seldom addressed in theoretical simulation studies. The challenge encountered in modeling viability selection is mainly caused by unavailability of phenotype data on the traits. Luo et al. (2005) considered an imaginary trait, liability, invisible to the investigators but visible to nature. If the individual's trait exceeds a threshold, then the individual survives, otherwise will be eliminated. There is no doubt that the work of Luo and his colleagues is a reminiscence of both modeling segregation distortion from viability selection and comparing the adjusted values of recombination fractions using the cited programs with the true values. In order to validate the reliability and the correctness of the adjusted values of the recombination fractions between distorted markers, we carried out intensive simulations that mimic the segregation distortion derived from viability selection, and compared the results obtained from the new method with those from the cited programs, i.e., MapManager and Mapmaker.

Materials and methods

Genetic model

Let z_j be the liability of the jth individual in the F_2 population under study. It can be indicated by the following linear model

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

where g_j is the genotypic value of the jth individual for the SDL considered, and ε_j is a normally distributed residual variable with mean zero and standard deviation 1.0, which accounts for polygenes that are linked to the markers and for environmental variation (Luo et al. 2005). Three genotypes at this locus, AA, Aa and aa, are assumed to have genotypic values $\sqrt{2}a - d, d,$ and $-\sqrt{2}a - d$, respectively, with a and d indicating additive and dominant effects. We hypothesize that the liability is subject to natural selection. An individual will survive if $z_j \geq 0$ and will be eliminated from the population if $z_j < 0$, since all the sampled individuals have survived from the viability selection, the liability of each genotype will follow a truncated distribution with a cumulative probability, $G_j = k$ ($k = 1, 2, 3$), with

$$f_k = \Pr(z_j \geqslant 0 | G_j = k) = \Phi\left[(2-k)\sqrt{2}a + (-1)^k d\right] \tag{2}$$

where k indexes the genotype at the SDL, and f_k is also referred to as the relative fitness of the kth genotype of the locus. The expected frequencies of three genotypes AA, Aa and aa, respectively, at the SDL will be:

$$p_{AA} = \frac{0.25f_1}{0.25f_1 + 0.5f_2 + 0.25f_3} = \frac{f_1}{f_1 + 2f_2 + f_3} \tag{3}$$

similarly

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

Reconstruction of genetic linkage maps

Given that all markers are neutral and codominant, the observed segregation distortion of markers is caused by one or more SDL near the markers. Other possible sources, such as partial manifestation or structural rearrangements like translocations, will not be considered in this study. As for gametic selection, for convenience, we always assume that only male gametes are affected. The assumption seems to be realistic because pollen grains are more often affected by differential

viability or by different capacity to fertilization, than are ovules.

Let the order of the m markers on a same chromosome be M_1, M_2, ..., M_m, and 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 kth marker, respectively. Similarly, make z_k be indicator for the phenotype of the kth marker M_k. Provided that there is no crossing-over interference among the markers on the considered linkage group, SDL may cause segregation distortion of some or all markers linked to SDL on the chromosome. We assume that the distorted markers are caused by viability selection and three genotypes at each marker locus have different viability coefficients. In other words, the viability coefficients of M_k m_k and m_k m_k relative to M_k M_k at the kth marker are $s_{k,1}$ and $s_{k,2}$, respectively ($0 \le s_{k,1} < +\infty$ and $0 \le s_{k,2} < +\infty$ for $k = 1, 2, ..., m$). Therefore, it is necessary to estimate the $2m$ coefficients if the zygotic selections of all the markers are to occur. As for F_2 population, it is possible to determine which type of selection occurred at a locus by using two successive Chi-square tests (Pham et al. 1990; Lorieux et al. 1995b). In our model, gametic selection is a special case of the zygotic selection. If $s_{k,1}^{(1)} = (1 + s_{k,2})^{(1)}/2 (k = 1,2, ..., m)$, this becomes a gametic selection model. The case of $s_{k,1} = s_{k,2} = 1 (k = 1, 2, ..., m)$ shows usual Mendelian segregation.

If we incorporate the above viability model into the methods of both Lander and Green (1987) and Jiang and Zeng (1997), the logarithm likelihood is defined by

$$\log L = \sum_{i=1}^{n} \log[q'_{z_{i1}} H_{z_{i2}}(r_1) H_{z_{i3}}(r_2), \ldots, H_{z_{im}}(r_{m-1}) c] \tag{4}$$

where n represents population size, $q_{z_{i1}}$ denotes the column vector of the prior probability $\Pr(x_1)$, $q'_{z_{i1}} =$ [Pr ($x_1 = 1$), Pr ($x_1 = 0$), Pr($x_1 = -1$)] and $c' = [1, 1, 1]$, $'$ denotes transpose of a matrix or vector, r_k is the recombination fraction between the kth and (k+1)th markers, and the transition probability matrix $H(r_k)$ from marker M_k to M_{k+1} is

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

It is noted that we should specify appropriate marker matrix elements of zero depending on the information of the markers, i.e, only one element in the transition matrix takes value of 1 and all other equal zero when fully informative markers occur (see also Jiang and Zeng 1997).

There are several methods to obtain the MLEs of both r_k ($k = 1, ..., m-1$) and $s_{k,j}$ ($j = 1, 2$ and $k = 1, 2, ..., m$), we here adopt an Expectation–Maximization (EM) algorithm (Dempster et al. 1977). The procedures are summarized as:

Step 1. *Initialization*: The initial value of the recombination fraction r_k ($k = 1, 2, ..., m-1$) along with the order of markers on the linkage group is obtained by the software Mapmaker 3.0 (Lander et al. 1987). The viability coefficients, $s_{k,1}^{(0)}$ and $s_{k,2}^{(0)}$ for $k = 1, 2, ..., m$, are initialized with 1.

Step 2. *Updating the matrix* A_k : Let $A_k = (a_{k,icd})$, denoted by $P(x_k\ x_{k+1}\ |z_1, ..., z_m)$ in a 3×3 matrix form, where $a_{k,icd}$ is the cth row and dth column element of A_k for the ith individual. Using the multipoint method, the posterior probabilities $P(x_k\ x_{k+1}\ |z_1, ...,z_m)$ for each individual can be calculated by

$$P(x_k x_{k+1}|z_1, \ldots, z_m) = \frac{p(x_k x_{k+1}) p(z_1, \ldots, z_m|x_k, x_{k+1})}{\sum_{x_o,\, x_{o+1}} p(x_o x_{o+1}) p(z_1, \ldots, z_m|x_o, x_{o+1})} \tag{5}$$

Step 3. *Updating the estimate of recombination fraction*: The recombination fraction between the kth and (k+1)th markers can be updated using

$$r_k^{(1)} = \frac{1}{2n}\sum_{i=1}^{n}\left[a_{k,i12} + 2a_{k,i13} + a_{k,i21} + \frac{2r_k^{(0)2}}{r_k^{(0)2} + (1-r_k^{(0)})^2}a_{k,i22} + a_{k,i23} + 2a_{k,i31} + a_{k,i32}\right] - \frac{r_k^{(0)}(1-r_k^{(0)})\left[r_k^{(0)}(s_{k,2}^{(0)} + s_{k+1,2}^{(0)}) - (1-r_k^{(0)})(1 + s_{k,2}^{(0)}s_{k+1,2}^{(0)}) + (1-2r_k^{(0)})(s_{k,1}^{(0)}(1 - s_{k+1,1}^{(0)} + s_{k+1,2}^{(0)}) + (1 - s_{k,1}^{(0)} + s_{k,2}^{(0)})s_{k+1,1}^{(0)})\right]}{D_k} \quad (6)$$

Step 4. *Updating the viability coefficients*: the selection coefficients can be updated as

$$s_{k,1}^{(1)} = \frac{X(k)\sum_i^n (a_{k,i21} + a_{k,i22} + a_{k,i23}) + W(k)\sum_i^n (a_{k-1,i12} + a_{k-1,i22} + a_{k-1,i32})}{2n\left\{X(k)\left[(1 + s_{k+1,2}^{(0)})r_k^{(1)}(1 - r_k^{(1)}) + s_{k+1,1}^{(0)}\left(1 - 2r_k^{(1)} + 2r_k^{(1)2}\right)\right]/D_k + W(k)\left[\left(1 + s_{k-1,2}^{(0)}\right)r_{k-1}^{(1)}\left(1 - r_{k-1}^{(1)}\right) + s_{k-1,1}^{(0)}\left(1 - 2r_{k-1}^{(1)} + 2r_{k-1}^{(1)2}\right)\right]/D_{k-1}\right\}} \quad (7)$$

$$s_{k,2}^{(1)} = \frac{X(k)\sum_i^n (a_{k,i31} + a_{k,i32} + a_{k,i33}) + W(k)\sum_i^n (a_{k-1,i13} + a_{k-1,i23} + a_{k-1,i33})}{n\left\{X(k)\left[r_k^{(1)2} + 2r_k^{(1)}(1 - r_k^{(1)})s_{k+1,2}^{(0)} + s_{k+1,1}^{(0)}\left(1 - r_k^{(1)2}\right)\right]/D_k + W(k)\left[r_{k-1}^{(1)2} + 2r_{k-1}^{(1)}\left(1 - r_{k-1}^{(1)}\right)s_{k-1,2}^{(0)} + s_{k-1,1}^{(0)}\left(1 - r_{k-1}^{(1)2}\right)\right]/D_{k-1}\right\}} \quad (8)$$

where

$$D_k = (1 + s_{k,2}^{(0)}s_{k+1,2}^{(0)})(1 - r_k^{(0)})^2 + 2r_k^{(0)}(1 - r_k^{(0)})\left[s_{k,1}^{(0)}(1 + s_{k+1,2}^{(0)}) + s_{k+1,1}^{(0)}(1 + s_{k,2}^{(0)})\right] + r_k^{(0)2}(s_{k,2}^{(0)} + s_{k+1,2}^{(0)}) + 2(1 - 2r_k^{(0)} + 2r_k^{(0)2}) \times s_{k,1}^{(0)}s_{k+1,1}^{(0)}$$

$X(k)$ and $W(k)$ are indicator variables, $X(k) = 1$ if $1 \leq k < m$ otherwise takes a value of zero; $W(k) = 1$ if $1 < k \leq m$ otherwise takes a value of zero. In addition, we set $s_{k,1}^{(1)} = (1 + s_{k,2}^{(1)})/2$ when gametic selection occurs. As long as the kth marker shows normal Mendelian segregation ($k = 1, ..., m$), the $s_{k,1}^{(1)}$ and $s_{k,2}^{(1)}$ take a value of 1, and the Eq. 6 will be translated into our familiar expression of Jansen and Stam (1994) or Jiang and Zeng (1997). Step 2 is the E-step, and steps 3 and 4 are the M-step. The E-step and M-step are iterated until convergence occurs. The appendix gives a step-by-step derivation of Eqs. 6, 7 and 8.

Applications

Simulation model

Based on the method described in Luo et al. (2005), the genotypes of both distorted markers and SDL in F_2 population could be simulated under the zygotic selection model. The genetic variance (V_g) in an F_2 population is

$$V_g = p_{AA}(\sqrt{2}\,a - d)^2 + p_{Aa}d^2 + p_{aa}\left(-\sqrt{2}\,a - d\right)^2 - \left[p_{AA}\left(\sqrt{2}\,a - d\right) + p_{Aa}d + p_{aa}\left(-\sqrt{2}\,a - d\right)\right]^2 = \frac{4(f_1f_2 + 2f_1f_3 + f_2f_3)a^2 - 2\sqrt{2}(f_1 - f_3)(2f_2 - f_1 - f_3)ad + 8f_2(f_1 + f_3)d^2}{(f_1 + 2f_2 + f_3)^2} \quad (9)$$

and the broad heritability is

$$h_b^2 = V_g/(V_g + 1) \quad (10)$$

Let $dr = d/a$ be dominance ratio. Once broad heritability and dominance ratio of the SDL are known, the additive and dominant effects can be obtained using a numerical algorithm, as described for example by Press et al. (2001), in order to solve Eqs. 9 and 10.

Effect of SDL heritability on the estimation of genetic distances

Eleven equally spaced markers were simulated on a single-chromosome segment of length 100 cM. A single SDL with dominance ratio of 0.5 was located at position 25 cM. One hundred independent simulations were performed for each set of parameters, with sample size 200. The mean and the standard deviations obtained from 100 replicates were given. The differences between the estimated genetic distances considering SDL and those without considering SDL were evaluated by paired comparisons *t* test. The statistical significance threshold for rejecting H_0 was be chosen from central *t* distribution on the basis of Bonferroni argument due to multiple paired comparisons *t* tests on a chromosome (Lander and Botstein 1989). The results are listed in Table 1. The results show that the corrected genetic distances are closer to the corresponding true values than the corresponding uncorrected ones. The standard deviation among the adjusted estimates is smaller than that among the corresponding uncorrected ones. Among most of the intervals, the corrected genetic distance significantly differs from the uncorrected one but not from the corresponding true value. To our surprise and interest, a dramatic increase in the *t* values has been observed when the broad heritability increased from 0.05 to 0.15 and all the peaks of *t* values locate in the third interval where the true SDL resides.

Effect of multiple SDL on the estimation of genetic distances

Having demonstrated the superiority of the proposed method that takes into account the marker segregation distortion, we now implement the new method under the situation of multiple SDL. Eleven SDL were simulated with the same heritability of 0.02, the same dominance ratio of 1.0 and the locations at marker positions 0, 10, ..., 100 cM, respectively. The sample size was 200. The genetic distances between adjacent

Table 1 Effect of SDL heritability on genetic distance between adjacent markers (SDL dominance ratio 0.5, sample size 200 and replication 100)

h_b^2	True value	10	10	10	10	10	10	10	10	10
0.05	Estimate 1	10.16(1.70)	9.88(1.58)	9.90(1.92)	9.81(2.02)	10.03(1.67)	9.64(1.76)	10. 24(1.95)	9.95(1.75)	9.87(1.51)	10.07(1.78)
	Estimate 2	10.21(1.71)	9.83(1.59)	9.77(1.95)	9.68(2.03)	10.10(1.66)	9.69(1.77)	10.25(1.94)	9.96(1.75)	9.87(1.51)	10.07(1.78)
	t value	3.59	4.06	5.63	4.30	3.54	3.62	1.08	1.64	0.44	0.62
0.10	Estimate 1	10.07(1.84)	10.50(1.68)	10.06(1.90)	10.18(1.88)	10.02(1.78)	10.27(1.87)	10.03(1.72)	9.87(1.79)	9.79(1.91)	10.12(2.04)
	Estimate 2	10.19(1.88)	10.64(1.71)	10.35(1.99)	10.48(1.92)	10.16(1.81)	10.34(1.90)	10.08(1.76)	9.89(1.80)	9.80(1.92)	10.12(2.04)
	t value	6.53	7.01	8.39	7.38	5.30	3.18	2.49	2.04	1.74	1.65
0.15	Estimate 1	9.96(1.93)	9.89(2.00)	10.00(1.96)	9.81(1.89)	9.99(1.96)	9.68(1.72)	9.95(1.74)	10.28(1.79)	10.16(1.57)	9.68(1.84)
	Estimate 2	10.17(1.94)	9.64(2.00)	10.53(1.97)	10.42(1.91)	10.36(2.08)	9.81(1.76)	10.01(1.75)	10.35(1.83)	10.19(1.58)	9.68(1.84)
	t value	7.95	11.28	15.88	11.36	8.32	5.09	3.09	3.50	2.40	0.72

h_b^2 is the broad heritability of the liability trait; Estimate 1 and Estimate 2 are the estimates of genetic distance between adjacent markers with and without considering SDL, respectively, and the standard deviations are in parentheses. The difference between the estimates 1 and 2 under the same sample is tested by paired comparisons *t* test. The same is true for Tables 2 and 5

markers with and without considering segregation distortion were computed, respectively. The mean and the standard deviation obtained from 100 replicates are listed in Table 2. Obviously, t results are similar to those obtained for a single SDL. This denotes that the proposed method can correct the bias derived from multiple SDL. As compared with the true length, the total length of the chromosome without considering segregation distortion is usually underestimated.

Comparison of the new method with the methods in the MapManager and Mapmaker programs

In order to confirm the reliability and the correctness of the adjusted values based on our multipoint analysis, we compared the method described in this paper with the methods in the MapManager and Mapmaker programs, respectively. We simulated one chromosome of 100 cM long covered by 11 evenly spaced codominant markers and put two SDL at position 10 and 20 cM (exactly the second and the third marker loci), respectively. One hundred simulation runs were performed for a sample size of 300. Each of datasets was analyzed thrice by the proposed method here, the Mapmaker (Lander et al. 1987) and the MapManager programs (QTX for Windows, Manly et al. 2001). The results are summarized in Table 3. The results show that the corrected genetic distances based on the new method are closer to the true ones whereas the estimates of the genetic distances based on the MapManager and the Mapmaker programs are severely biased (Table 3). In addition, the MapManager QTX has an option called *Allow for segregation distortion* which causes it to use, where possible. There is no significant difference between genetic distances estimated with these two options (data not shown).

A working example

As a demonstration of the proposed method in this paper, we analyzed a real data of a F_2 population containing 157 individuals derived from the rice cross between CPSLO17 and W207-2. A set of 117 SSR markers and one RAPD marker covered 2,423 cM of the genome with an average marker interval of 23 cM. According to the two successive Chi-square tests, there are 39 markers which deviate from Mendelian segregation ratios ($P < 0.05$), accounting for 33.05% of the total markers. This indicates that there are many distorted markers. Therefore, it is necessary to correct the genetic distance between adjacent markers. As an illustrative example, only the data from chromosome 8 were used to demonstrate our method (Table 4).

The genetic distances between adjacent markers are computed twice with and without considering distorted markers, respectively, and some results with segregation distortion of markers are listed in Table 5. The results from Table 5 show that the genetic distances between distorted markers considering segregation distortion are usually larger than those without considering segregation distortion although there are one or two exceptions. In order to clarify the genetic reason for the differences, the genotypic frequencies of two distorted markers were calculated. Given that SDL resided on these distorted markers, the additive and dominant effects of these putative SDL are estimated by using the method of Luo et al. (2005). The results show that the additive and dominant effects of the putative SDL on the fourth marker of chromosome 8 are –0.4851 and –0.4388, respectively, and those on the sixth marker are –0.2788 and –0.0198, respectively. It is obvious that the former is close to the complete dominant model, and the latter close to the additive model. The difference in heritance mode is probably the cause of the different bias of the estimate of genetic distance. Moreover, in order to further understand the behavior of the estimates of the genetic distances between adjacent markers considering segregation distortion, 1,000 bootstrap samples of chromosome 8 as an entity are simulated. The results are listed in Table 6. The standard error and 95% confidence interval considering distorted markers are usually smaller than the corresponding values without taking into account the segregation distortion.

Table 2 Effect of eleven SDL on the genetic distance between adjacent markers

Estimate	Genetic distance between adjacent markers (cM)									
	10	10	10	10	10	10	10	10	10	10
Estimate 1	9.76(1.78)	9.85(1.85)	9.96 (1.54)	9.81(1.88)	9.91(1.82)	9.90(1.82)	9.96(1.64)	9.98(1.57)	9.92(1.57)	9.88(1.74)
Estimate 2	8.90(1.90)	9.00(1.98)	8.76 (1.64)	9.19(2.07)	8.97(2.03)	8.63(2.08)	9.08(1.78)	8.85(1.69)	9.15(1.63)	9.09(1.84)
t value	8.52	8.53	10.23	10.90	8.45	7.72	8.75	7.53	6.73	5.24

For each SDL, its heritability and dominance ratio were set at 2% and 1.0, respectively. Sample size was 200 and replication was 100

Discussion

Segregation distortion of markers is a common phenomenon observed in QTL analysis (Lyttle 1991). Most quantitative geneticists interested in QTL mapping hesitate to use these distorted markers for QTL mapping because the basic assumption of Mendelian segregation is violated. Too many distorted markers will cause tremendous information loss in QTL mapping if these markers are removed from the marker maps (Luo et al. 2005). Furthermore, viability selection is known to bias the estimation of recombination values between consecutive markers and the order of the markers on linkage groups (Lorieux et al. 1995a, b). We describe here an alternative algorithm for reconstructing genetic linkage maps given a specified gene order while considering segregation distortion from viability selection and incomplete information at markers. Our analysis is similar to the method of Lander and Green's (1987) in spirit, but we extend the construction of genetic linkage maps to a more general situation of simultaneously taking into account distorted, dominant and missing markers. The simulated investigations show that the adjusted genetic distances based on the proposed method here are close to the true ones and demonstrate the superiority of the new method over the existing ad hoc method in cited programs, i.e. MapManager, Mapmaker.

The analysis, of course, depends on the assumption of no crossing-over interference. When there is negative crossing-over interference, which is a usual case, the probability of double or triple crossing-over events will be lower than that considered in the algorithm. However, those probabilities are of a lower magnitude in the analysis, and the effect of the assumption of no crossing-over interference is likely to be small for the algorithm.

The accuracy of any maximum-likelihood method of ordering loci is directly related to the quality of the estimation of the recombination frequencies. The genetic distances between adjacent markers can be calculated based on our multipoint method. However, gene order is typically not known. Generally speaking, combinatorial optimization techniques, i.e. simulation annealing algorithm (Kirkpatrick et al. 1983) and heuristic algorithm (Ansari and Hou 1999), can be used together with the methods described above to find the best gene order yielding maximum-likelihood maps with the highest likelihood.

With our approach, it is possible to efficiently construct genetic linkage maps considering distorted, dominant and missing markers at the same time. Analyses can be extended to a general full-sib family or

Table 3 Comparison of the new method with the methods in the Mapmaker and MapManager programs

a_1	d_1	a_2	d_2	True value	10	10	10	10	10	10	10	10
0.5	0.0	0.5	0.0	Estimate 1	10.13(1.25)	9.88(1.02)	10.22(1.41)	10.06(1.50)	10.27(1.53)	10.07(1. 46)	9.97(1.56)	10.06(1.63)	10.05(1.46)	9.97(1.34)
				Estimate 2	9.99(1.43)	7.27(1.38)	10.00(1.52)	9.91(1.48)	10.16(1.52)	10.01(1.45)	9.94(1.55)	10.04(1.62)	10.05(1.46)	9.97(1.34)
				Estimate 3	9.62(1.29)	7.33(1.31)	10.25(1.46)	10.04(1.36)	9.96(1.49)	10.20(1.42)	9.63(1.61)	10.14(1.40)	10.18(1.77)	10.03(1.68)
0.0	0.5	0.0	–0.5	Estimate 1	10.16(1.52)	10.05(1.27)	10.09(1.49)	9.81(1.47)	10.24(1.36)	10.00(1.51)	10.17(1.55)	10.06(1.50)	9.94(1.49)	9.85(1.52)
				Estimate 2	10.16(1.72)	17.04(2.41)	10.09(1.59)	9.81(1.48)	10.24(1.36)	10.00(1.51)	10.17(1.55)	10.06(1.50)	9.94(1.49)	9.75(1.65)
				Estimate 3	10.79(1.61)	16.66(1.97)	10.42(1.57)	10.18(1.61)	10.08(1.33)	10.28(1.59)	10.71(1.49)	10.27(1.65)	10.14(1.35)	10.25(1.65)

Estimate 1, Estimate 2 and Estimate 3 are the estimates obtained from our new method, Mapmaker and MapManager, respectively. Sample size was 300 and replication was 100

Table 4 The two contingency analyses for the markers on the eighth chromosome in the real data analysis

Marker	Number of three genotypes			The allelic test		The genotypic test	
	$n_1(MM)$	$n_2(Mm)$	$n_3(mm)$	χ_1^2	P value	$\chi_2{}^2$	P value
RM506	29.00	74.72	53.28	3.76	0.056	7.88	0.005
RM152	36.18	63.62	57.20	2.81	0.093	11.26	0.001
RM6863	34.25	62.70	60.05	4.24	0.040	14.83	0.000
RM5068	32.69	53.67	70.64	9.17	0.003	34.059	5.4e-9
RM547	32.48	70.84	53.68	2.86	0.091	7.22	0.007
RM331	27.81	77.30	51.89	3.69	0.055	7.42	0.006
RM223	36.11	73.32	47.57	0.83	0.361	2.35	0.125
RM284	32.19	73.69	51.12	2.28	0.131	5.16	0.023
RM256	25.00	78.00	54.00	5.36	0.020	10.72	0.001
RM264	29.23	71.54	56.23	4.64	0.031	10.52	0.001
RM281	29.04	72.93	55.03	4.30	0.038	9.40	0.002

the selfing of outcrossing individuals by changing the transition probabilities between adjacent loci. Although these changes are trivial, they will complicate presentation substantially and the missing phase information needs to be considered. In addition, the multipoint algorithm is relatively important for the full-sib

Table 5 The estimated genetic distances between adjacent markers for eight chromosomes with and without considering distorted markers

Chromosome		Marker interval													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	Estimate 1	7.19	16.96	40.47	10.63	45.30	30.27	31.53	20.10	27.83					
	Estimate 2	6.32	15.18	40.49	10.63	45.30	30.27	31.53	20.10	27.83					
3	Estimate 1	22.32	14.82	22.14	36.37	15.12	50.37	7.93	55.41	13.88	3.38	39.22	5. 93	14.58	5.98
	Estimate 2	22.32	14.82	22.14	36.37	15.12	50.30	7.43	53.17	14.00	3.12	35.95	5. 68	13.56	5.26
4	Estimate 1	33.53	9.19	38.50	14.29	28.09	18.83	15.49							
	Estimate 2	31.85	9.19	38.49	14.11	27.65	18.83	15.49							
6	Estimate 1	9.57	21.26	43.30	11.85	15.05	34.84	48.67	10.39						
	Estimate 2	9.65	20.76	41.99	10.86	15.05	34.84	48.67	10.39						
8	Estimate 1	7.66	15.55	29.12	36.13	25.13	18.01	11.14	34.71	44.91	9.57				
	Estimate 2	7.28	15.50	23.86	32.64	28.50	17.76	11.18	35.06	42.69	8.96				
9	Estimate 1	31.83	31.98	16.52	20.73	16.47	1.00	41.04							
	Estimate 2	27.30	28.06	15.53	20.74	16.47	1.00	41.04							
10	Estimate 1	9.39	16.89	22.69	19.57	40.97	46.31	9.39	16.89	22.69					
	Estimate 2	9.35	16.85	22.63	19.58	39.66	43.55	9.35	16.85	22.63					
12	Estimate 1	17.97	50.51	16.06	21.81	9.58	45.65	4.85	17.97						
	Estimate 2	17.69	50.56	16.06	21.81	7.58	40.79	4.41	17.69						

Table 6 Average (Ave), standard error (SE) and 95% confidence interval (CI) of the genetic distances from 1,000 bootstrap samples for the eighth chromosome in the real data analysis

Marker interval	The corrected genetic distance			The uncorrected genetic distance		
	Ave	SE	CI (95%)	Ave	SE	CI (95%)
1	7.74	0.05	(4.90, 10.64)	7.38	0.05	(5.04, 12.46)
2	15.65	0.08	(10.62, 20.14)	15.64	0.09	(10.31, 20.45)
3	29.77	0.18	(20.30, 39.55)	14.30	0.16	(5.97, 23.04)
4	37.30	0.25	(24.25, 51.92)	33.34	0.32	(17.27, 50.83)
5	25.70	0.16	(17.03, 34.73)	28.66	0.24	(17.23, 41.34)
6	18.12	0.10	(12.53, 23.54)	17.68	0.10	(12.20, 22.95)
7	11.21	0.08	(6.73, 15.67)	10.96	0.08	(6.63, 15.23)
8	35.72	0.22	(23.88, 48.09)	35.88	0.24	(22.68, 48.95)
9	45.44	0.25	(31.18, 58.88)	43.79	0.25	(30.36, 57.07)
10	9.54	0.07	(5.64, 13.13)	9.00	0.07	(5.13, 14.46)

design. For codominant markers, it is possible to determine what type of selection occurred at a locus by using two successive Chi-square tests, the allele test and the genotype test, respectively . As for partial information markers, i.e. dominant or missing markers, the posterior probabilities of three genotypes at marker locus for each individual are computed by multipoint method, prior to conducting the two successive Chi-square tests. By contrast, for a single marker test, the estimates of alleles of allelic frequencies are biased in case of incompletely informative markers (Pham et al. 1990). There is no doubt that considerable number of false positive SDLs in the single marker Chi-square test are present, yet this complication would be solved by a multipoint likelihood ratio test. The related investigations will be discussed in a companion paper.

Segregation distortion is a common occurrence within species and inter specific hybridizations (Whitkus 1998). This may be caused by structural rearrangements, i.e. inversions, which constitute a pre-fertilization mechanism. As an alternative to selected genes, structural rearrangements such as translocations may affect the viability of gametes (Faure et al. 1993). The models described above do not apply in this case, and the probable answer is not a statistical issue.

The source codes for a C++ program, with which the above calculations can be performed and they are available for scientific use from Dr. Zhu (cszhu@sina.com) or Dr. Zhang (soyzhang@njau.edu.cn).

Acknowledgments We are grateful to Dr Charcosset, Prof Melchinger and two anonymous reviewers for their thoughtful criticisms, comments and suggestions, which have been helpful in improving the presentation of the paper and in removing several ambiguities. The research was supported in part by: (1) the National Natural Science Foundation of China (No. 30470998, No. 30671333), Jiangsu Natural Science Foundation (No. BK2005087), NCET (NCET-05-0489), 973 program (2006CB101708) and the Talent Foundation of Nanjing Agricultural Univiversity to Dr. Zhang; (2) China and Jiangsu Postdoctoral Science Foundation to Dr. Zhu (No. 2005038246); and (3) the Program for Changjiang Scholars and Innovative Research Team in University, the Ministry of Education.

Appendix

We use here the F_2 design as an example to infer the maximum likelihood estimate of recombinant fraction between two markers.

One-gene model

Provided that only one marker, M_1, displays zygotic viability selection, the viabilities of genotypes M_1 m_1 and m_1 m_1 relative to M_1 M_1 are s_1 and s_2, respectively. Thus, the frequencies for the three genotypes among the survival individuals after selection are $1/D$ for M_1 M_1, $2s_1/D$ for M_1 m_1, and s_2/D for m_1 m_1, respectively, where $D = 1 + 2s_1 + s_2$. If another marker, M_2, is linked to the marker M_1 with recombinant fraction r. The expected frequencies of nine F_2 genotypes are a function of the viability coefficients and the recombination fraction, arrayed by Fig. a

$$\begin{array}{c} \\ M_1M_1 \\ M_1m_1 \\ m_1m_1 \end{array} \begin{array}{c} \begin{array}{ccc} M_2M_2 & M_2m_2 & m_2m_2 \end{array} \\ \begin{bmatrix} (1-r)^2/D & 2r(1-r)/D & r^2/D \\ 2s_1r(1-r)/D & 2s_1(1-2r+2r^2)/D & 2s_1r(1-r)/D \\ s_2r^2/D & 2s_2r(1-r)/D & s_2(1-r)^2/D \end{bmatrix} \end{array} \tag{A1}$$

The MLE of r is obtained by the EM algorithm for a normal F_2. This is because the first derivative of the likelihood function with respect to recombination fraction contains no information about the viability coefficients s_1 and s_2. It indicates that the estimate of r is not affected by the viability coefficients. Thus we can estimate r directly using the familiar formula in the M step of Jansen and Stam (1994)

$$\hat{r} = \frac{1}{2n}\left[n_{12}+2n_{13}+n_{21}+\frac{2r^2}{r^2+(1-r)^2}n_{22}+n_{23}+2n_{31}+n_{32}\right] \tag{A2}$$

where $n = \sum_{i=1}^{3}\sum_{j=1}^{3} n_{ij}$, these n_{ij} were the number of the nine genotypes above in matrix A1. Parameters s_1 and s_2 are obtained by

$$\hat{s}_1 = \frac{n_{21}+n_{22}+n_{23}}{2(n_{11}+n_{12}+n_{13})} \quad \hat{s}_2 = \frac{n_{31}+n_{32}+n_{33}}{n_{11}+n_{12}+n_{13}} \tag{A3}$$

Based on the Fisher's information matrix, the sample variance of MLE of the recombination fraction can be indicated by

$$V(\hat{r}) = \frac{Dr(1-r)(1-2r+2r^2)}{2n\left[D(1-2r)^2(1-2r+2r^2)+4(1+s_2)r(1-r)(1-2r+2r^2)+4s_1r(1-r)(1-2r)^2\right]} \tag{A4}$$

It is obvious that the sample variance of the recombination fraction is affected by the viability coefficients.

Two-gene model

Provided that two linked markers with recombinant fraction r, say M_1 and M_2, display zygotic viability selection, so the viabilities of genotypes M_1 m_1 and m_1 m_1 relative to M_1 M_1 are $s_{1,1}$ and $s_{1,2}$, respectively; for the marker M_2, similarly, they are $s_{2,1}$ and $s_{2,2}$, respectively. Thus, the expected frequencies of nine F_2 genotypes are a function of the viability coefficients and the recombination fraction, arrayed by

$$F = \begin{array}{c} \\ M_1M_1 \\ M_1m_1 \\ m_1m_1 \end{array} \begin{array}{c} \begin{array}{ccc} M_2M_2 & M_2m_2 & m_2m_2 \end{array} \\ \begin{bmatrix} (1-r)^2/D & 2s_{2,1}r(1-r)/D & s_{2,2}r^2/D \\ 2s_{1,1}r(1-r)/D & 2s_{1,1}s_{2,1}(1-2r+2r^2)/D & 2s_{1,1}s_{2,2}r(1-r)/D \\ s_{1,2}r^2/D & 2s_{1,2}s_{2,1}r(1-r)/D & s_{1,2}s_{2,2}(1-r)^2/D \end{bmatrix} \end{array} \tag{B1}$$

$$= \begin{pmatrix} (1-r)^2 & 2r(1-r) & r^2 \\ 2r(1-r) & (1-2r+2r^2) & 2r(1-r) \\ r^2 & 2r(1-r) & (1-r)^2 \end{pmatrix} \circ \begin{pmatrix} 1/D & s_{2,1}/D & s_{2,2}/D \\ s_{1,1}/D & s_{1,1}s_{2,1}/D & s_{1,1}s_{2,2}/D \\ s_{1,2}/D & s_{1,2}s_{2,1}/D & s_{1,2}s_{2,2}/D \end{pmatrix} \tag{B2}$$

$$= F_r \circ F_{s_{1,1}, s_{1,2}, s_{2,1}, s_{2,2}}$$

where $\circ$ stands for the component-wise product between the two matrices, the first (F_r) only associated with r and the second with $s_{i,j}(i,j=1,2)$ and $D = (1+s_{1,2}s_{2,2})(1-r)^2 + 2r(1-r)[s_{1,1}(1+s_{2,2}) + s_{2,1}(1+s_{1,2})] + r^2(s_{1,2} + s_{2,2}) + 2(1-2r+r^2)s_{1,1}s_{2,1}$. The EM algorithm can be used to obtain the MLE of r based on matrix B1, but this will be difficult to derive because the coefficients within each cell of this matrix contain r. By dividing matrix B1 into two component matrices in B2, however, we can simplify this derivation process. Based on the results of Wu et al. (2005), similarly, the MLE of r can be expressed by

$$\hat{r} = \frac{1}{2n}\left[n_{12} + 2n_{13} + n_{21} + \frac{2r^2}{r^2 + (1-r)^2}n_{22} + n_{23} + 2n_{31} + n_{32}\right] - \frac{r(1-r)}{2D}\frac{\partial D}{\partial r} \tag{B3}$$

where

$$\frac{\partial D}{\partial r} = 2\Big\{r(s_{1,2} + s_{2,2}) - (1-r)(1 + s_{1,2}s_{2,2}) + (1-2r)\big[s_{1,1}(1 - s_{2,1} + s_{2,2}) + s_{2,1}(1 - s_{1,1} + s_{1,2})\big]\Big\}$$

The four viability coefficients $\hat{s}_{i,j}(i,j=1,2)$ can be estimated simultaneously

$$\begin{aligned} \hat{s}_{1,1} &= \frac{D(n_{21} + n_{22} + n_{23})}{2n\big[(1+s_{2,2})r(1-r) + s_{2,1}(1-2r+2r^2)\big]} \\ \hat{s}_{1,2} &= \frac{D(n_{31} + n_{32} + n_{33})}{n\big[r^2 + 2r(1-r)s_{2,2} + s_{2,1}(1-r^2)\big]} \\ \hat{s}_{2,1} &= \frac{D(n_{12} + n_{22} + n_{32})}{2n\big[(1+s_{1,2})r(1-r) + s_{1,1}(1-2r+2r^2)\big]} \\ \hat{s}_{2,2} &= \frac{D(n_{13} + n_{23} + n_{33})}{n\big[r^2 + 2r(1-r)s_{1,2} + s_{1,1}(1-r^2)\big]} \end{aligned} \tag{B4}$$

As for multiple viability loci, the viability coefficients are affected by its adjacent marker loci (i.e. the left and the right viability loci). Thus the coefficients of viability can be expressed as presented above.

References

Ansari N, Hou E (1999) The design and analysis of computer algorithm. Addison-Wesley, Reading

Carr DE, Dudash MR (2003) Recent approaches into the genetic basis of inbreeding depression in plants. Phil Trans R Soc Lond 358:1071–1084

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

Edwards AWF (1972) Likelihood. The John Hopkins University Press, Baltimore

Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics. Longman, London

Faure S, Noyer JL, Horry JP, Bakry F, Lanaud C, Gonzalez de Leon (1993) A molecular marker-based linkage map of diploid bananas. Theor Appl Genet 87:517–526

Garcia-Dorado A, Gallego A (1992) On the use of the classical tests for detecting linkage. J Heredity 83:143–146

Jansen RC, Stam P (1994) High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136:1447–1455

Jiang CJ, Zeng ZB (1997) Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica 101:47–56

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680

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

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367

Lander E, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and nature populations. Genomics 1:174–181

Lorieux MB, Perrier GX, Gonzalez de Leon D, Lanaud C (1995a) Maximum likelihood models for mapping genetic

markers showing segregation distortion. 1. Backcross population. Theor Appl Genet 90:73–80

Lorieux M, Perrier X, Goffinet B, Lanaud C, Gonzalez de Leon D (1995b) Maximum likelihood models for mapping genetic markers showing segregation distortion. 2. F_2 population. Theor Appl Genet 90:81–89

Luo L, Zhang YM, Xu S (2005) A quantitative genetics model for viability selection. Heredity 94: 347–355

Lyttle TW (1991) Segregation distortion. Ann Rev Genetics 25:511–557

Manly KF, Cudmore RH, Meer JM (2001) Map Manager QTX cross-platform software for genetic mapping. Mammal Genome 12:930–932

Pham JL, Glaszmann JC, Sano R, Barbier P, Ghesquiere A, Second G (1990) Isozyme markers in rice: genetic analysis and linkage relationships. Genome 33:348–359

Press WH, Flanner BP, Teukolsky SA, Vellerting WT (2001) Numerical recipes in C++: the art of scientific computing, 2nd version. Cambridge University Press, New York

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

Wu RL, Ma CX, Casella G (2005) Statistical genomics of complex traits. Springer, Berlin Heidelberg New York