Heredity (2014) 112, 479–488

www.nature.com/hdy

ORIGINAL ARTICLE

Linkage group correction using epistatic distorted markers in F_2 and backcross populations

S-Q Xie[1], J-Y Feng[1] and Y-M Zhang

Epistasis has been frequently observed in all types of mapping populations. However, relatively little is known about the effect of epistatic distorted markers on linkage group construction. In this study, a new approach was proposed to correct the recombination fraction between epistatic distorted markers in backcross and F_2 populations under the framework of fitness and liability models. The information for three or four markers flanking with an epistatic segregation distortion locus was used to estimate the recombination fraction by the maximum likelihood method, implemented via an expectation–maximisation algorithm. A set of Monte Carlo simulation experiments along with a real data analysis in rice was performed to validate the new method. The results showed that the estimates from the new method are unbiased. In addition, five statistical properties for the new method in a backcross were summarised and confirmed by theoretical, simulated and real data analyses.
Heredity (2014) **112,** 479–488; doi:10.1038/hdy.2013.127; published online 5 March 2014

INTRODUCTION

The non-Mendelian segregation of markers, known as distorted segregation, is a common biological phenomenon and has been reported since the early twentieth century (Mangelsdorf and Jones, 1926; Sandler *et al.*, 1959; Rick, 1966; McCouch *et al.*, 1988; Paterson *et al.*, 1988; Brummer *et al.*, 1993; Xu *et al.*, 1997; Kaló *et al.*, 2000; Lu *et al.*, 2002; Barchi *et al.*, 2010). It may lead to a biased estimate of the recombination fraction and affect the accuracy of linkage groups (Lorieux *et al.*, 1995a,b). For example, slight but significant segregation distortion results in a reduced estimate of the recombination fraction (Cloutier *et al.*, 1997; Kaló *et al.*, 2000), and an overwhelming number of heterozygous individuals in the F_2 population leads to a false genetic linkage of markers (Kaló *et al.*, 2000) and the overestimation of the recombination fraction (Lashermes *et al.*, 2001). These conclusions are not contradictory and can be clearly explained. More specifically, two linked segregation distortion loci (SDL) underestimate the recombinant fraction in most cases and overestimate the recombinant fraction under an additive model with opposite additive effects (Zhu *et al.*, 2007). Therefore, the importance of accurate genetic linkage groups necessitates an in-depth study of marker segregation distortion.

To date, several approaches have been proposed to construct linkage groups. Lander and Green (1987) developed a multi-point method using a Hidden Markov chain model. Jiang and Zeng (1997) extended the multi-point method suitable for dominant and missing markers. However, a question remains how can distorted markers be utilised in the construction of linkage groups? The simplest method is to exclude significantly distorted markers from linkage groups, but this treatment usually reduces the coverage and saturation of the genome (Wang *et al.*, 2005). The most common method is to insert distorted markers into a linkage group. If the new linkage group is seriously different from the old one, the recombination fraction between distorted markers should be re-estimated. However, the traditional approach does not work well because a new variable, selection coefficient, is involved (Kärkkäinen *et al.*, 1996; Kreike and Stiekema, 1997; Faris *et al.*, 1998). To overcome this issue, Lorieux *et al.* (1995a,b) regarded the selection coefficient as a parameter and adopted the maximum likelihood method to estimate the recombination fraction and selection coefficient simultaneously under a fitness model. Compared with the traditional method, this approach leads to more precise linkage groups, and new software, named MapDisto, is available (Lorieux, 2012). Recently, Zhu *et al.* (2007) further extended the multi-point method suitable for distorted, dominant and missing markers under the framework of a quantitative genetics model for viability selection (Luo *et al.*, 2005). However, epistatic distorted markers have been not considered in the above methods.

Epistasis, the interaction between loci, has been shown to have a strong association with segregation distortion (Bomblies *et al.*, 2007; Alheit *et al.*, 2011). Epistatic SDL has a significant implication for inbreeding depression (Phillips, 2008), which is mainly manifested as hybrid male or female sterility. Törjék *et al.* (2006) reported that marker segregation distortion is due to reduced fertility caused by epistasis. Kubo *et al.* (2008) showed that hybrid male sterility is caused by epistasis between two novel genes, *S24* and *S35*, on rice chromosomes 5 and 1. Similar results have also been found in *Drosophila* (Chang and Noor, 2010), alfalfa (Li *et al.*, 2011), rice (Xie and Chen, 2012; Yang *et al.*, 2012) and *Arabidopsis lyrata* (Leppälä *et al.*, 2013). Thus, the Dobzhansky–Muller model, in which hybrid inviability is assumed to be caused by epistasis (Dobzhansky, 1936; Muller, 1942), has been widely accepted. In addition, McMullen *et al.*

Section on Statistical Genomics, State Key Laboratory of Crop Genetics and Germplasm Enhancement/Collaborative Innovation Center for Modern Crop Production, Department of Crop Genetics and Breeding, College of Agriculture, Nanjing Agricultural University, Nanjing, China
[1]These authors contributed equally to this work.
Correspondence: Dr Y-M Zhang, College of Agriculture, Nanjing Agricultural University, Nanjing 210095, China.
E-mail: soyzhang@hotmail.com or soyzhang@njau.edu.cn
Received 22 February 2013; revised 29 September 2013; accepted 28 October 2013; published online 5 March 2014

(2009) investigated genome-wide segregation distortion among nested association mapping populations and indicated that epistasis affected fitness. Therefore, epistatic SDL should be considered in the construction of precise linkage groups.

In this study, we integrated the fitness model for viability selection with the liability model and developed a new method to correct the recombination fraction between epistatic distorted markers in backcross and F_2 populations. A series of simulated data sets along with a real data set was analysed to validate the proposed method, and the statistical properties of the new method were summarised and confirmed.

MATERIALS AND METHODS

Genetic model in a backcross population

The new method in this study was developed on the basis of a backcross population. The extension to F_2 populations is mentioned briefly in a subsequent section. In this study, the recombinant fraction between epistatic distorted markers was corrected, and the molecular marker information from all n individuals was used to detect the epistatic SDL under the liability and fitness models. The gametic and zygotic selections in the backcross are the same. Thus, the two cases are discussed together.

Liability model. If the selection in a backcross is controlled by two linked SDL, with a recombinant fraction of r, the liability z_j of the jth individual may be described by the following model:

$$z_j = x_{j1}a_1 + x_{j2}a_2 + x_{j1}x_{j2}i + \varepsilon_j \quad (1)$$

where a_k is the main effect of the kth SDL ($k=1, 2$); i is the epistatic effect between the two SDL; two genotypes for any one locus are assumed to be *SS* and *Ss*, respectively; x_{jk} is the dummy variable defined as $x_{jk}=1$ for SDL homozygote *SS* and as $x_{jk}=-1$ for SDL heterozygote *Ss*; and $\epsilon_j \sim N(0, \sigma^2)$ is a normally distributed residual error. In addition, set $\sigma^2=1$ for convenience (Luo *et al.*, 2005). The model (1) can be simply expressed as

$$z_j = X_j b + \varepsilon_j \quad (2)$$

We hypothesise that the liability is subject to natural selection. An individual will survive if $z_j \geqslant 0$ and will be eliminated from the population if $z_j < 0$. As all of the sampled individuals have survived from the viability selection, the liability of each observed individual will follow a truncated normal distribution with a cumulative probability:

$$\Pr(z_j \geq 0) = \Phi(X_j b) \quad (3)$$

This result may be considered to be the relative fitness for individual j and is denoted by $\Phi(X_j b)$. Because four possible genotypes for two linked SDL exist, the relative fitness f_l^B ($l=1,\ldots,4$) can be easily defined. Therefore, the expected frequencies p_l^{Lb} of the four genotypes after selection are easily calculated and are listed in Table 1.

Fitness model. In the fitness model, the viability coefficients for the S_1s_2, s_1S_2 and s_1s_2 gametes relative to S_1S_2 are defined to be v, u and x, respectively, which means that the fitnesses for $S_1S_1S_2S_2$, $S_1S_1S_2s_2$, $S_1s_1S_2S_2$ and $S_1s_1S_2s_2$ in the backcross are 1, v, u and x, respectively. The case $u=v=x=1$ indicates no selection, which is a typical Mendelian segregation. Therefore, the expected frequencies p_l^{Fb} ($l=1,\ldots,4$) of the above four genotypes among surviving individuals are also easily calculated and are listed in Table 1.

Table 1 Expected frequencies of four genotypes under the liability and fitness models in a backcross population

Genotype	*Relative fitness* (f_l^B)	p_l^{Lb} *in liability model*	p_l^{Fb} *in fitness model*
S_1S_2/S_1S_2	$\Phi(a_1+a_2+i)$	$(1-r)f_1^B/d$	$(1-r)/D$
S_1s_2/S_1S_2	$\Phi(a_1-a_2-i)$	rf_2^B/d	r/D
s_1S_2/S_1S_2	$\Phi(-a_1+a_2-i)$	rf_3^B/d	r/D
s_1s_2/S_1S_2	$\Phi(-a_1-a_2+i)$	$(1-r)f_4^B/d$	$(1-r)x/D$

$D=(1-r)(x+1)+r(u+v)$; $d=(1-r)(f_1^B+f_4^B)+r(f_2^B+f_3^B)$

Relationship between parameters in the above two models. The expected frequencies of one genotype under the liability and fitness models should be the same, that is, $p_l^{Lb}=p_l^{Fb}$ ($l=1,\ldots,4$). Therefore, the relationship between parameters in the two models can be expressed as

$$[u \quad v \quad x] = \begin{bmatrix} \frac{\Phi(-a_1+a_2-i)}{\Phi(a_1+a_2+i)} & \frac{\Phi(a_1-a_2-i)}{\Phi(a_1+a_2+i)} & \frac{\Phi(-a_1-a_2+i)}{\Phi(a_1+a_2+i)} \end{bmatrix} \quad (4)$$

Likelihood function and parameter estimation in a backcross

Although the genotypes of two SDL in the above two models are unobserved, the genotypes of markers flanking with the SDL are observed. Assume that two loci, S_1 and S_2, are located between markers A and B and between markers C and D, respectively, and that the recombination fractions between A and S_1, between S_1 and B, between B and C, between C and S_2 and between S_2 and D are r_1, r_2, r_{BC}, r_3 and r_4, respectively. The expected frequencies of the 16 observed genotypes of markers A, B, C and D are calculated and listed in Table 2.

Let n_k and p_k ($k=1,\ldots,16$) be the observed number and expected frequencies of the kth genotype for the four markers and $n=\sum_{k=1}^{16} n_k$ be the total number of all individuals. The likelihood function in a backcross is

$$L = \frac{n!}{\prod_k n_k!} \prod_k p_k^{n_k} \quad (5)$$

However, the maximum likelihood estimate in equation (5) is complicated. Thus, the complete information that includes all 64 genotypes for four markers and two SDL was used to construct the likelihood function, which is expressed as

$$L = \frac{n!}{\prod_{k,l} n_{kl}!} \prod_{k,l} p_{kl}^{n_{kl}} \quad (6)$$

where p_{kl} and n_{kl} ($k=1,\ldots,16$; $l=1,\ldots,4$) are the expected frequency and the observed number for the kth marker genotype and the lth SDL genotype, respectively, and $n_{kl}=\frac{p_{kl}}{p_k}\times n_k$. Theoretically, the Newtow–Raphson method may be used to obtain the maximum likelihood estimates in equation (6). Here, we adopt the expectation–maximisation (EM) algorithm (Dempster *et al.*, 1977). The logarithm likelihood function is

$$\begin{aligned}
\ln L = & \left(\sum_{k=1}^{8}\sum_{l=1}^{2} n_{kl} + \sum_{k=9}^{16}\sum_{l=3}^{4} n_{kl}\right)\ln(1-r_1) \\
& + \left(\sum_{k=1}^{8}\sum_{l=3}^{4} n_{ij} + \sum_{k=9}^{16}\sum_{l=1}^{2} n_{kl}\right)\ln(r_1) \\
& + \left[\sum_{l=1}^{2}\left(\sum_{k=1}^{4} n_{kl} + \sum_{k=9}^{12} n_{kl}\right) + \sum_{l=3}^{4}\left(\sum_{k=5}^{8} n_{kl} + \sum_{k=13}^{16} n_{kl}\right)\right]\ln(1-r_2) \\
& + \left[\sum_{l=1}^{2}\left(\sum_{k=5}^{8} n_{kl} + \sum_{k=13}^{16} n_{kl}\right) + \sum_{l=3}^{4}\left(\sum_{k=1}^{4} n_{kl} + \sum_{k=9}^{12} n_{kl}\right)\right]\ln(r_2) \\
& + \left(\sum_{k=1}^{2} n_k + \sum_{k=7}^{10} n_k + \sum_{k=15}^{16} n_k\right)\ln(1-r_{BC}) + \left(\sum_{k=3}^{6} n_k + \sum_{k=11}^{14} n_k\right)\ln(r_{BC}) \\
& + \left(\sum_{k=1}^{4}\sum_{l=1}^{2}\left(n_{4k-3,2l-1} + n_{4k-2,2l-1} + n_{4k-1,2l} + n_{4k,2l}\right)\right)\ln(1-r_3) \\
& + \left[\sum_{k=1}^{4}\sum_{l=1}^{2}\left(n_{4k-3,2l} + n_{4k-2,2l} + n_{4k-1,2l-1} + n_{4k,2l-1}\right)\right]\ln(r_3) \\
& + \left[\sum_{k=1}^{8}\sum_{l=1}^{2}\left(n_{2k,2l-1} + n_{2k-1,2l}\right)\right]\ln(1-r_4) \\
& + \left[\sum_{k=1}^{8}\sum_{l=1}^{2}\left(n_{2k-1,2l} + n_{2k,2l-1}\right)\right]\ln(r_4) + \sum_{k=1}^{16} n_{k3}\ln(u) \\
& + \sum_{k=1}^{16} n_{k2}\ln(v) + \sum_{k=1}^{16} n_{k4}\ln(x) - n\ln(d)
\end{aligned} \quad (7)$$

where $d=(1-r)(f_1^B+f_4^B)+r(f_2^B+f_3^B)$. The maximum likelihood estimate of each parameter is found by setting its partial derivative to zero and solving

Table 2 Expected frequencies of the 16 genotypes of markers A, B, C and D under the epistatic SDL genetic model in a backcross population

Genotype S_1S_2	S_1s_2	s_1S_2	s_1s_2	*Observed count* n_i
ABCD $(1-r_1)(1-r_2)(1-r_{BC})(1-r_3)(1-r_4)/d$	$(1-r_1)(1-r_2)(1-r_{BC})r_3r_4v/d$	$r_1r_2(1-r_{BC})(1-r_3)(1-r_4)u/d$	$r_1r_2(1-r_{BC})r_3r_4x/d$	n_1
ABCd $(1-r_1)(1-r_2)(1-r_{BC})(1-r_3)r_4/d$	$(1-r_1)(1-r_2)(1-r_{BC})r_3(1-r_4)v/d$	$r_1r_2(1-r_{BC})(1-r_3)r_4u/d$	$r_1r_2(1-r_{BC})r_3(1-r_4)x/d$	n_2
ABcD $(1-r_1)(1-r_2)r_{BC}r_3(1-r_4)/d$	$(1-r_1)(1-r_2)r_{BC}(1-r_3)r_4v/d$	$r_1r_2r_{BC}r_3(1-r_4)u/d$	$r_1r_2r_{BC}(1-r_3)r_4x/d$	n_3
ABcd $(1-r_1)(1-r_2)r_{BC}r_3r_4/d$	$(1-r_1)(1-r_2)r_{BC}(1-r_3)(1-r_4)v/d$	$r_1r_2r_{BC}r_3r_4u/d$	$r_1r_2r_{BC}(1-r_3)(1-r_4)x/d$	n_4
AbCD $(1-r_1)r_2r_{BC}(1-r_3)(1-r_4)/d$	$(1-r_1)r_2r_{BC}r_3r_4v/d$	$r_1(1-r_2)r_{BC}(1-r_3)(1-r_4)u/d$	$r_1(1-r_2)r_{BC}r_3r_4x/d$	n_5
AbCd $(1-r_1)r_2r_{BC}(1-r_3)r_4/d$	$(1-r_1)r_2r_{BC}r_3(1-r_4)v/d$	$r_1(1-r_2)r_{BC}(1-r_3)(1-r_4)u/d$	$r_1(1-r_2)r_{BC}r_3(1-r_4)x/d$	n_6
AbcD $(1-r_1)r_2(1-r_{BC})r_3(1-r_4)/d$	$(1-r_1)r_2(1-r_{BC})(1-r_3)r_4v/d$	$r_1(1-r_2)(1-r_{BC})r_3(1-r_4)u/d$	$r_1(1-r_2)(1-r_{BC})(1-r_3)r_4x/d$	n_7
Abcd $(1-r_1)r_2(1-r_{BC})r_3r_4/d$	$(1-r_1)r_2(1-r_{BC})(1-r_3)(1-r_4)v/d$	$r_1(1-r_2)(1-r_{BC})r_3r_4u/d$	$r_1(1-r_2)(1-r_{BC})(1-r_3)(1-r_4)x/d$	n_8
aBCD $r_1(1-r_2)(1-r_{BC})(1-r_3)(1-r_4)/d$	$r_1(1-r_2)(1-r_{BC})r_3r_4v/d$	$(1-r_1)r_2(1-r_{BC})(1-r_3)(1-r_4)u/d$	$(1-r_1)r_2(1-r_{BC})r_3r_4x/d$	n_9
aBCd $r_1(1-r_2)(1-r_{BC})(1-r_3)r_4/d$	$r_1(1-r_2)(1-r_{BC})r_3(1-r_4)v/d$	$(1-r_1)r_2(1-r_{BC})(1-r_3)r_4u/d$	$(1-r_1)r_2(1-r_{BC})r_3(1-r_4)x/d$	n_{10}
aBcD $r_1(1-r_2)r_{BC}r_3(1-r_4)/d$	$r_1(1-r_2)r_{BC}(1-r_3)r_4v/d$	$(1-r_1)r_2r_{BC}r_3(1-r_4)u/d$	$(1-r_1)r_2r_{BC}(1-r_3)r_4x/d$	n_{11}
aBcd $r_1(1-r_2)r_{BC}r_3r_4/d$	$r_1(1-r_2)r_{BC}(1-r_3)(1-r_4)v/d$	$(1-r_1)r_2r_{BC}r_3r_4u/d$	$(1-r_1)r_2r_{BC}(1-r_3)(1-r_4)x/d$	n_{12}
abCD $r_1r_2r_{BC}(1-r_3)(1-r_4)/d$	$r_1r_2r_{BC}r_3r_4v/d$	$(1-r_1)(1-r_2)r_{BC}(1-r_3)(1-r_4)u/d$	$(1-r_1)(1-r_2)r_{BC}r_3r_4x/d$	n_{13}
abCd $r_1r_2r_{BC}(1-r_3)r_4/d$	$r_1r_2r_{BC}r_3(1-r_4)v/d$	$(1-r_1)(1-r_2)r_{BC}(1-r_3)r_4u/d$	$(1-r_1)(1-r_2)r_{BC}r_3(1-r_4)x/d$	n_{14}
abcD $r_1r_2(1-r_{BC})r_3(1-r_4)/d$	$r_1r_2(1-r_{BC})(1-r_3)r_4v/d$	$(1-r_1)(1-r_2)(1-r_{BC})r_3(1-r_4)u/d$	$(1-r_1)(1-r_2)(1-r_{BC})(1-r_3)r_4x/d$	n_{15}
abcd $r_1r_2(1-r_{BC})r_3r_4/d$	$r_1r_2(1-r_{BC})(1-r_3)(1-r_4)v/d$	$(1-r_1)(1-r_2)(1-r_{BC})r_3r_4u/d$	$(1-r_1)(1-r_2)(1-r_{BC})(1-r_3)(1-r_4)x/d$	n_{16}

$d=[r_{BC}(r_2+r_3-2r_2r_3)+(1-r_{BC})(1-r_2-r_3+2r_2r_3)](1+x)+[(1-r_{BC})(r_2+r_3-2r_2r_3)+r_{BC}(1-r_2-r_3+2r_2r_3)](v+u)$.

the equation to obtain

$$
\begin{aligned}
r_1 &= \frac{1}{n}\sum_{k=1}^{8}\sum_{l=3}^{4} n_{kl} + \frac{1}{n}\sum_{k=9}^{16}\sum_{l=1}^{2} n_{kl}\\
r_2 &= \frac{t_{22}[(r_3+r_{BC}-2r_3r_{BC})(u+v-x-1)+x+1]}{t_{21}[(r_3+r_{BC}-2r_3r_{BC})(x+1-u-v)+u+v]+t_{22}[(r_3+r_{BC}-2r_3r_{BC})(u+v-x-1)+x+1]}\\
r_{BC} &= \frac{t_{32}[(r_2+r_3-2r_2r_3)(u+v-x-1)+x+1]}{t_{31}[(r_2+r_3-2r_2r_3)(x+1-u-v)+u+v]+t_{32}[(r_2+r_3-2r_2r_3)(u+v-x-1)+x+1]}\\
r_3 &= \frac{t_{42}[(r_2+r_{BC}-2r_2r_{BC})(u+v-x-1)+x+1]}{t_{41}[(r_2+r_{BC}-2r_2r_{BC})(x+1-u-v)+u+v]+t_{42}[(r_2+r_{BC}-2r_2r_{BC})(u+v-x-1)+x+1]}\\
r_4 &= \frac{1}{n}\sum_{k=1}^{8}\sum_{l=1}^{2}\left(n_{2k-1,2l}+n_{2k,2l-1}\right)\\
u &= \frac{(1-r_2-r_3-r_{BC}+2r_2r_3+2r_2r_{BC}+2r_3r_{BC}-4r_2r_3r_{BC})\sum_{k=1}^{16} n_{k3}}{(r_2+r_3+r_{BC}-2r_2r_3-2r_2r_{BC}-2r_3r_{BC}+4r_2r_3r_{BC})\left(n-\sum_{k=1}^{16}\sum_{l=2}^{4} n_{kl}\right)}\\
v &= \frac{(1-r_2-r_3-r_{BC}+2r_2r_3+2r_2r_{BC}+2r_3r_{BC}-4r_2r_3r_{BC})\sum_{k=1}^{16} n_{k2}}{(r_2+r_3+r_{BC}-2r_2r_3-2r_2r_{BC}-2r_3r_{BC}+4r_2r_3r_{BC})\left(n-\sum_{k=1}^{16}\sum_{l=2}^{4} n_{kl}\right)}\\
x &= \frac{\sum_{k=1}^{16} n_{k4}}{n-\sum_{k=1}^{16}\sum_{l=2}^{4} n_{kl}}
\end{aligned}
\qquad (8)
$$

where $t_{21}=\sum_{l=1}^{2}\left(\sum_{k=1}^{4} n_{kl}+\sum_{k=9}^{12} n_{kl}\right)+\sum_{l=3}^{4}\left(\sum_{k=5}^{8} n_{kl}+\sum_{k=13}^{16} n_{kl}\right)$, $t_{22}=\sum_{l=1}^{2}\left(\sum_{k=5}^{8} n_{kl}+\sum_{k=13}^{16} n_{kl}\right)+\sum_{l=3}^{4}\left(\sum_{k=1}^{4} n_{kl}+\sum_{k=9}^{12} n_{kl}\right)$, $t_{31}=\sum_{k=1}^{2} n_k+\sum_{k=7}^{10} n_k+\sum_{k=15}^{16} n_k$, $t_{32}=\sum_{k=3}^{6} n_k+\sum_{k=11}^{14} n_k$, $t_{41}=\sum_{k=1}^{4}\sum_{l=1}^{2}\left(n_{4k-3,2l-1}+n_{4k-2,2l-1}+n_{4k-1,2l}+n_{4k,2l}\right)$ and $t_{42}=\sum_{k=1}^{4}\sum_{l=1}^{2}\left(n_{4k-3,2l}+n_{4k-2,2l}+n_{4k-1,2l-1}+n_{4k,2l-1}\right)$. The estimates for r_1 and r_2 were used to correct the recombination fraction between markers A and B: $r_{AB}=r_1+r_2-2r_1r_2$; similarly, $r_{CD}=r_3+r_4-2r_3r_4$. When m markers are located in a linkage group, the number of estimates for r_{AB} is C_m^2. Among these estimates, some may be overestimated and some may be underestimated; in this study, the median is our suggested estimate, which is validated by Monte Carlo simulation experiments. Although only selection parameters u, v and x were estimated, these parameters in the fitness model can be transferred to those in the liability model using equation (4). Therefore, only the estimates of parameters in the fitness model are given in this study.

Variance of recombination fraction. The expected Fisher's information score of the recombination fraction is given by

$$I(r) = -E\left(\frac{\partial^2 \ln L}{\partial r^2}\right) \qquad (9)$$

Where $\ln L=(n_{AB}+n_{ab})\ln(1-r)+(n_{Ab}+n_{aB})\ln r+n_{Ab}\ln v+n_{aB}\ln u+n_{ab}\ln x-n\ln[(1-r)(x+1)+r(u+v)]$. For large samples, the variance of r was estimated by

$$Var(\hat{r}) = \frac{1}{I(r)} = \frac{r(1-r)[(1-r)(x+1)+r(u+v)]^2}{n(x+1)(u+v)} \qquad (10)$$

Genetic model under zygotic selection in the F_2 population

Liability model. The liability z_j of the jth F_2 individual under study could be described by the following model:

$$z_j = x_{j11}a_1+x_{j12}d_1+x_{j21}a_2+x_{j22}d_2+x_{j11}x_{j21}i+x_{j11}x_{j22}j_{12}+x_{j12}x_{j21}j_{21}+x_{j12}x_{j22}l+\varepsilon_j \qquad (11)$$

where a_k and d_k are the additive and dominant effects of the kth SDL ($k=1, 2$), respectively; i, j_{12}, j_{21} and l are the additive-by-additive, additive-by-dominant, dominant-by-additive and dominant-by-dominant interaction effects of the two SDL, respectively; x_{j}.. is the dummy variable defined as $x_{jk1}=1$ and $x_{jk2}=0$ for SDL homozygote *SS*, $x_{jk1}=0$ and $x_{jk2}=1$ for SDL heterozygote *Ss* and $x_{jk1}=-1$ and $x_{jk2}=0$ for SDL homozygote *ss* ($k=1, 2$); and the other variables are similar to those in model (1). As nine possible genotypes for two linked SDL exist, the relative fitness f_l ($l=1,\ldots,9$) can be easily calculated, and both the results and the expected frequencies p_l^{Lz} are listed in Table 3.

Fitness model. Two SDL under study are linked with a recombination fraction of r. In zygotic selection, the viabilities of $S_1S_1S_2S_2$, $S_1S_1s_2s_2$, $S_1s_1S_2S_2$, $S_1s_1S_2s_2$, $S_1s_1s_2S_2$, $s_1s_1S_2S_2$, $s_1s_1S_2s_2$, $s_1s_1s_2S_2$ and $s_1s_1s_2s_2$ relative to $S_1S_1S_2S_2$ are assumed to be v_2, v_1, u_2, x_4, x_3, u_1, x_2 and x_1, respectively. Their expected frequencies p_l^{Fz} ($l=1,\ldots,9$) are also listed in Table 3.

Relationship between parameters in the above two models. The expected frequencies of one genotype under the liability and fitness models should be the same, that is, $p_l^{Lz}=p_l^{Fz}$ ($l=1,\ldots,9$). Therefore, the relationship between

Table 3 Expected frequencies of the nine genotypes under zygotic and gametic selection in the F_2 population

Genotype	*Relative fitness* f_l	*Zygotic selection*		*Gametic selection*	
		p_l^{Fz} *in fitness model*	p_l^{Lz} *in liability model*	p_l^{Fg} *in fitness model*	p_l^{Lg} *in liability model*
$S_1S_1S_2S_2$	$\Phi(a_1+a_2+i)$	$(1-r^2)/D_1$	$(1-r)^2\ f_1/d_1$	$(1-r)^2/D_2$	$(1-r)^2\ f_1/d_2$
$S_1S_1S_2s_2$	$\Phi(a_1+d_2+j_{12})$	$2r(1-r)v_2/D_1$	$2r(1-r)f_2/d_1$	$r(1-r)(v_g+1)/D_2$	$r(1-r)(f_1+f_3)/d_2$
$S_1S_1s_2s_2$	$\Phi(a_1-a_2-i)$	r^2v_1/D_1	r^2f_3/d_1	r^2v_g/D_2	r^2f_3/d_2
$S_1s_1S_2S_2$	$\Phi(d_1+a_2+j_{21})$	$2r(1-r)u_2/D_1$	$2r(1-r)f_4/d_1$	$r(1-r)(u_g+1)/D_2$	$r(1-r)(f_1+f_2)/d_2$
$S_1s_1S_2s_2$	$\Phi(d_1+d_2+l)$	$2(1-2r+2r^2)x_4/D_1$	$2(1-2r+2r^2)f_5/d_1$	$[(1-r)^2(x_g+1)+r^2(u_g+v_g)]/D_2$	$[(1-r)^2(f_1+f_9)+r^2(f_3+f_7)]/d_2$
$S_1s_1s_2s_2$	$\Phi(d_1-a_2-j_{21})$	$2r(1-r)x_3/D_1$	$2r(1-r)f_6/d_1$	$r(1-r)(v_g+x_g)/D_2$	$r(1-r)(f_3+f_7)/d_2$
$s_1s_1S_2S_2$	$\Phi(-a_1+a_2-i)$	r^2u_1/D_1	r^2f_7/d_1	r^2u_g/D_2	r^2f_7/d_2
$s_1s_1S_2s_2$	$\Phi(-a_1+d_2-j_{12})$	$2r(1-r)x_2/D_1$	$2r(1-r)f_8/d_1$	$r(1-r)(u_g+x_g)/D_2$	$r(1-r)(f_7+f_9)/d_2$
$s_1s_1s_2s_2$	$\Phi(-a_1-a_2+i)$	$(1-r)^2x_1/D_1$	$(1-r)^2f_9/d_1$	$(1-r)^2x_g/D_2$	$(1-r)^2\ f_9/d_2$

$D_1=(1-r)^2(x_1+1)+2r(1-r)(u_2+v_2+x_2+x_3)+r^2(u_1+v_1)+2(1-2r+2r^2)x_4$.
$D_2=2(1-r)(x_g+1)+2r(u_g+v_g)$.
$d_1=(1-r)^2(f_1+f_9)+2r(1-r)(f_2+f_4+f_6+f_8)+r^2\ f_3+2(1-2r+2r^2)f_5$.
$d_2=2(1-r)(f_1+f_9)+2r(f_3+f_7)$.

parameters in the two models can be expressed as

$$\begin{bmatrix} u_1 & u_2 & v_1 & v_2 \\ x_1 & x_2 & x_3 & x_4 \end{bmatrix} = \begin{bmatrix} \frac{\Phi(-a_1+a_2-i_{12})}{\Phi(a_1+a_2+i_{12})} & \frac{\Phi(d_1+a_2+j_{21})}{\Phi(a_1+a_2+i_{12})} & \frac{\Phi(a_1-a_2-i_{12})}{\Phi(a_1+a_2+i_{12})} & \frac{\Phi(a_1+d_2+j_{12})}{\Phi(a_1+a_2+i_{12})} \\ \frac{\Phi(-a_1-a_2+i_{12})}{\Phi(a_1+a_2+i_{12})} & \frac{\Phi(-a_1+d_2-j_{12})}{\Phi(a_1+a_2+i_{12})} & \frac{\Phi(d_1-a_2-j_{21})}{\Phi(a_1+a_2+i_{12})} & \frac{\Phi(d_1+d_2+l_{12})}{\Phi(a_1+a_2+i_{12})} \end{bmatrix} \quad (12)$$

Genetic model under gametic selection in the F_2 population

Liability model. The genetic model under gametic selection is the same as model (11). Assume that female gametes and their mating processes are normal, that is, the frequencies of female gametes S_1S_2, S_1s_2, s_1S_2 and s_1s_2 are $(1-r)/2$, $r/2$, $r/2$ and $(1-r)/2$, respectively. If the marker segregation ratio shows deviation from the Mendelian ratio, the distortion is derived from the male gametes of an F_1 plant. Note that the frequencies of the nine genotypes under gamete selection are same as those under zygotic selection and that genotypes $S_1S_1S_2S_2$, $S_1S_1s_2s_2$, $s_1s_1S_2S_2$ and $s_1s_1s_2s_2$ are uniquely derived from the crosses S_1S_2/S_1S_2, S_1s_2/S_1s_2, S_1S_2/s_1S_2 and s_1s_2/s_1s_2, respectively. Thus, the frequencies of male gametes S_1S_2, S_1s_2, s_1S_2 and s_1s_2 are $2(1-r)f_1/d_1$, $2rf_3/d_1$, $2rf_7/d_1$ and $2(1-r)f_9/d_1$, respectively, and the expected frequencies p_l^{Lg} of the nine genotypes in F_2 can be calculated as in Supplementary Table S1 (listed in Table 3). If we compare columns 4 and 6 in Table 3, the following equations can be found: $2f_2=f_1+f_3$, $2f_4=f_1+f_7$, $2f_6=f_3+f_9$, $2f_8=f_7+f_9$ and $2f_5=f_1+f_9=f_3+f_7$.

Fitness model. Let the viabilities of male gametes S_1s_2, s_1S_2 and s_1s_2 relative to S_1S_2 be v_g, u_g and x_g, respectively. The expected frequencies p_l^{Fg} of the nine genotypes under gametic selection are listed in Table 3.

Relationship between parameters in the above two models. The expected frequencies of one genotype under the liability and fitness models should be the same, that is, $p_l^{Lg}=p_l^{Fg}$ $(l=1,\ldots,9)$. The relationship between parameters in the two models can be expressed as

$$\begin{bmatrix} u_g & v_g & x_g \end{bmatrix} = \begin{bmatrix} \frac{\Phi(-a_1+a_2-i)}{\Phi(a_1+a_2+i)} & \frac{\Phi(a_1-a_2-i)}{\Phi(a_1+a_2+i)} & \frac{\Phi(-a_1-a_2+i)}{\Phi(a_1+a_2+i)} \end{bmatrix} \quad (13)$$

which is the same as equation (4) in the backcross. In fact, this relationship is reasonable. Under the situation of gametic selection, selection occurs during the gamete production stage but not the mating process. As we know, these gametes are formed in the F_1 plant stage, which is similar to a backcross.

Likelihood function and parameter estimation in the *F_2 population*

If p_l and n_l $(l=1,\ldots,9)$ are the expected frequencies and observed number of the kth genotype for the two SDL, and $n=\sum_{l=1}^{9} n_l$ is the total number of individuals, then the general likelihood function in F_2 can be expressed as

$$L=\frac{n!}{\prod_l n_l!}\prod_l p_l^{n_l} \quad (14)$$

where p_l is p_l^{Fg} under gametic selection or p_l^{Fz} under zygotic selection.

Parameter estimation under zygotic selection. The genotypes of an SDL are unobserved if the SDL does not reside at the marker position. As described in a backcross, the information for four markers flanking with the two SDL can be used to estimate all of the parameters. However, there are 4096 (64×64) gamete combinations and 729 genotypes for four markers and two SDL. Using this calculation, it is time consuming to estimate the parameters. To reduce the running time, the information for three markers (A, B and C) flanking with the two SDL (S1 and S2) is utilised. The order of these loci are A, S_1, B, S_2 and C, and the recombination fractions for the four linked intervals are r_1, r_2, r_3 and r_4, respectively. There are 27 genotypes (observed) for three markers and nine genotypes (unobserved) for the two SDL. Thus, the complete information likelihood function is

$$L=\frac{n!}{\prod_{k,l} n_{kl}!}\prod_{k,l} p_{kl}^{n_{kl}} \quad (15)$$

where p_{kl} and n_{kl} $(k=1,\ldots,27;\ l=1,\ldots,9)$ are the expected frequencies and observed numbers of the kth marker genotype and the lth SDL genotype, respectively. The logarithm likelihood function and the partial derivative for each parameter are given in Supplementary Material A. The estimations of the parameters are

$$\begin{aligned} r_i &= \frac{T_{r_i}}{2n}-\frac{r_i(1-r_i)}{2d}\times\frac{\partial d}{\partial r_i} \\ u_j &= \frac{dT_{u_j}}{n}\times\left(\frac{\partial d}{\partial u_j}\right)^{-1} \\ v_j &= \frac{dT_{v_j}}{n}\times\left(\frac{\partial d}{\partial v_j}\right)^{-1} \\ x_i &= \frac{dT_{x_i}}{n}\times\left(\frac{\partial d}{\partial x_i}\right)^{-1} \end{aligned} \quad (16)$$

where d, T_{1-r_i}, T_{r_i}, T_{v_j}, T_{u_j} and T_{x_i} $(i=1,2,3,4;\ j=1,2)$ are listed in Supplementary File zygotic.xls. The estimates for r_1 and r_2 are used to estimate the recombination fraction between markers A and B: $r_{AB}=r_1+r_2-2r_1r_2$, which is the corrected recombinant fraction. When m markers are located in a linkage group, the number of estimates for r_{AB} is $m-2$. Similarly, the median is the suggested estimate.

Parameter estimation under gametic selection. Four parameters, r, u_g, v_g and x_g, under gametic selection need to be estimated. The procedures and algorithm for the parameter estimation are similar to those under zygotic

selection. Similarly, we obtain

$$\begin{aligned} r_i &= \frac{T_{r_i}}{2n} - \frac{r_i(1-r_i)}{2d} \times \frac{\partial d}{\partial r_i} \\ u_g &= \frac{dT_{u_g}}{n} \times \left(\frac{\partial d}{\partial u_g}\right)^{-1} \\ v_g &= \frac{dT_{v_g}}{n} \times \left(\frac{\partial d}{\partial v_g}\right)^{-1} \\ x_g &= \frac{dT_{x_g}}{n} \times \left(\frac{\partial d}{\partial x_g}\right)^{-1} \end{aligned} \tag{17}$$

where d, T_{r_i} $(i=1,2,3,4)$, T_{v_g}, T_{u_g} and T_{x_g} are listed in Supplementary File gametic.xls. The strategy for estimate of r is the same as that under zygotic selection.

Variance of recombination fraction. The variances of recombination fraction r under gametic and zygotic selection in the F_2 population are

$$Var(r_g) = \frac{r(1-r)\left[(1-r)(x_g+1)+r(u_g+v_g)\right]^2}{n\left[(u_g+v_g+x_g+1)(ru_g+rv_g-rx_g-r+x_g+1)-r(1-r)(u_g+v_g-x_g-1)^2\right]}$$

$$Var(r_z) = \frac{D}{n}\left[4(u_1+v_1+x_1+1)+(u_2+v_2+x_2+x_3)\frac{2(2r-1)^2}{r(1-r)}+\frac{8x_4(2r-1)^2}{2r^2-2r+1}-\frac{K^2}{D}\right]^{-1}$$

respectively, where $K=2r(u_1+v_1+x_1+1)+2(2r-1)(2x_4-u_2-v_2-x_2-x_3)-2(x_1+1)$ and $D=(1-r)^2(x+1)+2r(1-r)(u_2+v_2+x_2+x_3)+r^2(u_1+v_1)+2(1-2r+2r^2)x_4$.

Detection of selection type in the F_2 population

The χ^2-test of Pham *et al.* (1990) is used to determine whether the numbers of *AA*, *Aa* and *aa* genotypes in F_2, n_{AA}, n_{Aa} and n_{aa} follow the Mendelian segregation ratio of 1: 2: 1

$$\chi^2 = \frac{4n_{AA}^2+2n_{Aa}^2+4n_{aa}^2}{n} - n \tag{18}$$

If the test is significant, selection exists. To further clarify the selection type (that is, gametic vs zygotic), Lorieux *et al.* (1995b) suggest two χ^2 tests,

$$\chi_1^2 = \frac{(2n\hat{p}-n)^2+(2n\hat{q}-n)^2}{2n} \tag{19}$$

$$\chi_2^2 = \frac{(n_{AA}-n\hat{p}^2)^2}{n\hat{p}^2} + \frac{(n_{Aa}-2n\hat{p}\hat{q})^2}{2n\hat{p}\hat{q}} + \frac{(n_{aa}-n\hat{q}^2)^2}{n\hat{q}^2} \tag{20}$$

where $\hat{p}$ is the allelic frequency of *A* in F_2. Gametic selection occurs if χ_1^2 but not χ_2^2 is significant; zygotic selection occurs if χ_2^2 is significant (Lorieux *et al.*, 1995b).

Statistical properties

At present, there are three approaches available. The first is the method that does not consider the effect of distorted markers, named method I, implemented by MapMaker v3.0 (Lander *et al.*, 1987) or JoinMap v4.0. The second is the method that considers the effect of distorted markers, named method II (Lorieux *et al.*, 1995a, b; Zhu *et al.*, 2007). The third is the new method described in this study, which considers the effect of epistatic distorted markers. Compared with methods I and II, some properties of the new method in a backcross population are summarised below:

(a) The new method is equal to method I when $u=v=x=1$, and the new method is equal to method II when $u\neq 1$, $v\neq 1$ and $x=1$. This finding means that the new method is general and that methods I and II are specific.
(b) An unbiased estimate for the recombinant fraction can be obtained when $x+1=u+v$ or $f_1^B+f_4^B=f_2^B+f_3^B$ for method I; $x=uv$ or $f_1^Bf_4^B=f_2^Bf_3^B$ for method II; and for all situations for the new method.
(c) The overestimate for the recombinant fraction occurs when $f_2^B+f_3^B>f_1^B+f_4^B$ or $u+v>x+1$ for method I and $f_2^Bf_3^B>f_1^Bf_4^B$ or $uv>x$ for method II. The underestimate for the recombinant fraction occurs when $f_2^B+f_3^B<f_1^B+f_4^B$ or $u+v<x+1$ for method I and $f_2^Bf_3^B<f_1^Bf_4^B$ or $uv<x$ for method II.
(d) Two linked SDL affect the estimates of the recombinant fraction for all marker intervals within the two linked SDL. The evidence is shown in Supplementary Material B.
(e) The variance of recombinant fraction r for the new method is equal to and less than that for method I when $u+v=x+1$ and $u+v<x+1$, respectively. If $u+v>x+1$, two situations occur: the variance of r for the new method is greater and less than that for method I when $\hat{r}>\frac{\sqrt{x+1}}{\sqrt{u+v}+\sqrt{x+1}}$ and $\hat{r}<\frac{\sqrt{x+1}}{\sqrt{u+v}+\sqrt{x+1}}$, respectively. The evidence is shown in Supplementary Material C.

RESULTS

Monte Carlo simulation

Effect of heritability, marker interval length and sample size on the estimate of map distance. Nine equally spaced markers were simulated on a single-chromosome segment in a backcross population. Two SDL were placed at the middle of the second and seventh marker intervals. Three levels were set up for each factor in Monte Carlo simulated experiments: (1) SDL heritability, 2, 5 and 10%; (2) interval length between adjacent markers, 5, 10 and 15 cM; and (3) sample size, 100, 200 and 300. All of the simulated parameters are shown in Supplementary Table S2. For each parameter combination, 200 replicated experiments were conducted, and the absolute bias and s.d. among the estimates from the 200 replicates were used to estimate the precision. All of the results for the backcross population are listed in Figures 1 and 2. The results showed that all of the estimates from the new method were unbiased (Figure 1). The two linked SDL do not affect the estimates of the map distances for marker intervals 1 and 8 (outside the two SDL) but do affect the estimates of the map distances for marker intervals 2–7 (within the two SDL) when methods I and II are adopted (Figure 1). In addition, the absolute bias and s.d. of the new method increase as the SDL heritability and marker interval length increase, and they decrease as the sample size increases (Figures 1 and 2). Similar results are also observed in F_2 (Supplementary Figures S1 and S2).

Effect of the SDL genetic model on linkage map construction. Eight genetic modes of SDL, listed in Figure 3 and Supplementary Table S3, were set up to investigate the effect of the SDL genetic model on the map distance under the liability and fitness models. The sample size was 300, and the marker interval length was 10 cM. The other parameters were the same as those in the above simulated experiment. All the results in the backcross are listed in Figure 3. The results were as follows: (1) all the estimates from the new method were unbiased. (2) Using method I, the estimates under SDL genetic modes 5–8 were unbiased because $f_2^b+f_3^b=f_1^b+f_4^b$ and $u+v=x+1$ (Figures 3e–h, and Supplementary Table S3); the estimates under SDL genetic modes 1 and 3 were underestimated because $f_2^b+f_3^b<f_1^b+f_4^b$ and $u+v<x+1$ (Figures 3a and c, and Supplementary Table S3); and the estimates under SDL genetic modes 2 and 4 were overestimated because $f_2^b+f_3^b>f_1^b+f_4^b$ and $u+v>x+1$ (Figures 3b and d, and Supplementary Table S3). Using method II, the estimates under SDL genetic modes 7 and 8 were unbiased because $f_2^bf_3^b=f_1^bf_4^b$ and $uv=x$ (Figures 3g and h, and Supplementary Table S3); the estimates under SDL genetic modes 1, 3 and 5 were underestimated because $f_2^bf_3^b<f_1^bf_4^b$ and $uv>x$ (Figures 3a, c and e, and Supplementary Table S3); and the estimates under SDL genetic modes 2, 4 and 6 were overestimated because $f_2^bf_3^b>f_1^bf_4^b$ and $uv>x$ (Figures 3b, d and f, and Supplementary Table S3). (3) The bias was proportional to the above related size difference. For example, the bias of the estimates from method I in Figure 3d is larger than that in Figure 3b because

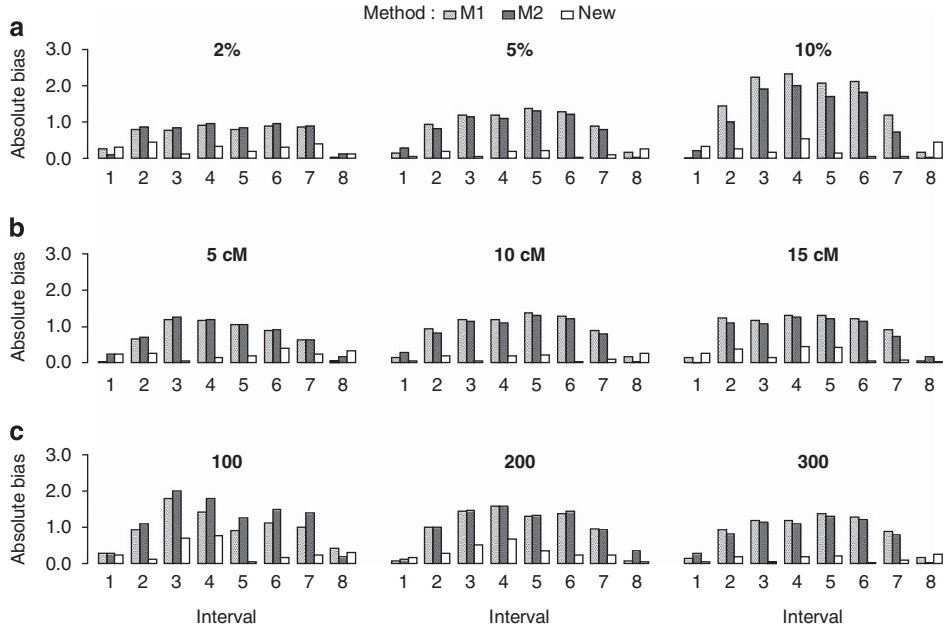

Figure 1 Effect of SDL heritability (**a**), interval length (**b**) and sample size (**c**) on the estimate of map distance in a backcross population. (**a**) Interval length, 10 cM; sample size, 300; (**b**) SDL heritability, 5%; sample size, 300; and (**c**) SDL heritability, 5%; interval length, 10 cM.

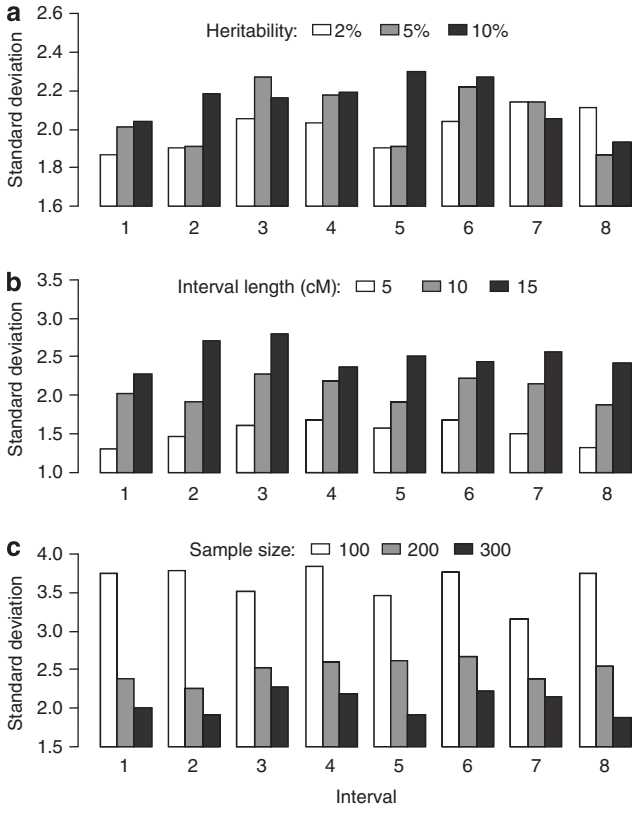

Figure 2 Effect of SDL heritability (**a**), interval length (**b**) and sample size (**c**) on the s.d. of the estimates from the new method in a backcross population. (**a**) Interval length, 10 cM; sample size, 300; (**b**) SDL heritability, 5%; sample size, 300; and (**c**) SDL heritability, 5%; interval length, 10 cM.

$(f_2^b + f_3^b) - (f_1^b + f_4^b) = 0.76$ in Figure 3d is larger than 0.62 in Figure 3b.

Effect of selection type in the F_2 population on linkage map construction. Gametic and zygotic selections of SDL in the F_2 population were simulated to investigate the effect of the selection type on map distance. SDL heritability was set at 5%, sample size was 300 and marker interval length was 10 cM. The other parameters were the same as those in the first simulated experiment. All of the data sets were first analysed by the χ^2-test to determine the selection type. The results are listed in Table 4 and are consistent with the theoretical results. Each data set was then analysed twice: once under gametic selection and once under zygotic selection. The purpose was to determine which method was better in the case of inconsistent selection types of adjacent markers. The results are listed in Table 5. The results showed that new method works well under consistent selection types of adjacent markers. If gametic selection occurs at the ith locus and zygotic selection occurs at the $(i+1)$th locus, how to select the method for parameter estimation was unclear. As a result, the absolute bias under zygotic selection is less than that under gametic selection. Therefore, we recommend the zygotic selection approach to address this case.

Real data analysis in rice

To further demonstrate the new method, a real data set for a backcross population (*Oryza sativa/Oryza longistaminata//O. sativa*) (Causse *et al.*, 1994) was downloaded from the McCouch RiceLab website (http://ricelab.plbr.cornell.edu/Causse_at_al_1994) and re-analysed. The data set is composed of 617 markers on 12 chromosomes. On the basis of 12 linkage groups constructed by Mapdisto v1.7.7 (Lorieux, 2012), all of the map distances between flanking markers were corrected by software package DistortedMap of the new method (Supplementary file DistortedMap). All of the results are listed in Supplementary Table S4 and Supplementary Figure S3.

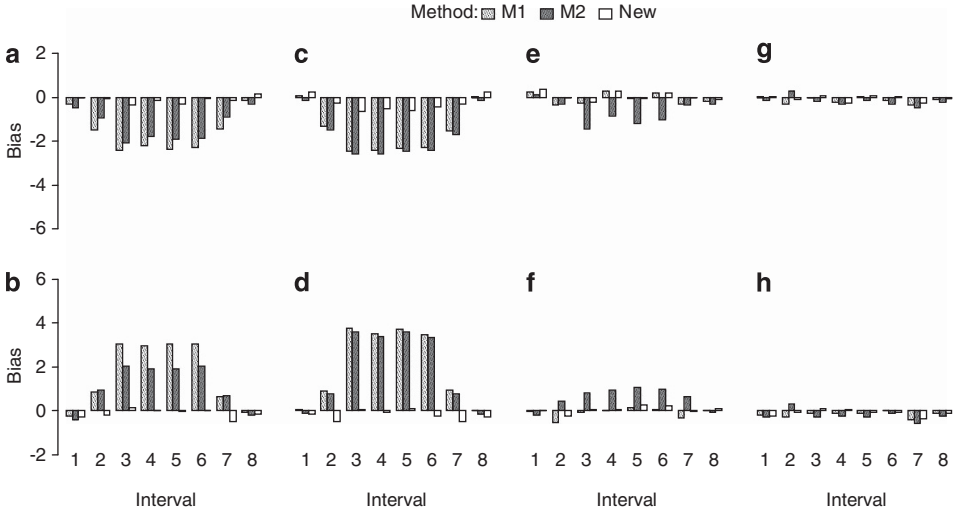

Figure 3 Effect of SDL genetic mode on the estimate of map distance in a backcross population. The SDL parameters are as follows: (**a**) $a_1 = a_2 = i = 0.5$; (**b**) $a_1 = -a_2 = -i = 0.5$; (**c**) $a_1 = a_2 = 0, i = 0.5$; (**d**) $a_1 = a_2 = 0, i = -0.5$; (**e**) $a_1 = -a_2 = 0, i = 0.5$; (**f**) $a_1 = a_2 = 0.5, i = 0$; (**g**) $a_1 = -0.5, a_2 = i = 0$; and (**h**) $a_1 = 0.5, a_2 = i = 0$. The relationship among the parameters in the liability and fitness models is shown in Supplementary Table S3: $f_2^b + f_3^b < f_1^b + f_4^b$ and $u + v < x + 1$ (**a**, **c**); $f_2^b + f_3^b > f_1^b + f_4^b$ and $u + v > x + 1$ (**b**, **d**); $f_2^b + f_3^b = f_1^b + f_4^b$ and $u + v = x + 1$ (**e**–**h**); $f_2^b f_3^b < f_1^b f_4^b$ and $uv < x$ (**a**, **c**, **e**); $f_2^b f_3^b > f_1^b f_4^b$ and $uv > x$ (**b**, **d**, **f**); $f_2^b f_3^b = f_1^b f_1^b$ and $uv = x$ (**g** and **h**).

Table 4 χ^2-tests for marker segregation distortion, and gametic and zygotic selection

Marker	*χ^2 for marker segregation distortion*		*χ_1^2 for gametic selection*		*χ_2^2 for zygotic selection*	
	Gametic selection	*Zygotic selection*	*Gametic selection*	*Zygotic selection*	*Gametic selection*	*Zygotic selection*
1	11.65**(6.35)	13.65**(6.29)	5.31* (3.10)	5.78*(3.05)	1.10(1.47)	2.99(3.41)
2	16.57**(7.32)	21.83**(7.54)	7.78**(3.59)	8.77**(3.79)	1.25(1.63)	6.71*(4.62)
3	18.93**(8.48)	23.60**(7.76)	8.93**(4.03)	9.85**(3.83)	1.36(1.65)	6.59*(4.63)
4	16.59**(7.78)	19.54**(7.68)	7.80**(3.74)	8.69**(3.71)	1.15(1.43)	3.67(3.92)
5	16.02**(7.47)	18.54**(7.34)	7.55**(3.59)	8.30**(3.64)	1.06(1.41)	3.18(3.51)
6	16.03**(6.88)	19.33**(7.21)	7.50**(3.33)	8.60**(3.69)	1.24(1.58)	3.57(3.33)
7	18.58**(7.79)	23.53**(7.27)	8.77**(3.72)	9.76**(3.69)	1.27(1.68)	6.62*(4.65)
8	16.83**(7.74)	21.14**(7.13)	7.90**(3.77)	8.49**(3.47)	1.18(1.56)	6.51*(4.55)
9	11.95**(6.39)	13.61**(5.93)	5.43*(3.15)	5.65*(2.81)	1.14(1.80)	3.25(3.15)

* and **: significances at the 0.05 and 0.01 levels, respectively. The s.d. among 200 replicates are in parentheses.

Table 5 Comparison of the gametic and zygotic model methods under gametic and zygotic selection in the F_2 population

Selection type	*Method*	*Marker interval*							
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Gametic	gm	9.96(1.34)	10.08(1.29)	9.88(1.31)	9.79(1.26)	9.93(1.33)	9.88(1.33)	9.97(1.33)	9.90(1.39)
	zm	9.92(1.33)	9.59(1.18)	9.41(1.20)	9.33(1.16)	9.47(1.22)	9.41(1.21)	9.49(1.21)	9.85(1.38)
Zygotic	gm	10.35(1.49)	12.67(2.10)	11.83(1.99)	11.94(2.22)	12.13(1.87)	11.88(1.89)	12.81(2.26)	10.25(1.54)
	zm	10.05(1.41)	10.01(1.39)	9.50(1.30)	9.59(1.46)	9.71(1.25)	9.50(1.27)	10.08(1.52)	9.96(1.47)

Abbreviations: gm, gametic model method, zm: zygotic model method. Marker interval length, 10 cM.

To further illustrate the new method, all of the map distances on chromosome 3, with several severely distorted segregation regions, were calculated by methods I, II and new (Table 6). As a result, in regions with normal Mendelian segregation, the estimates of the recombinant fraction by the above three methods were similar, such as for markers CDO375, RCH and RZ696, and almost all the estimates for u, v and x were closer to 1 than those in regions with distorted segregation. In the distorted segregation region between markers RZ585 and RZ284, the χ^2-test for marker RZ284 was significant ($\chi^2 = 18.60$, $P = 1.61e-5$), and the map distances of 4.75 and 5.29 cM, calculated by methods I and II, respectively, were less than 6.52 cM, calculated by the new method, indicating that the results from methods I and II were underestimates because $u + v = 0.93 < x + 1 = 1.30$ for method I and $uv = 0.19 < x = 0.30$ for method II.

Table 6 Comparison of the map distances on chromosome 3 from methods I, II and the new method in a rice data analysis

Marker	χ^2	*Selection*	*Map distances (cM) from various methods*			*Segregation distortion loci effect*		
			I	*II*	*New*	*u*	*v*	*x*
RZ497	3.57	No	—	—	—	—	—	—
RZ25	10.89**	Gametic/zygotic (g/z)	1.14	0.15	0.15	11.45	0.05	0.53
RZ22	11.56**	g/z	0.00	0.00	0.00	0.71	0.71	0.50
RZ18	5.38*	g/z	1.17	0.17	1.26	0.01	1.43	0.55
CDO375	0.29	No	0.00	0.00	0.00	0.84	0.84	0.71
RCH	1.90	No	8.95	8.99	8.95	0.91	0.91	0.82
RZ696	1.08	No	9.76	9.81	9.76	0.73	0.97	0.69
RZ394	4.57*	g/z	2.80	0.68	0.68	7.07	0.10	0.71
RZ452	3.85*	g/z	0.00	0.00	0.00	0.80	0.80	0.65
RZ251	3.32	No	1.22	0.23	0.23	9.34	0.08	0.73
RZ585	8.89**	g/z	2.80	0.49	0.49	8.58	0.06	0.49
RZ284	18.6**	g/z	4.75	5.29	6.52	0.62	0.31	0.30
RZ672	12.49**	g/z	3.07	3.48	4.70	0.44	0.43	0.36
CDO938	8.05**	g/z	3.10	3.26	5.26	0.44	0.44	0.52
RG745	8.78**	g/z	0.00	0.00	0.00	0.75	0.75	0.56
RZ574	6.37*	g/z	0.00	0.00	0.00	0.79	0.79	0.63
RZ1000	12.52**	g/z	1.65	0.28	2.44	1.04	0.01	0.56
CDO608	4.74*	g/z	2.03	0.53	2.46	0.02	1.32	0.63
RG227	3.08	No	0.00	0.00	0.00	0.77	0.77	0.59
RZ678	11.12**	g/z	0.00	0.00	0.00	0.85	0.85	0.72
BCD734	11.56**	g/z	0.00	0.00	0.00	0.72	0.72	0.52
BCD1092	8.38**	g/z	0.00	0.00	0.00	0.72	0.72	0.52
CD1053X	13.79**	g/z	0.00	0.00	0.00	0.73	0.73	0.53
RZ399	7.19**	g/z	0.00	0.00	0.00	0.73	0.73	0.53
RZ517	12.3**	g/z	0.00	0.00	0.00	0.74	0.74	0.55
RZ16	10.12**	g/z	0.00	0.00	0.00	0.70	0.70	0.49
CDO260	16.64**	g/z	0.00	0.00	0.00	0.69	0.69	0.48
CDO33X	12.96**	g/z	0.00	0.00	0.00	0.67	0.67	0.44
RG191	9.24**	g/z	1.24	0.36	2.28	0.80	0.02	0.52
RG224	3.96*	g/z	0.00	0.00	0.00	0.78	0.78	0.61
CD1387B	12.96**	g/z	0.00	0.00	0.00	0.85	0.85	0.72
CDO1395	13.5**	g/z	0.00	0.00	0.00	0.68	0.68	0.46
RZ313	11.71**	g/z	0.00	0.00	0.00	0.69	0.69	0.48
RG450	15.36**	g/z	0.00	0.00	0.00	0.69	0.69	0.47
RG369A	11.33**	g/z	3.30	3.34	4.63	0.34	0.69	0.47
RG100	10.13**	g/z	2.65	0.50	3.95	0.99	0.01	0.51
RZ545	4.76*	g/z	0.00	0.00	0.00	0.72	0.72	0.51
RZ742B	5.45*	g/z	4.66	4.54	5.16	0.47	0.96	0.59
CDO1069	14.44**	g/z	3.38	0.34	5.29	0.91	0.00	0.46
RZ993X	12.63**	g/z	2.29	2.49	3.52	0.46	0.45	0.43
RZ891	7.51**	g/z	2.29	0.30	3.03	0.00	1.13	0.51
RZ987	5.76*	g/z	2.52	0.45	3.10	1.30	0.01	0.62
RZ329	8.49**	g/z	2.49	2.57	2.56	0.77	0.79	0.60
RG944	8.38**	g/z	7.57	7.55	8.00	0.48	0.95	0.51
RG348	9.91**	g/z	6.83	5.93	6.91	1.16	0.29	0.47
RG104	2.85	No	5.69	5.11	5.11	0.46	1.36	0.63
CDO20	14.82**	g/z	7.64	6.92	6.92	1.31	0.44	0.58
CDO481	5.31*	g/z	22.82	23.62	25.97	0.37	0.82	0.37
RG432	2.47	No	27.10	18.80	18.80	0.37	2.56	0.96

* and **: significances at the 0.05 and 0.01 levels, respectively.

DISCUSSION

Although the new method proposed for linkage map correction in this study is based on the epistatic genetic model of SDL, it is suitable not only for the above model but also for normal (Supplementary Table S5) and distorted (Figure 3) markers. When no SDL is identified in a linkage group, the estimates for map distances by the above three approaches are almost unbiased and close to the true values (Supplementary Table S5). We also calculated the s.d. of the estimates by two approaches: one using Fisher's information (SD1) and the other using the variation of the estimates across 200 replicates

(SD2). For the former, similar results were observed across the above three methods; for the latter, slightly increased results were found from method I to II and from method II to the new method (Supplementary Table S5). These findings are reasonable because the number of parameters gradually increased in the above three methods, and the accumulated errors from their corresponding estimates were also increased. If multi-SDL are considered in a linkage group, the corrected results for the genetic distance are more accurate using the new method than using methods I and II (Supplementary Table S5). However, SD1 and SD2 are slightly larger for the new method than for methods I and II. The theoretical evidence is given in Supplementary Material C.

With respect to statistical properties 2 and 3 in the backcross, the evidence exists. Using method I, the recombinant fraction is estimated by $\frac{n_2+n_3}{n}$. If the expectations of n_2, n_3 and n in the liability and fitness models are used to estimate n_2, n_3 and n, respectively, the two properties can be demonstrated. In the liability model, $\hat{r}=\frac{r(f_2^b+f_3^b)}{r(f_2^b+f_3^b)+(1-r)(f_1^b+f_4^b)}$. If $f_2^b+f_3^b=f_1^b+f_4^b$, then $\hat{r}=r$, which is an unbiased estimate; if $f_2^b+f_3^b>f_1^b+f_4^b$, then $\hat{r}>r$, an overestimate; and if $f_2^b+f_3^b<f_1^b+f_4^b$, then $\hat{r}<r$, an underestimate. In the fitness model, $\hat{r}=\frac{r(u+v)}{r(u+v)+(1-r)(x+1)}$. By using a similar approach, the statistical properties 2 and 3 can be obtained. Using method II, the recombinant fraction is estimated by $\frac{\sqrt{n_2n_3}}{\sqrt{n_2n_3}+\sqrt{n_1n_4}}$ (Lorieux *et al.*, 1995a). In the fitness model, the estimate is changed to $\hat{r}=\frac{r\sqrt{uv}}{r\sqrt{uv}+(1-r)\sqrt{x}}$. If $uv=x$, then $\hat{r}=r$, which is an unbiased estimate; if $uv>x$, then $\hat{r}>r$, an overestimate; and if $uv<x$, then $\hat{r}<r$, an underestimate. In the liability model, $\hat{r}=\frac{r\sqrt{f_2^bf_3^b}}{r\sqrt{f_2^bf_3^b}+(1-r)\sqrt{f_1^bf_4^b}}$. By using a similar approach, the statistical properties 2 and 3 can also be obtained. These properties have been confirmed by the Monte Carlo simulation studies and real data analysis in this study.

If two SDL of one SDL-by-SDL interaction are placed in the same linkage group, this interaction does not affect the estimate of the recombinant fraction outside the SDL intervals. In Supplementary Material B, the estimates for the recombinant fractions outside the SDL intervals are obtained as $R_1=(n_{Ab}+n_{aB})$, $R_2=(n_{Bc}+n_{bC})/n$, $R_3=(n_{De}+n_{dE})/n$ and $R_4=(n_{Ef}+n_{eF})/n$. Obviously, the four estimates are independent of the viability parameters, which are evidence of the above result. Similar evidence was also observed in the Monte Carlo simulation studies. If two SDL of one SDL-by-SDL interaction are placed in different linkage groups, this interaction does not affect the estimates of the recombinant fraction. In Supplementary Material B, the estimates for the two recombinant fractions involved this interaction are obtained as $r_1=\sum_{i=5}^{12}n_i/.n$ and $r_2=\sum_{i=2}^{3}(n_i+n_{i+4}+n_{i+8}+n_{i+12})/.n$. Obviously, the two estimates for both r_1 and r_2 are independent of the viability parameters, representing evidence of the above result. In addition, the results in Figures 2g and h showed that the estimate for the recombination fraction is unaffected by the distorted markers due to only one SDL in one linkage group. This finding is consistent with the previous results in Bailey (1949), Lorieux *et al.* (1995a,b) and Zhu *et al.* (2007).

In linkage group construction, some multi-point approaches are widely used. Among these approaches, Lander and Green (1987) first proposed a Markov chain multi-point approach to utilise missing markers. Jiang and Zeng (1997) then extended the method of Lander and Green (1987) to address dominant and missing markers, and Zhu *et al.* (2007) further extended the multi-point method to address distorted, dominant and missing markers. In this study, epistatic distorted markers are also considered as well. In fact, once the transition probability matrix $H(r)$ from markers A to B is determined, the multi-point method including epistatic distorted markers should work well. Here, we provide these matrices as follows:

$$H_{BC}(r)=\begin{bmatrix}\frac{1-r}{1-r+rv} & \frac{rv}{1-r+v}\\ \frac{ru}{(1-r)x+ru} & \frac{(1-r)x}{(1-r)x+ru}\end{bmatrix}$$

for a backcross population,

$$H^g_{F_2}(r)=\begin{bmatrix}\frac{(1-r)^2}{(1-r)+rv_g} & \frac{r(1-r)(v_g+1)}{(1-r)+rv_g} & \frac{r^2v_g}{(1-r)+rv_g}\\ \frac{r(1-r)(u_g+1)}{(1-r)(x_g+1)+r(u_g+v_g)} & \frac{(1-r)^2(x_g+1)+r^2(u_g+v_g)}{(1-r)(x_g+1)+r(u_g+v_g)} & \frac{r(1-r)(v_g+x_g)}{(1-r)(x_g+1)+r(u_g+v_g)}\\ \frac{r^2u_g}{(1-r)x_g+ru_g} & \frac{r(1-r)(x_g+u_g)}{(1-r)x_g+ru_g} & \frac{(1-r)^2x_g}{(1-r)x_g+ru_g}\end{bmatrix}$$

for the gametic selection approach in F_2, and

$$H^z_{F_2}(r)=\begin{bmatrix}\frac{(1-r)^2}{(1-r)^2+2r(1-r)v_2+r^2v_1} & \frac{2r(1-r)v_2}{(1-r)^2+2r(1-r)v_2+r^2v_1} & \frac{r^2v_1}{(1-r)^2+2r(1-r)v_2+r^2v_1}\\ \frac{r(1-r)u_2}{r(1-r)(u_2+x_3)+(2r^2-2r+1)x_4} & \frac{(2r^2-2r+1)x_4}{r(1-r)(u_2+x_3)+(2r^2-2r+1)x_4} & \frac{r(1-r)x_3}{r(1-r)(u_2+x_3)+(2r^2-2r+1)x_4}\\ \frac{r^2u_1}{(1-r)^2x_1+2r(1-r)x_2+r^2u_1} & \frac{2r(1-r)x_2}{(1-r)^2x_1+2r(1-r)x_2+r^2u_1} & \frac{(1-r)^2x_1}{(1-r)^2x_1+2r(1-r)x_2+r^2u_1}\end{bmatrix}$$

for the zygotic selection approach in F_2.

In animal and plant genetics, epistasis for viability selection has been detected in the studies of Chang and Noor (2010), Li *et al.* (2011) and Kubo *et al.* (2008). Thus, the method in this study should be developed. This method may be extended to additional biparental populations, for example, recombination inbred lines. The new method deals only with recombinant fraction correction, not with linkage group construction. With respect to this construction, the Mapmaker, Mapmanager, Joinmap and Mapdisto programs are available. Once linkage groups have been constructed and distorted markers exist, the new method can be used to correct bias.

DATA ARCHIVING

All simulated datasets are available from the public website: http://jpkc.njau.edu.cn/swtj/show.asp?classid=44&classtype=26. The real dataset can be retrieved from: http://ricelab.plbr.cornell.edu/Causse_at_al_1994.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank Susan R McCouch, Department of Plant Breeding and Genetics, Cornell University, for providing the rice backcross population mapping data. This research was funded by the National Key Basic Research Program of China (2011CB109306), a Specialised Research Fund for the Doctoral Program of Higher Education (20100097110035 and 20120097110023), and a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

Alheit KV, Reif JC, Maurer HP, Hahn V, Weissmann EA, Miedaner T *et al*. (2011). Detection of segregation distortion loci in triticale (x *triticosecale* Wittmack) based on a high-density DArT marker consensus genetic linkage map. *BMC Genomics* **12**: 380.

Bailey NTJ (1949). The estimation of linkage with differential viability, II and III. *Heredity* **3**: 220–228.

Barchi L, Lanteri S, Portis E, Stàgel A, Valè G, Toppino L *et al*. (2010). Segregation distortion and linkage analysis in eggplant (*Solanum melongena* L.). *Genome* **53**: 805–815.

Bomblies K, Lempe J, Epple P, Warthmann N, Lanz C, Dangl JL *et al.* (2007). Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. *PLoS Biol* **5**: 1962–1972.

Brummer EC, Bouton JH, Kochert G (1993). Development of an RFLP map in diploid alfalfa. *Theor Appl Genet* **86**: 329–332.

Causse MA, Fulton TM, Cho YG, Ahn SN, Chunwongse J, Wu K *et al.* (1994). Saturated molecular map of the rice genome based on an interspecific backcross population. *Genetics* **138**: 1251–1274.

Chang AS, Noor MAF (2010). Epistasis modifies the dominance of loci causing hybrid male sterility in the drosophila pseudoobscura species group. *Evolution* **64**: 253–260.

Cloutier S, Cappadocia M, Landry BS (1997). Analysis of RFLP mapping inaccuracy in *Brassica napus L. Theor Appl Genet* **95**: 83–91.

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

Dobzhansky T (1936). Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. *Genetics* **21**: 113–135.

Faris JD, Laddomada B, Gill BS (1998). Molecular mapping of segregation distortion loci in *Aegilops tauschii*. *Genetics* **49**: 319–327.

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.

Kaló P, Endre G, Zimányi L, Csanádi G (2000). Construction of an improved linkage map of diploid alfalfa (*Medicago sativa*). *Theor Appl Genet* **100**: 641–657.

Kreike DD, Stiekema WJ (1997). Reduced recombination and distorted segregation in a Solanum tuberosum (2×) × S.spegazzinii (2×) hybrid. *Genome* **40**: 180–187.

Kubo T, Yamagata Y, Eguchi M, Yoshimura A (2008). A novel epistatic interaction at two loci causing hybrid male sterility in an inter-subspecific cross of rice (*Oryza sative* L.). *Genes Genet. Syst* **83**: 443–453.

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

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

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

Lashermes P, Combes MC, Prakash NS, Trouslot P, Lorieux M, Charrier A (2001). Genetic linkage map of Coffea canephora: Effect of segregation distortion and analysis of recombination rate in male and female meiosis. *Genome* **44**: 589–596.

Leppälä J, Bokma F, Savolainen O (2013). Investigating incipient speciation in *Arabidopsis lyrata* from patterns of transmission ratio distortion. *Genetics* **194**: 697–708.

Li XH, Wang XJ, Wei YL, Brummer EC (2011). Prevalence of segregation distortion in diploid alfalfa and its implications for genetics and breeding applications. *Theor Appl Genet* **123**: 667–679.

Lorieux M (2012). MapDisto: fast and efficient computation of genetic linkage maps. *Mol Breed* **30**: 1231–1235.

Lorieux M, Goffinet B, Perrier X, González-de-león 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, González-de-león D (1995b). Maximum-likelihood models for mapping genetic markers showing segregation distortion. 2. F_2 populations. *Theor Appl Genet* **90**: 81–89.

Lu H, Romero-Severson J, Bernardo R (2002). Chromosomal regions associated with segregation distortion in maize. *Theor Appl Genet* **105**: 622–628.

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

Mangelsdorf PC, Jones DF (1926). The expression of Mendelian factors in the gametophyte of maize. *Genetics* **11**: 423–455.

McCouch SR, Kochert G, Yu Z, Wang Z, Khush GS, Coffman WR *et al.* (1988). Molecular mapping of rice chromosomes. *Theor Appl Genet* **76**: 815–829.

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q *et al.* (2009). Genetic properties of the maize nested association mapping population. *Science* **325**: 737–740.

Muller H (1942). Isolating mechanisms, evolution, and temperature. *Biol Symp* **6**: 71–125.

Paterson AH, Lander ES, Hewitt J, Peterson S, Lincoln SE (1988). Resolution of quantitative traits into mendelian factors by using a complete map of restriction fragment length polymorphism. *Nature* **335**: 721–726.

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.

Phillips PC (2008). Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* **9**: 855–867.

Rick CM (1966). Abortion of male and female gametes in the tomato determined by allelic interaction. *Genetics* **53**: 85–96.

Sandler L, Hiraizumi Y, Sandler I (1959). Meiotic drive in natural populations of *Drosophila melanogaster*. I. the cytogenetic basis of segregation-distortion. *Genetics* **44**: 233–250.

Törjék O, Witucka-Wall H, Meyer RC, Korff M, Kusterer B, Rautengarten C *et al.* (2006). Segregation distortion in Arabidopsis C24/Col-0 and Col-0/C24 recombinant inbred line populations is due to reduced fertility caused by epistatic interaction of two loci. *Theor Appl Genet* **113**: 1551–1561.

Wang CM, Zhu CS, Zhai HQ, Wan JM (2005). Mapping segregation distortion loci and quantitative trait loci for spikelet sterility in rice (*Oryza sative* L.). *Genet Res* **86**: 97–106.

Xie SQ, Chen JG (2012). A statistical method for genetic mapping of sterility genes that exhibit epistasis in remote hybridization of plants using molecular markers in an F_2 population. *Chin Sci Bull* **57**: 2681–2687.

Xu Y, Zhu L, Xiao L, Huang N, McCouch SR (1997). Chromosomal regions associated with segregation distortion of molecular markers in F2, backcross, double haploid, and recombinant inbred populations in rice (*Oryza sativa* L.). *Mol Gen Genet* **253**: 535–545.

Yang J, Zhao X, Cheng K, Du H, Ouyang Y, Chen J *et al.* (2012). A killer-protector system regulates both hybrid sterility and segregation distortion in rice. *Science* **337**: 1336–1340.

Zhu CS, Wang CM, Zhang YM (2007). Modeling segregation distortion for viability selection I. Reconstruction of genetic linkage maps with distorted markers. *Theor Appl Genet* **114**: 295–305.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)