Theor Appl Genet (2010) 121:1071–1082
DOI 10.1007/s00122-010-1372-z

ORIGINAL PAPER

Effects of missing marker and segregation distortion on QTL mapping in F_2 populations

Luyan Zhang · Shiquan Wang · Huihui Li · Qiming Deng · Aiping Zheng · Shuangcheng Li · Ping Li · Zhonglai Li · Jiankang Wang

Received: 4 March 2010 / Accepted: 21 May 2010 / Published online: 10 June 2010

Abstract Missing marker and segregation distortion are commonly encountered in actual quantitative trait locus (QTL) mapping populations. Our objective in this study was to investigate the impact of the two factors on QTL mapping through computer simulations. Results indicate that detection power decreases with increasing levels of missing markers, and the false discovery rate increases. Missing markers have greater effects on smaller effect QTL and smaller size populations. The effect of missing markers can be quantified by a population with a reduced size similar to the marker missing rate. As for segregation distortion, if the distorted marker is not closely linked with any QTL, it will not have significant impact on QTL mapping; otherwise, the impact of the distortion will depend on the degree of dominance of QTL, frequencies of the three marker types, the linkage distance between the distorted marker and QTL, and the mapping population size. Sometimes, the distortion can result in a higher genetic variance than that of non-distortion, and therefore benefits the detection of linked QTL. A formula of the ratio of genetic variance explained by QTL under distortion and non-distortion was given in this study, so as to easily determine whether the segregation distortion marker (SDM) increases or decreases the QTL detection power. The effect of SDM decreases rapidly as its linkage relationship with QTL becomes looser. In general, distorted markers will not have a great effect on the position and effect estimations of QTL, and their effects can be ignored in large-size mapping populations.

Communicated by E. Carbonell.

L. Zhang, S. Wang and H. Li contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-010-1372-z) contains supplementary material, which is available to authorized users.

L. Zhang · Z. Li
School of Mathematical Sciences,
Beijing Normal University, Beijing 100875, China

L. Zhang · H. Li · J. Wang (✉)
Institute of Crop Science, The National Key Facility for Crop Gene Resources and Genetic Improvement, CIMMYT China,
Chinese Academy of Agricultural Sciences,
No. 12 Zhongguancun South Street, Beijing 100081, China
e-mail: wangjk@caas.net.cn; jkwang@cgiar.org

S. Wang · Q. Deng · A. Zheng · S. Li · P. Li
Rice Research Institute, Sichuan Agricultural University,
Chengdu 611130, China

Introduction

Quantitative trait locus (QTL) mapping has become a routine approach in genetic studies of complex traits in plants, animals, and humans (Paterson et al. 1991; Lynch and Walsh 1998; Mackay 2001; Barton and Keightley 2002; Doerge 2002), where linkage mapping and association mapping are often referred. In comparison with association mapping, QTL linkage mapping in animals and human beings is normally based on pedigree data, but in plants it is more often based on biparental genetic populations. Backcrosses, doubled haploids, recombination inbred lines, and F_2 populations are commonly used biparental genetic populations in QTL linkage mapping.

In actual QTL mapping populations, genotypes with missing markers are common problems for various reasons (Martínez and Curnow 1994). If the missing rate of a marker is high, it may be not reliable and should not be used in QTL mapping. To avoid ill-conditioned variance–covariance

matrices in regression analysis (Little 1992), missing genotypes can be estimated by the genotypes of flanking markers using least squares, maximum likelihood, Bayesian methods, multiple imputation and other methods. In large-scale genotyping studies, imputation is also used to infer missing genotypic data, but the methods are more complicated, especially in human research (Yu and Schaid 2007; Browning 2008). Several methods have been proposed for QTL mapping with missing markers in various genetic populations (Martínez and Curnow 1994; Jiang and Zeng 1997; Butruille et al. 1999). These methods can help recover the information of individuals with missing markers. It is generally agreed that missing markers could have some negative impacts on QTL mapping. But less research has been conducted on how missing markers affect QTL mapping efficiency, i.e., the detection power, and QTL position and effect estimations, especially for the most recently developed methods.

Segregation distortion is another commonly encountered phenomenon in mapping populations. Segregation distortion loci (SDL) make the adjacent markers show distortion from the expected Mendelian ratio, which has been reported in several crop species (Xu 2008; Tai et al. 2000). Markers showing segregation distortion are called segregation distortion markers (SDMs), or simply distorted markers. Segregation distortion is always related to gamete gene, sterile gene and chromosome translocation (Hedrick and Muona 1990; Lorieux et al. 1995a; Luo and Xu 2003; Zhu et al. 2007). A number of statistical methods have been developed to detect the genes or loci related to distortion, i.e., to estimate the positions and effects of SDL. For example, a flanking-marker analysis method in selfing organisms was proposed by Hedrick and Muona (1990), and a maximum likelihood method was proposed by Luo and Xu (2003). The impact of SDL on constructing genetic linkage maps has been investigated (Garcia-Dorado and Gallego 1992; Lorieux et al. 1995a, b; Zhu et al. 2007). Relatively fewer studies have been conducted to investigate the influence of segregation distortion on QTL mapping. Under the additive model, Xu (2008) demonstrated that SDL are beneficial to QTL mapping about 44% of the time when segregation distortion of a locus is a random event. But knowledge of when SDM may benefit QTL mapping and when it may have negative effects on QTL with both additive and dominance effects is still lacking.

In populations consisting of two distinct genotypes, i.e., backcross populations, doubled haploids, and recombination inbred lines, only additive effects can be estimated. F_2 populations have three distinct genotypes, providing the probability of estimating additive and dominance effects simultaneously; however, the inclusion of more genetic effects also complicates the QTL mapping procedure. Zhang et al. (2008) found that the dominance effect of a QTL can cause interactions between markers in F_2 populations, and proposed an inclusive linear model that includes marker variables and marker interactions so as to completely control both additive and dominance effects of QTL. The inclusive linear model provides the theoretical basis for inclusive composite interval mapping (abbreviated as ICIM) in F_2 populations. ICIM highlights the importance of model selection and interval testing in QTL linkage mapping, and is a useful step forward for QTL mapping in biparental populations (Li et al. 2007; Wang 2009). Our objective in this study was to use ICIM to investigate the effects of missing marker and segregation distortion on QTL mapping. Though an F_2 population was used, major conclusions from this study can be extended to other populations.

Materials and methods

Imputation of missing markers

In ICIM (Li et al. 2007; Zhang et al. 2008; Wang 2009), all missing marker types were assigned to either the P_1 type, the F_1 type, or the P_2 type through an imputation algorithm that considers linkage relationship between markers, where P_1 and P_2 are the two parents of the mapping population. To perform missing marker imputation, we assume: (1) a linkage map has been constructed, which means markers have been grouped in order; (2) M (from P_1) and m (from P_2) are the two alleles at the missing marker locus (denoted as M); and (3) x is a random number from the uniform distribution $U(0, 1)$. Imputation is performed from the first marker on each chromosome, so the marker types at the marker loci before the current marker are always known. Taking the F_2 population as an example, one missing marker point can be classified into three cases.

Case I

The current missing marker is not linked with any non-missing markers, either on its left or its right side. In this case, no linkage information can be used, and the genotype for the missing marker is imputed randomly, depending on the Mendelian frequencies of the three types, i.e., 0.25, 0.50 and 0.25 in F_2 populations. That is to say, the missing marker point is assigned to MM if the random number $x > 0.75$, and to mm if $x < 0.25$; otherwise, it is assigned to Mm.

Case II

At least one non-missing marker exists on one side of the current missing marker. Let us assume that A and a are the two alleles of the nearest non-missing marker locus (denoted

as locus A), and the recombination frequency between marker loci M and A is c. In this case, the frequencies of marker types MM, Mm, and mm are dependent on the three types at the non-missing marker (Table S1 of Supplementary material). For example, if the non-missing marker type is AA, the conditional probabilities of genotypes MM, Mm and mm are $(1-c)^2$, $2c(1-c)$, and c^2, respectively. Thus, the missing marker type is assigned to MM if the random number $x < (1-c)^2$, to Mm if $(1-c)^2 < x < (1-c)^2 + 2c(1-c)$, and otherwise to mm.

Case III

Non-missing markers exist on both sides of the current missing marker. In this case, the frequencies of missing marker types are dependent on the nine marker types at the two non-missing markers (Table S2 of Supplementary material), which are then used to assign the missing marker types.

Genetic variations explained by one QTL under distortion and non-distortion

Assuming a and d are the additive and dominance effects of one QTL, and f_{QQ}, f_{Qq} and f_{qq} are the frequencies of three QTL genotypes QQ, Qq and qq, respectively. When there is no segregation distortion linked with this QTL, f_{QQ}, f_{Qq} and f_{qq} are expected to be equal to 0.25, 0.5 and 0.25, respectively, in an F_2 population, and the genetic variance of the QTL is

$$\sigma^2 = \frac{1}{2}a^2 + \frac{1}{4}d^2. \quad (1)$$

Let f_{MM}, f_{Mm} and f_{mm} be the frequencies of three marker types MM, Mm and mm of one SDM, with the recombination frequency c between the SDM and QTL. Due to linkage between SDM and QTL, the distortion at the marker locus will cause the distortion at the QTL locus as well. Thus, f_{QQ}, f_{Qq} and f_{qq} will deviate from the Mendelian frequencies 0.25, 0.5 and 0.25, depending on the recombination frequency between SDM and QTL, i.e., c, and frequencies of the three marker types, i.e., f_{MM}, f_{Mm} and f_{mm} (Tables 1, 2). Mean genotypic value and genetic variance of the QTL (Table 2) under distortion can be calculated as

$$\mu = m + (f_{QQ} - f_{qq})a + f_{Qq}d, \quad \text{and}$$
$$\sigma^2_{SD} = 1[f_{QQ} + f_{qq} - (f_{QQ} - f_{qq})^2]a^2 - 2f_{Qq}(f_{QQ} - f_{qq})ad + 1(f_{Qq} - f^2_{Qq})d^2. \quad (2)$$

Setting the degree of dominance of the QTL to $r = d/a$, from Eqs. 1 and 2, the ratio (k) of the variance under distortion to the variance of no distortion can be calculated as

Table 1 Frequencies of marker types and QTL genotypes under non-distortion and distortion

Genotype	Frequency under non-distortion	Frequency under distortion
MMQQ	$\frac{1}{4}(1-c)^2$	$f_{MM}(1-c)^2$
MMQq	$\frac{1}{2}c(1-c)$	$2f_{MM}c(1-c)$
MMqq	$\frac{1}{4}c^2$	$f_{MM}c^2$
MmQQ	$\frac{1}{2}c(1-c)$	$f_{Mm}c(1-c)$
MmQq	$\frac{1}{2}(1-2c+2c^2)$	$f_{Mm}(1-2c+2c^2)$
Mmqq	$\frac{1}{2}c(1-c)$	$f_{Mm}c(1-c)$
mmQQ	$\frac{1}{4}c^2$	$f_{mm}c^2$
mmQq	$\frac{1}{2}c(1-c)$	$2f_{mm}c(1-c)$
mmqq	$\frac{1}{4}(1-c)^2$	$f_{mm}(1-c)^2$

The recombination frequency between the distortion marker and QTL is represented by c, and the frequencies of the three marker types are f_{MM}, f_{Mm} and f_{mm}, respectively, where $f_{MM} + f_{Mm} + f_{mm} = 1$

$$k = \frac{\sigma^2_{SD}}{\sigma^2} = \frac{4[f_{QQ} + f_{qq} - (f_{QQ} - f_{qq})^2] - 8f_{Qq}(f_{QQ} - f_{qq})r + 4(f_{Qq} - f^2_{Qq})r^2}{2 + r^2}. \quad (3)$$

Equation 3 will be used to explain the effect of SDM on QTL mapping observed in simulation. In addition, Eq. 3 also provides a general way to explain the effect of SDM. In statistics, greater variation makes greater significance probability, and therefore has greater power to be detected. Therefore, if $k > 1$, distorted marker will increase the QTL detection power; if $k < 1$, distorted marker will decrease the QTL detection power; otherwise, distorted marker will not change the QTL detection power.

One actual F_2 population in rice

The actual rice F_2 population used in this study, consisting of 180 individuals, was derived by the Rice Research Institute, Sichuan Agricultural University, China (Ye et al. 2005, 2007). The cross was made between *indica* rice variety PA64s and *japonica* rice variety Nipponbare in Chengdu, China, in July 2002. Nipponbare was completely sequenced, and PA64s was partially sequenced in 2002. The F_1 population was planted in Hainan, China, in December 2002, and the F_2 was planted in Chengdu, China, in April 2003 for genotyping and phenotyping. A total of 137 polymorphism markers were screened in the F_2 population; the linkage groups built cover the 12 chromosomes of rice, with an average density of 17.1 cM. There is a total of 24,660 (i.e., 180 × 137) marker data points in the F_2 population, of which 2,240 are missing, accounting for 9.08% of the total marker data points.

Ten traits were used in QTL mapping for both additive and dominance effects: plant height, heading date, panicle

Table 2 Frequencies and genotypic values of three QTL genotypes, and mean and variance of the QTL under additive and dominance model

QTL genotype	Frequency	Genotypic value
QQ (f_{QQ})	$f_{MM}(1-c)^2 + f_{Mm}c(1-c) + c^2 f_{mm}$	$m + a$
Qq (f_{Qq})	$2(f_{MM} + f_{mm})c(1-c) + (1-2c+2c^2)f_{Mm}$	$m + d$
qq (f_{qq})	$f_{MM}c^2 + f_{Mm}c(1-c) + f_{mm}(1-c)^2$	$m - a$
Mean (μ)	$m + (f_{QQ} - f_{qq})a + f_{Qq}d$	
Variance (σ^2_{SD})	$[f_{QQ} + f_{qq} - (f_{QQ} - f_{qq})^2]a^2 - 2f_{Qq}(f_{QQ} - f_{qq})ad + (f_{Qq} - f^2_{Qq})d^2$	
Variance under no linked SDM (σ^2)	$\frac{1}{2}a^2 + \frac{1}{4}d^2$	

length, flag leaf length, spikelets per panicle, thousand kernel weight, density of panicle, grain length, grain width, and ratio of grain length to width. In ICIM, the two probabilities for entering and removing variables of stepwise regression in the first step were set at 0.01 and 0.02, respectively, and in the second step the LOD threshold was set at 3.0 to declare significant QTL. QTL mapping results for plant height and heading date were then used to build genetic models so as to investigate the effects of missing markers and segregation distortion on QTL mapping.

Simulation experimental design on the effects of missing marker and segregation distortion

For the simulation experiment in this study, the linkage map of the actual rice F_2 population and QTL identified for plant height and heading date were used, and two levels of population size, i.e., 180 (same as the actual mapping population) and 500, were considered. Seven missing levels, i.e., 0 (no missing marker), 5, 10, 15, 20, 25 and 30% were simulated to evaluate the effects of missing marker. No segregation distortion was included. Therefore, a total of 28 (i.e., 2 traits × 2 population sizes × 7 missing levels) simulation experiments were designed for studying the effect of missing marker on QTL mapping.

For investigating the effect of segregation distortion, two scenarios were considered, and missing markers were not included. In the first scenario, each SDM was simulated using the fitness identified in the actual mapping population; the other 136 markers were assumed to follow the Mendelian segregation ratio. All markers were located at the same positions as defined by the linkage map of the actual mapping population. This scenario simulates both linkage and independent inheritance between marker and QTL. As nine significant SDM and two population sizes (i.e., 180 and 500) were considered, and non-distortion was used as a control, a total of 40 simulation experiments (i.e., 2 traits × 2 population sizes × 10 distortions) were designed for this scenario.

In the second scenario, each SDM identified in the actual mapping population was assumed to be located at the nearest left marker linked with each identified QTL, and the other 136 markers were assumed to follow the Mendelian segregation ratio. Nine significant SDM, and 15 plant height and heading date QTL were identified in the actual mapping population. Non-distortion was used as a control. Thus, there is a total of 300 simulation experiments (i.e., 15 QTLs × 2 population sizes × 10 segregation distortions) in the second scenario.

For each simulation experiment, 1,000 F_2 populations were generated by the genetics and breeding simulation tool of QuLine, formerly called QuCim (Wang et al. 2003, 2004). QTL mapping was conducted by the software QTL IciMapping (available from http://www.isbreeding.net).

Calculating power and false discovery rate (FDR) from 1,000 simulated populations

In calculating power and false positives, each simulated QTL was assigned to a confidence interval of 15 cM centered at the QTL location estimated in the actual rice F_2 population. In each simulated population, a QTL is assumed to be correctly identified (or true positive) if the significant LOD peak is observed in the 15 cM confidence interval. In the confidence interval, if multiple significant peaks occur, only the highest one will be counted. In other chromosome regions, all peaks higher than the LOD threshold of 3.0 are counted as false positives.

From the 1,000 simulated populations in each experiment, power of each QTL is calculated as the proportion of the number of populations where the QTL is identified within the 15 cM confidence interval. FDR of each experiment is calculated as the proportion of total false positives to the total positives (i.e., summation of true positives and false positives in the 1,000 simulated populations). Thus, in each simulation experiment, power is calculated per QTL, but there is only one FDR calculated for each experiment, due to the difficulty determining which QTL causes which false positive in QTL mapping.

Results

QTL identified by ICIM in the actual rice F_2 population

A total of 61 QTLs were identified for the 10 agronomic traits; 16 of them were approximately additive, 15 were

Table 3 Estimated location, additive and dominance effects of QTL identified by ICIM for plant height and heading date in the rice F_2 population

QTL	Marker interval	Distance[a] (cM)	Additive effect	Dominance effect	LOD score	PVE[b] (%)	Degree of dominance
Plant height							
qPH1-1	RM246–RP2	12.0	−0.57	−7.98	8.04	12.03	13.96
qPH1-2	RP82–RP3	19.5	−8.59	0.59	15.54	25.57	−0.07
qPH3-1	RM523–RM251	16.9	4.35	−4.86	6.51	13.30	−1.12
qPH3-2	RP242–RM520	11.4	−4.69	−1.00	5.04	6.84	0.21
qPH4	RP67–OSR15	13.7	−3.56	−2.09	4.61	5.53	0.59
qPH5	RM159–RP299	13.0	−0.44	−4.48	3.13	3.86	10.24
qPH6	RP199–RM276	6.2	−0.79	−5.05	3.17	4.96	6.36
qPH7	RM82–RM180	7.0	0.26	6.48	5.27	7.56	25.24
qPH12	RM19–RM247	2.4	−1.66	3.93	3.98	5.44	−2.36
Heading date							
qHD1	RM212–RP82	22.1	1.74	−0.30	3.65	7.27	−0.17
qHD3	RM523–RM251	19.9	0.88	−3.70	6.04	21.09	−4.19
qHD4	RP68–RP397	0.2	−0.77	1.85	3.58	5.24	−2.41
qHD5	RM430–RM163	5.7	−1.41	−1.46	4.79	8.20	1.04
qHD8	RP379–RM331	0.3	−1.78	−0.80	4.85	7.21	0.45
qHD11	RP222–RP165	6.2	0.15	−3.03	5.71	11.70	−20.15

[a] Distance between QTL and its nearest left marker

[b] PVE for the phenotypic variation of the rice F_2 population explained by each QTL

partially dominant, 7 were dominant and 23 were overdominant (Table S3 of Supplementary material). Due to the dominance effect, R^2 of additive QTL is smaller than R^2 of additive and dominant QTL, especially for traits with more dominant and overdominant QTL, such as plant height and panicle length. Panicle length had the largest number of QTL. Among the 17 QTLs of panicle length, 5 were additive, 3 were partially dominant, 1 was dominant and 8 were overdominant. R^2 of additive QTL in the stepwise regression was 25.58%, while R^2 of additive and dominant QTL was 61.26%, indicating the large variation caused by dominant effects. Grain length was found to have the lowest number of QTL, explaining 41.96% of the phenotypic variation of grain length. One of the two QTL was additive, and the other was partially dominant. Among the 61 identified QTL, one from panicle length had the largest phenotypic variation explained (PVE), i.e., 28.19%, and one from plant height had the smallest PVE, i.e., 3.86%. Fig. S1 of Supplementary material shows the PVE distribution of the 61 QTLs. Most QTLs explain 5.0–12.5% of the phenotypic variation, with an average PVE of 9.94%.

Details of the nine QTLs identified for plant height and six for heading date are given in Table 3; they were used to investigate the effects of missing markers and segregation distortion on QTL mapping. Two plant height QTLs were each located on chromosomes 1 and 3, and one QTL each was located on chromosomes 4–7, and 12. *qPH1-2*, located between markers RP82 and RP3 on chromosome 1, had the largest PVE (25.57%) and was approximately additive, i.e., the estimated additive effect was −8.59, and the dominance effect was 0.59. One heading date QTL each was located on chromosomes 1, 3–5, 8 and 11. *qHD3*, located between markers RM523 and RM251 on chromosome 3, had the largest PVE (21.09%) and was significantly overdominant, i.e., the additive effect was 0.88, and the dominant effect was −3.70.

Effect of missing markers on QTL mapping

The more markers are missing, the more QTL detection power decreases, regardless of the size of simulated populations (Fig. 1). However, a large amount of missing markers has greater effects on smaller effect QTL and smaller size populations. Take plant height as an example. When the size of simulated populations was 180, same as the size of the actual population, the detection powers at the seven levels of missing markers were 93.5, 92.7, 91.8, 89.8, 87.9, 84.7 and 84.7% for *qPH1-2* (having the largest PVE of 25.57%), and 9.5, 8.8, 7.7, 6.3, 5.4, 4.1 and 3.5% for *qPH5* (having the smallest PVE of 3.86%), respectively (Fig. 1a). When the size of simulated populations was 500, little reduction in power was observed for *qPH1-2* as more markers are missing (Fig. 1b). In comparison, the detection power of *qPH5* decreased from 48.6 to 40.8, 33.8, 28.7, 24.8, 20.7 and 15.1%, as the level of missing markers

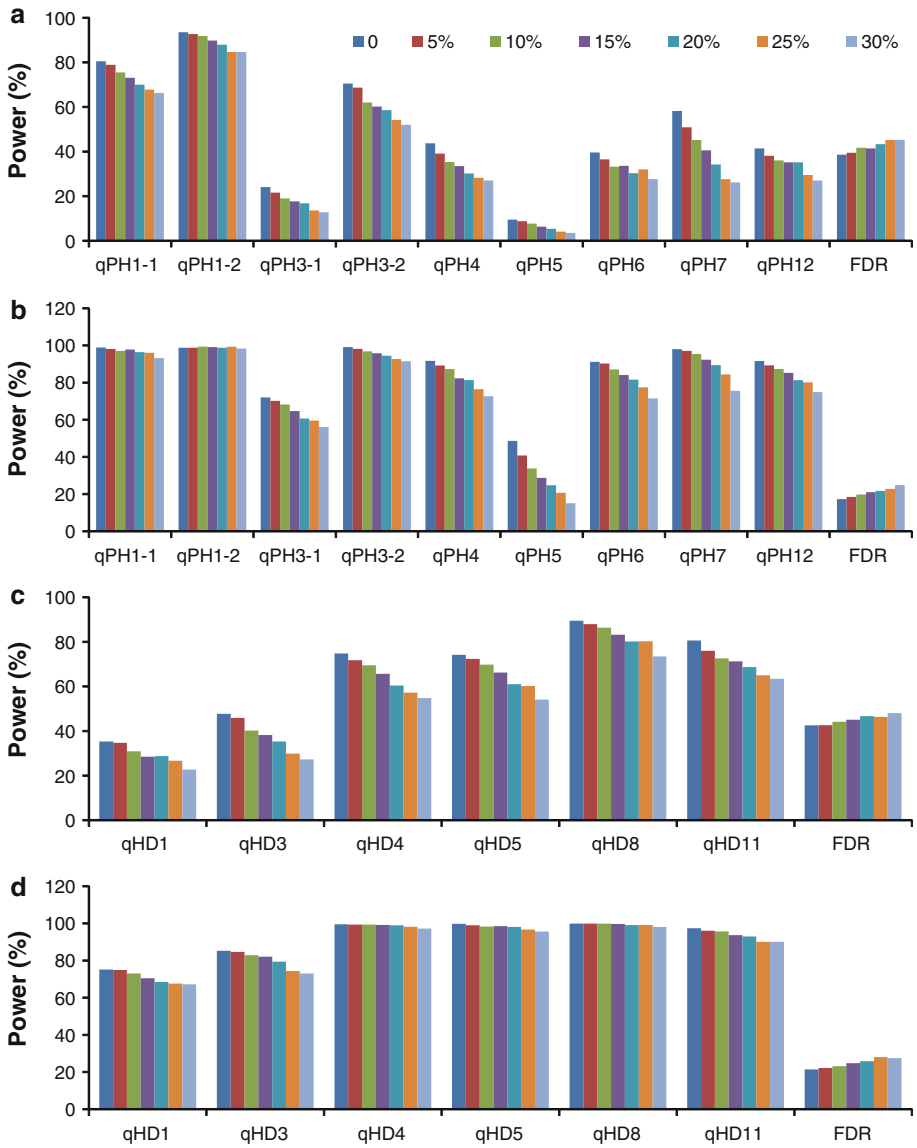

Fig. 1 Power analysis in 1,000 simulated F_2 populations for different levels of missing marker. **a** Plant height and population size 180, **b** plant height and population size 500, **c** heading date and population size 180, **d** heading date and population size 500. The confidence interval for each QTL was set at 15 cM, and the LOD threshold was 3.0. The *last bar* set in each figure represents the false discovery rate (FDR)

increased from 0 to 5–30% (Fig. 1b). Similar trends can be seen for heading date QTL (Fig. 1c, d).

Missing markers also increase the FDR in QTL mapping (Fig. 1). For the seven missing levels, the respective FDR values for plant height were 38.6, 39.5, 41.7, 41.4, 43.3, 45.2 and 45.3% when the size of the simulated populations was 180, and 17.3, 18.5, 19.8, 21.1, 21.8, 22.8 and 24.9% when the size of the simulated populations was 500.

To further illustrate the effect of missing markers, the estimated locations and effects of plant height QTL for population size 180 are shown in Table S4 of Supplementary material. The estimated positions, additive and dominant effects of QTL for populations with missing marker points were similar to those for populations with no missing marker points. Taking *qPH3-2* as an example (Table S4 of Supplementary material), the true distance between *qPH3-2* and its left marker was 11.40, while the true additive and dominant effects were −4.69 and −1.00. For the seven respective levels of missing markers, detection powers of *qPH3-2* were 70.5, 68.7, 62.0, 60.2, 58.6, 54.2 and 52.0%; the estimated distance between QTL and its left marker was 12.56, 12.55, 12.45, 12.27, 12.18, 12.58 and 12.02; the estimated additive effects were −5.11, −5.10, −5.06, −5.12, −5.09, −5.14 and −5.05; the estimated dominance effects were −0.67, −0.66, −0.72, −0.77, −0.64, −0.62 and −0.81. The estimated positions, additive effects and dominant effects of QTL were also similar to those of populations with no missing marker for population size 500 (results not shown).

The detection power of QTL in a population with size n and marker missing level p is similar to the power in a population whose size is $n \times p$ smaller than n, i.e., size of $n(1 - p)$ with no missing markers. To show this point, we also simulated populations with various sizes (Fig. S2 of

Supplementary material). For example, detection powers of *qPH1-1* for population size 180 at the missing levels 5–30% were 78.9, 75.5, 73.1, 70.0, 67.8 and 66.3%, respectively (Fig. 1a); for population size 5–30% smaller than 180 (i.e., 171, 162, 153, 144, 135 and 126) with no missing markers and no distortion, detection powers of *qPH1-1* were 78.3, 77.3, 73.6, 70.9, 66.0 and 66.8%, respectively (Fig. S2a of Supplementary material). Therefore, the effect of missing markers can be quantified by a reduced size of the mapping population where no missing marker is present.

Segregation distortion in the actual mapping population

From the significance test against the Mendelian segregation ratio 1:2:1 in the actual F_2 population (Fig. 2), the largest value of the significance test represented by the negative logarithm of probability was 20.02, corresponding to marker RP129 on chromosome 12. When the type I error was 0.05, 33 of the 137 markers showed significant segregation distortion. Nine significant markers are shown in Table 4. Other markers around RM304 and RP129 were significant likely due to linkage and therefore were not included in the simulation experiment. The top nine SDMs were distributed on chromosomes 2, 3, 5, 8, 10, 11 and 12 (Fig. 2).

Effect of segregation distortion on QTL mapping

In the first distribution scenario of SDM, the size of simulated populations was 180. When an individual distorted marker was assumed to be at the same location in the genome as shown in the actual linkage map, QTL detection power and FDR were similar to non-distortion; this holds true except in the case of *qPH12* (Fig. 3). This QTL was linked with RP129, which resulted in an increased detection power compared with non-distortion (Fig. 3a). The estimated effects and standard errors were similar, whether the QTL was linked with a SDM or not (results not shown). RP129 was located on chromosome 12, which was closely linked with *qPH12* at a distance of 15 cM. The power of *qPH12* is 63.5% for RP129, which is obviously higher than non-segregation distortion and other SDM. RM159, located on chromosome 5, was closely linked with *qPH5* at a distance of 13 cM. However, the detection power of *qPH5* for each SDM was too low to observe the impact of segregation distortion on QTL mapping. For other SDM, the corresponding markers were not closely linked with any plant height or heading date QTL, and no significant impact on detection power was observed.

In the second distribution scenario, for both plant height and heading date QTL, we assume that each of the nine SDMs was located at its nearest left marker locus. Significant difference in detection power was observed for a number of QTLs for each population size (i.e., 180 and 500). When the size of simulated populations was 180, the detection powers of *qPH1-1*, *qPH5*, *qPH7* and *qHD11* were similar to or lower than non-distortion (Table 5). Taking *qPH7* as an example, we assumed each distorted marker was located at RM82, which is closely linked with *qPH7*. The difference in power of the distorted marker and that of non-distortion was −13.7, −4.2, −0.6, −4.0, −1.6, −13.7, −5.3, −5.1 and −7.3% for the nine distorted markers, respectively. All SDM resulted in reduced detection power compared with non-distortion.

However, for other QTL, some SDMs may result in higher detection power, while others result in lower detection power. Whether one SDM results in higher or lower detection power also depends on the linked QTL. For example, for *qPH3-2*, the deviations in power of the nine SDMs from the power of non-segregation distortion were 0.7, 0.6, 9.2, −8.3, −4.5, 8.8, −14.1, −27.0 and 8.8%. RM44, RM304, RM552 and RP129 resulted in lower power than non-segregation distortion, while other SDMs resulted in similar or even higher power. But for *qHD4*, RM159, RM147 and RM491 resulted in lower power than non-segregation distortion, while other SDMs resulted in similar or higher power.

Similar trends can be observed for simulated populations of size 500 for all 15 QTLs (Table S5 of Supplementary material). However, the impact of SDM was not as obvious as the simulated populations of size 180, and the

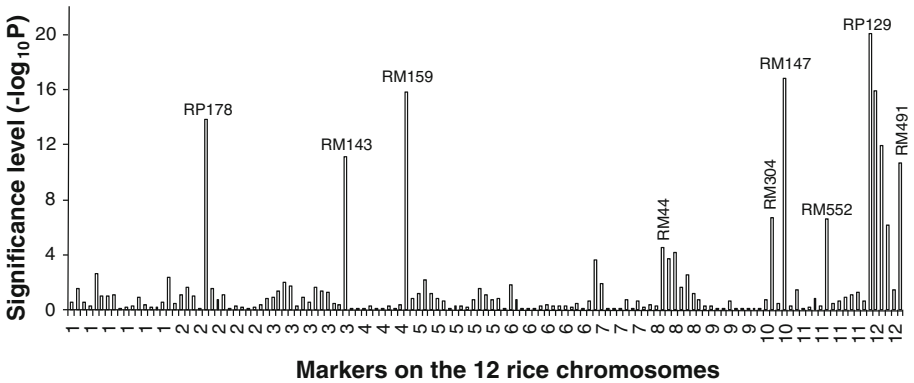

Fig. 2 Test of segregation distortion for the 137 markers in the rice F_2 population. The top nine peaks were marked and then used to investigate the effect of distortion on QTL mapping

Table 4 Nine most significant segregation distortion markers in the rice F_2 population

Marker name	Chr.	Sample size			M:m	Negative logarithm of probability	Fitness[a]		
		MM	Mm	mm			MM	Mm	mm
RP178	2	72	33	27	2.03	13.83	1.00	0.23	0.38
RM143	3	85	64	27	1.98	11.14	1.00	0.38	0.32
RM159	5	32	33	79	0.51	15.84	0.41	0.21	1.00
RM44	8	64	83	22	1.66	4.54	1.00	0.65	0.34
RM304	10	75	78	25	1.78	6.69	1.00	0.52	0.33
RM147	10	39	27	78	0.57	16.80	0.50	0.17	1.00
RM552	11	57	110	13	1.65	6.60	1.00	0.96	0.23
RP129	12	92	87	1	3.04	20.02	1.00	0.47	0.01
RM491	12	27	28	60	0.55	10.69	0.45	0.23	1.00

[a] Assume marker allele M is from parent PA64s, and m is from Nipponbare. Let n_{MAX} be the maximum value of n_{MM}, $0.5n_{Mm}$ and n_{mm}, where n_{MM}, n_{Mm} and n_{mm} are the sample sizes of the three marker types MM, Mm and mm, respectively. Fitness is n_{MM}/n_{MAX} for marker type MM, $0.5n_{Mm}/n_{MAX}$ for marker type Mm, and n_{mm}/n_{MAX} for marker type mm

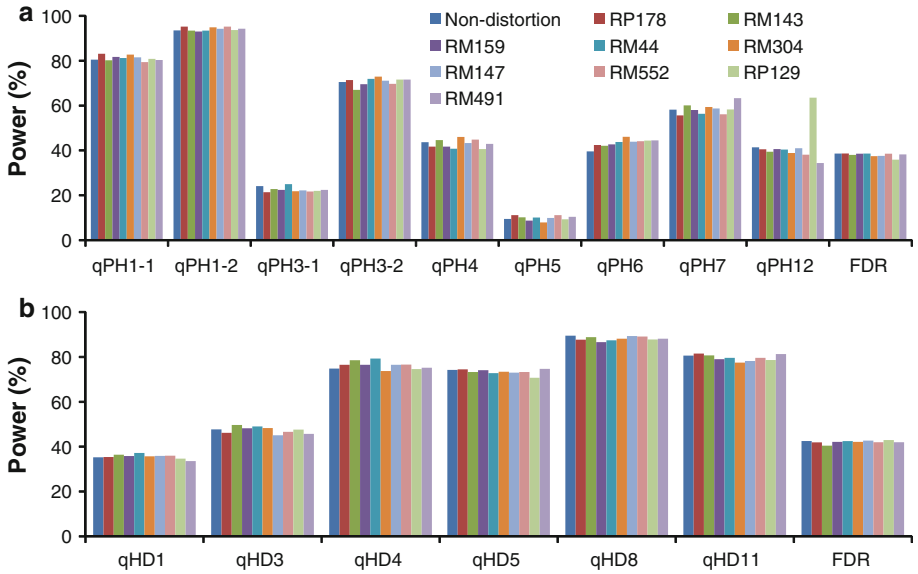

Fig. 3 Power analysis in 1,000 simulated F_2 populations of size 180 for the first distribution scenario of segregation distortion markers. In this scenario, each segregation distortion marker was simulated using the fitness identified in the actual mapping population, and the other 136 markers were assumed to follow the normal Mendelian segregation ratio. All markers were located at same positions as defined by the linkage map of the actual mapping population. **a** Plant height QTL and **b** heading date QTL. The confidence interval for each predefined QTL was set at 15 cM, and the LOD threshold was 3.0. The *last bar* set in each figure represents the false discovery rate (FDR)

difference in power for the nine SDMs was not as large as population of size 180. The difference in detection power between the nine SDMs and non-segregation distortion was also smaller than population size 180. Let us take *qPH7* and *qHD5* as examples, whose PVE were 7.56 and 8.20%, respectively. For *qPH7*, the powers of all SDM were consistently high (i.e., 92.6–98.5%), while the maximum difference in power between the nine SDMs and non-segregation distortion was only 5.4%. For *qHD5*, the powers of RM159, RM147 and RM491 were similar to or higher than that of non-segregation distortion, and lower than non-segregation distortion for other SDM, especially RM552 and RP129. For power of *qHD5* in RP129, the power difference between no distortion and RP129 was 70.8 and 81.4% for population sizes 180 and 500, respectively. However, the powers of *qHD5* in no distortion for population size 180 and 500 were 74.2 and 99.80%, respectively.

To further understand the effect of distortion, the estimated locations and effects of the plant height QTL are shown in Table S6 of Supplementary material for the

Table 5 Simulated QTL detection power and ratio of variance (k) for the second distribution scenario of segregation distortion markers

QTL	*qPH1-1*	*qPH1-2*	*qPH3-1*	*qPH3-2*	*qPH4*	*qPH5*	*qPH6*	*qPH7*	*qPH12*	*qHD1*	*qHD3*	*qHD4*	*qHD5*	*qHD8*	*qHD11*
Power															
Non-distortion	80.5	93.5	24.1	70.5	43.7	9.5	39.6	58.2	41.4	35.3	47.7	74.8	74.2	89.5	80.6
RP178	−8.3	1.4	6.7	0.7	0.4	−2.7	−10.2	−13.7	3.1	3.0	3.2	4.4	−7.0	−0.6	−6.7
RM143	−0.9	0.5	6.0	0.6	−4.5	−3.1	−5.7	−4.2	11.2	−0.8	6.6	10.3	−19.6	−7.9	−3.8
RM159	−1.1	−0.7	−2.0	9.2	18.0	−0.1	0.7	−0.6	−18.0	−0.7	−5.8	−23.5	11.7	4.1	−12.0
RM44	−3.2	−0.3	3.6	−8.3	−9.9	−1.3	−1.2	−4.0	9.1	−0.3	1.5	11.9	−27.8	−13.8	0.6
RM304	−0.2	0.4	7.7	−4.5	−6.6	−0.4	−1.3	−1.6	12.9	0.1	3.6	12.0	−22.0	−9.6	−0.1
RM147	−1.5	0.3	−1.1	8.8	24.3	−1.7	−3.5	−13.7	−16.7	0.3	−6.3	−27.0	10.0	3.4	−11.9
RM552	−3.3	−1.2	5.1	−14.1	−16.5	−3.1	−7.5	−5.3	12.6	−4.0	0.5	13.3	−47.2	−37.5	0.6
RP129	−2.0	0.2	29.3	−27.0	−27.9	−2.4	−3.4	−5.1	23.0	−4.5	27.1	17.7	−70.8	−72.2	0.1
RM491	0.1	0.0	−2.4	8.8	16.4	−0.5	0.9	−7.3	−16.5	1.3	−10.9	−20.3	10.0	5.4	−10.6
Ratio of variance															
RP178	0.71	1.29	1.30	1.18	1.02	0.69	0.67	0.72	1.10	1.32	0.95	1.09	0.86	1.08	0.79
RM143	0.86	1.09	1.34	0.95	0.80	0.84	0.79	0.89	1.26	1.14	1.14	1.26	0.69	0.85	0.97
RM159	0.76	1.31	0.88	1.38	1.39	0.78	0.82	0.73	0.68	1.27	0.64	0.68	1.31	1.39	0.68
RM44	0.93	0.93	1.27	0.79	0.67	0.90	0.85	0.96	1.28	0.98	1.20	1.27	0.60	0.70	1.05
RM304	0.91	1.00	1.31	0.86	0.72	0.89	0.84	0.95	1.29	1.05	1.19	1.28	0.64	0.77	1.03
RM147	0.65	1.46	1.00	1.50	1.45	0.66	0.71	0.63	0.71	1.43	0.61	0.71	1.31	1.48	0.59
RM552	0.86	0.70	1.18	0.54	0.40	0.83	0.76	0.90	1.25	0.76	1.18	1.24	0.36	0.44	1.01
RP129	0.86	0.59	1.38	0.33	0.10	0.80	0.69	0.92	1.48	0.70	1.37	1.48	0.03	0.17	1.09
RM491	0.78	1.33	0.92	1.39	1.40	0.80	0.85	0.76	0.72	1.29	0.68	0.72	1.32	1.41	0.71

Power was determined from 1,000 simulated F_2 populations of size 180, same as that of the actual rice population. The values are underlined if the change of power under distortion compared with non-distortion was not coincident with whether $k > 1$ or $k < 1$

second distribution scenario and population size 180. The estimated positions, additive effects and dominance effects of QTL from populations with distortion were similar to those of QTL from populations with non-segregation distortion. Taking *qPH3-2* as another example (Table S6 of Supplementary material), the estimated distance between QTL and its left marker in non-distortion was 12.56, 12.53, 12.54, 12.45, 12.46, 12.29, 12.62, 12.60, 12.43 and 12.34 for the nine SDMs; the estimated additive effects were −5.11, −4.97, −5.06, −4.89, −5.10, −5.07, −4.90, −5.42, −5.64 and −4.93; and the estimated dominance effects were −0.67, −0.47, −0.59, −0.84, −0.71, −0.54, −0.77, −0.64, −0.49 and −0.88, respectively. The estimated positions, additive effects and dominance effects of QTL were also similar to those of populations with non-segregation distortion and population size 500 (Table S7 of Supplementary material).

Theoretical explanation on the effect of distortion by the change of variance explained by QTL

The effect of segregation distortion on QTL mapping can be quantified by the ratio (k) of two variances explained by QTL under distortion and non-distortion, given by Eq. 3. Theoretically, SDM will not affect the QTL detection power on QTL mapping when $k = 1$. If σ^2_{SD} was higher than σ^2, i.e., $k > 1$, SDM will benefit QTL mapping; if $k < 1$, SDM will reduce the QTL detection power. In the rice F_2 population, genetic variances explained by four QTLs (i.e., *qPH1-1*, *qPH5*, *qPH6* and *qPH7*) when inked with any of the nine SDMs were smaller than those when no distortion was linked (Table 5), i.e., $k < 1$. But for other QTLs, k was lower than 1 when linked with some SDMs, but was greater than 1 when linked with other SDMs. For example, k of *qPH3-2* was 1.18, 1.38, 1.50 and 1.39 when linked with RP178, RM159, RM147 and RM491, but was 0.95, 0.79, 0.86, 0.54 and 0.33 when linked with RM143, RM44, RM304, RM552 and RP129, respectively. Most changes in power compared with non-distortion were coincident with the change of genetic variance determined by the ratio k. There are only a few exceptions in our simulation study (Table 5). For example, the power of *qPH3-2* in RM143 for population size 180 is 0.6% higher than no distortion while the value, but $k = 0.95$. However, the change of powers from non-distortion to distortion is very small for all exceptions. Power of *qPH3-2* in non-distortion is 70.5% and the change of power from non-distortion to distortion is only

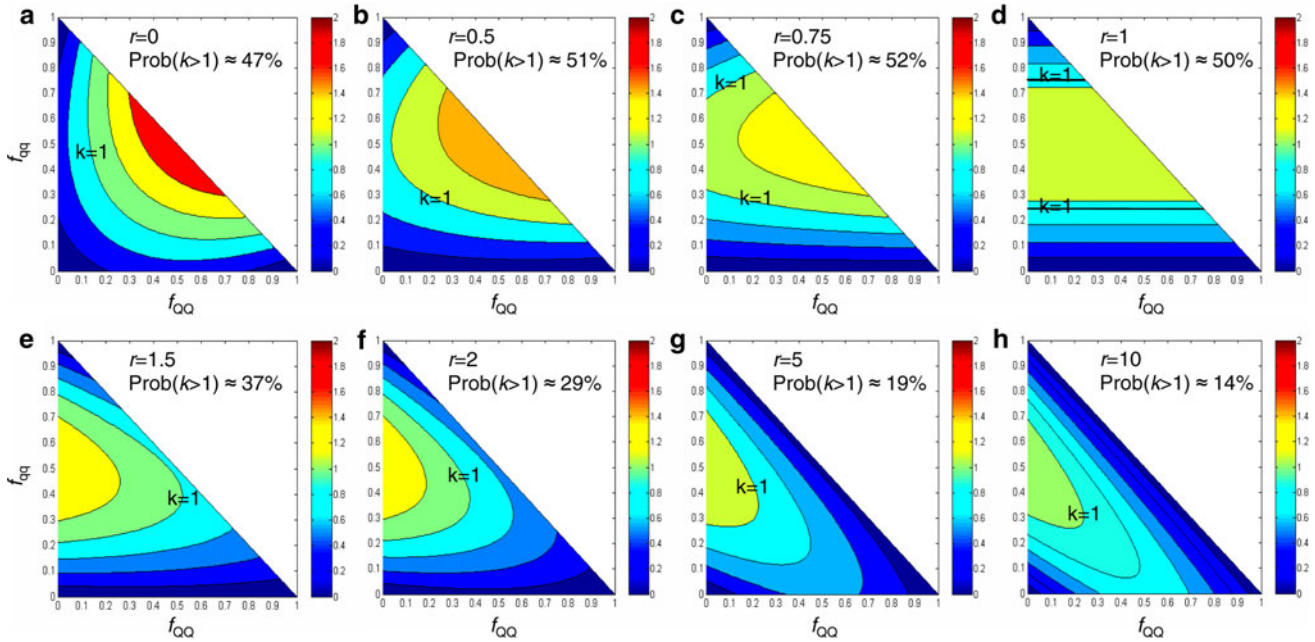

Fig. 4 Theoretical ratio of QTL variances under distortion and non-distortion for eight non-negative degrees of dominance of QTL. **a** $r = 0$, **b** $r = 0.5$, **c** $r = 0.75$, **d** $r = 1$, **e** $r = 1.5$, **f** $r = 2$, **g** $r = 5$, and **h** $r = 10$. Probability is approximated by the relative area where $k > 1$

0.6%. The small deviation may be caused by random errors in power simulation.

When $a = 0$, $\sigma^2_{SD} = (f_{Qq} - f^2_{Qq})d^2$, $\sigma^2 = \frac{1}{4}d^2$, and $k = 4(f_{Qq} - f^2_{Qq})$. It can be easily seen that $k \leq 1$ for any f_{Qq}, and the maximum value $k = 1$ is achieved when f_{Qq}=0.5. f_{Qq}=0.5 indicates that the sample size of Qq is equal to the combined sample size of QQ and qq, resulting in the same QTL detection power as that of non-distortion. Otherwise, the linked distortion will reduce QTL detection power. If $d = 0$ or equally $r = 0$, $k = 2[f_{QQ} + f_{qq} - (f_{QQ} - f_{qq})^2]$. It can be proved that the maximum value $k = 2$ is achieved when $f_{Qq} = 0$, and $f_{QQ} = f_{qq} = 0.5$. These are the genotypic frequencies in F_1-derived RIL populations. In other words, non-distortion F_1-derived RIL populations have the greatest power in detecting additive QTL.

Figure 4 shows the value of k for eight special values of the degree of dominance, i.e., $r = 0$, 0.5, 0.75, 1, 1.5, 2, 5 and 10. If the distortion is a random event, k is greater than 1 at chances about 47, 51, 52, 50, 37, 29, 19 and 14%, respectively (Fig. 4), estimated by the relative area where $k > 1$. The effect of f_{qq} on variance of QTL for $r = r_0$ is similar to the effect of f_{QQ} for $r = -r_0$ (Fig. S3 of Supplementary material). For example, when $r = 1$, k is larger than 1 when $0.25 < f_{qq} < 0.75$ (Fig. 4d); when $r = -1$, k is greater than 1 when $0.25 < f_{QQ} < 0.75$. The figure for $r = 2$ (Fig. 4f) is symmetrical to the figure for $r = -2$ (Fig. S3f of Supplementary material) on line $f_{QQ} = f_{qq}$.

Discussion

Missing marker and segregation distortion are common in most QTL mapping populations. We used plant height and heading date QTL identified in one rice F_2 population to investigate their effects on QTL mapping. We also used the actual linkage group in this population to design simulation experiments, thus avoiding over-simplifying putative genetic models commonly adopted in simulation studies. Though we used an F_2 population, the effects of missing markers and SDMs observed in this study are extendable to other biparental populations, e.g., backcross populations, doubled haploids and recombination inbred lines.

Missing markers reduce QTL detection power and increase the FDR. The influence of missing genotypes is more obvious in small-size populations than in large ones. However, the reduced power will not result in biased estimations of QTL location and effect. We believe this conclusion will also prove true when other populations are used. The effect of missing markers can be quantified by a reduced size of the mapping population. The detection power of QTL in a population with size n and missing rate p is similar to the power in a population with no missing markers whose size is reduced by p, i.e., $n(1 - p)$. Though commonly used for randomly missing markers in QTL mapping, the imputation algorithm is impossible to recover all linkage information lost by missing markers. When missing markers show schematic patterns, there may be other approaches which are able to better recover the missing linkage information.

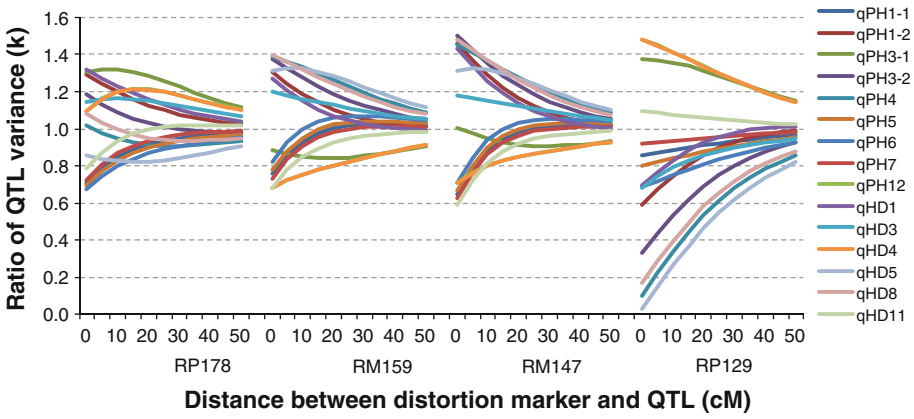

Fig. 5 Change in the theoretical ratio of the two QTL variances under distortion and non-distortion by the linkage distance (cM) between distorted marker and QTL. Four most distorted markers were used

Segregation distortion can be classified into different groups based on the frequencies of the three marker types (Table 4). For example, selection may sometimes occur not only for one marker type, but also for two marker types. The effect of segregation distortion on QTL mapping is complicated with F_2 populations, but can be quantified by the ratio (k) of the two variances explained by QTL under distortion and non-distortion. Theoretically, SDM will not affect the QTL detection power on QTL mapping when $k = 1$. For dominant QTL, Xu (2008) concluded that distortion is detrimental to the detection power. Here, dominant QTLs correspond to the type of QTL without additive effects in our study. For these QTLs, all distortions will cause k smaller than 1 and therefore reduce the QTL detection power. For additive QTL, Xu (2008) concluded that the SDLs are beneficial to QTL mapping ~44% of the time when segregation distortion of a locus is a random event, similar to the results for QTL without dominance effect (i.e., $r = 0$) in our study (Fig. 4a).

The effect of distortion is much simpler for populations consisting of two genotypes. Let frequencies of two QTL types be p and $1 - p$ in an actual population, and frequencies of two genotypes in Mendelian segregation ratio be f and $1 - f$. So, we can easily have

$$k = \frac{\sigma^2_{SD}}{\sigma^2} = \frac{p(1-p)}{f(1-f)} = \frac{1-(1-2p)^2}{1-(1-2f)^2}.$$

When two genotypes have the expected ratio 1:1 (i.e., $f = 0.5$), such as F_1-derived RIL or F_1-derived doubled haploids, any distortion will cause k smaller than 1, and therefore reduce QTL detection power. When two genotypes have other expected ratios, such as 3:1 (i.e., $f = 0.75$) in backcross-derived recombination inbred lines or backcross-derived doubled haploids, the distortion resulting in k larger than 1 will increase the QTL detection power. By introducing the variance ratio k, the effect of distortion in any biparental populations can be appropriately quantified.

How far one QTL can be affected by one linked SDM? We used the four most distorted markers, i.e., RP178, RM159, RM147 and RP129, to show the value of k by distance between distorted marker and QTL (Fig. 5). The value of k trends to 1 with the increase of distance between SDM and QTL for all 4 SDMs and 15 QTLs, especially when the distance is over 40 cM. Thus, SDM will not have obvious effect on QTL mapping when the distance between SDM and QTL is over 40 cM. Taking RP129 as an example, f_{QQ}, f_{Qq} and f_{qq} are equal to 0.51, 0.48 and 0.01, respectively. When the distance between SDM and QTL is 0, the maximum k is 1.48 for *qPH12* and the minimum k is 0.03 for *qHD5*. When the distance between SDM and QTL is 5, 10, …, and 50 cM, k is 1.45, 1.41, 1.38, 1.34, 1.30, 1.26, 1.23, 1.20, 1.17 and 1.14 for *qPH12*; and is 0.15, 0.26, 0.37, 0.46, 0.55, 0.62, 0.68, 0.74, 0.78 and 0.82 for *qHD5*, respectively. Clearly, if an SDM is linked closely with some QTL, the impact of the SDM will depend on distortion types and the effect of the QTL, i.e., f_{QQ}, f_{Qq}, f_{qq}, additive effect a and dominance effect d. If the variance ratio of k is larger than 1, distortion will be beneficial to QTL mapping; otherwise, it will be detrimental to QTL mapping. SDM will not have obvious effect on QTL mapping when the distance between SDM and QTL is over 40 cM.

In conclusion, segregation distortion affects the detection power of QTL when QTL and SDMs or loci are closely linked. Sometimes it makes QTL detection easier, but sometimes it reduces the QTL detection power, as is determined by the change of genetic variance explained by QTL. But in general, segregation distortion will not produce more false QTL, nor will it have significant impact on the estimation of QTL position and effect. So in practice, if the distortion is not extremely serious, the effect from distortion can be ignored in large-size mapping populations.

Acknowledgments This work was supported by the National 973 Projects of China (no. 2006CB101700) and Natural Science Foundation of China (no. 30771351).

References

Barton NH, Keightley PD (2002) Understanding quantitative genetic variation. Nat Rev Genet 3:11–21

Browning SR (2008) Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet 124:439–450

Butruille DV, Guries RP, Osborn TC (1999) Linkage analysis of molecular markers and quantitative trait loci in populations of inbred backcross lines of *Brassica napus* L. Genetics 153:949–964

Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet 3:43–52

Garcia-Dorado A, Gallego A (1992) On the use of the classical tests for detecting linkage. Heredity 83(2):143–146

Hedrick PW, Muona O (1990) Linkage of viability genes to marker loci in selfing organisms. Heredity 64:67–72

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

Li H, Ye G, Wang J (2007) A modified algorithm for the improvement of composite interval mapping. Genetics 175:361–374

Little RJA (1992) Regression with missing X's: a review. J Am Stat Assoc 87:1227–1237

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 population. Theor Appl Genet 90:81–89

Luo L, Xu S (2003) Mapping viability loci using molecular markers. Heredity 90:459–467

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Inc, Sunderland, MA

Mackay TFC (2001) Quantitative trait loci in *Drosophila*. Nat Rev Genet 2:11–20

Martínez O, Curnow RN (1994) Missing markers when estimating quantitative trait loci using regression mapping. Heredity 73:198–206

Paterson AH, Damon S, Hewitt JD, Zamir D, Rabinowitch HD, Lincoln SE, Lander ES, Tanksley SD (1991) Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. Genetics 127:181–197

Tai GCC, Seabrook JEA, Aziz AN (2000) Linkage analysis of anther-derived monoploids showing distorted segregation of molecular markers. Theor Appl Genet 101:126–130

Wang J (2009) Inclusive composite interval mapping of quantitative trait genes. Acta Agronom Sinica 35(2):239–245

Wang J, van Ginkel M, Podlich D, Ye G, Trethowan R, Pfeiffer W, DeLacy IH, Cooper M, Rajaram S (2003) Comparison of two breeding strategies by computer simulation. Crop Sci 43:1764–1773

Wang J, van Ginkel M, Trethowan R, Ye G, Delacy I, Podlich D, Cooper M (2004) Simulating the effects of dominance and epistasis on selection response in the CIMMYT Wheat Breeding Program using QuCim. Crop Sci 44:2006–2018

Xu S (2008) Quantitative trait locus mapping can benefit from segregation distortion. Genetics 180:2201–2208

Ye S, Zhang Q, Li J, Zhao B, Li P (2005) QTL mapping for yield component traits using (Pei'ai 64s/Nipponbare) F_2 population. Acta Agronom Sinica 31:1620–1627 (in Chinese with English abstract)

Ye S, Zhang Q, Li J, Zhao B, Yin D, Li P (2007) Mapping of quantitative trait loci for six agronomic traits of rice in Pei'ai 64s/Nipponbare F_2 population. Chin J Rice Sci 21(1):39–43 (in Chinese with English abstract)

Yu Z, Schaid DJ (2007) Methods to impute missing genotypes for population data. Hum Genet 122:495–504

Zhang L, Li H, Li Z, Wang J (2008) Interactions between markers can be caused by the dominance effect of quantitative trait loci. Genetics 180:1177–1190

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