Genet. Res., Camb. (2006), **88**, *pp.* 93–104. © 2006 Cambridge University Press
doi:10.1017/S0016672306008408 Printed in the United Kingdom

QTL mapping of grain length in rice (*Oryza sativa* L.) using chromosome segment substitution lines

JIANKANG WANG[1,2], XIANGYUAN WAN[1,3], JOSE CROSSA[2], JONATHAN CROUCH[2], JIANFENG WENG[3], HUQU ZHAI[1] AND JIANMIN WAN[1,3]*

[1] *Institute of Crop Science and The National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Beijing 100081, China*

[2] *Genetic Resources Enhancement Unit (GREU), Crop Research Informatics Laboratory (CRIL), International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico*

[3] *National Key Laboratory of Plant Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing 210095, China*

(*Received 24 April 2006 and in revised form 11 July 2006*)

Summary

Chromosome segment substitution (CSS) lines have the potential for use in QTL fine mapping and map-based cloning. The standard *t*-test used in the idealized case that each CSS line has a single segment from the donor parent is not suitable for non-idealized CSS lines carrying several substituted segments from the donor parent. In this study, we present a likelihood ratio test based on stepwise regression (RSTEP-LRT) that can be used for QTL mapping in a population consisting of non-idealized CSS lines. Stepwise regression is used to select the most important segments for the trait of interest, and the likelihood ratio test is used to calculate the LOD score of each chromosome segment. This method is statistically equivalent to the standard *t*-test with idealized CSS lines. To further improve the power of QTL mapping, a method is proposed to decrease multicollinearity among markers (or chromosome segments). QTL mapping with an example CSS population in rice consisting of 65 non-idealized CSS lines and 82 chromosome segments indicated that a total of 18 segments on eight of the 12 rice chromosomes harboured QTLs affecting grain length under the LOD threshold of 2·5. Three major stable QTLs were detected in all eight environments. Some minor QTLs were not detected in all environments, but they could increase or decrease the grain length constantly. These minor genes are also useful in marker-assisted gene pyramiding.

1. Introduction

The rapid progress in the development of polymorphic molecular markers has led to the intensive use of QTL mapping in genetics and plant breeding, and thus to the development of a number of statistical methods for QTL detection (Lander & Botstein, 1989; Haley & Knott, 1992; Martinez & Curnow, 1992; Jansen, 1994; Zeng, 1994; Churchill & Doerge, 1994; Whittaker *et al.*, 1996; Satagopan *et al.*, 1996; Sen & Churchill, 2001; Bogdan *et al.*, 2004). From a statistical perspective, methods for QTL mapping are generally based on three broad classes: least square (regression), maximum likelihood and Bayesian models. The simplest QTL mapping approach is based on linear regression at the marker position and is called single-marker or point analysis; it identifies the association of a phenotypic trait with marker classes by contrasting the means of marker types. In single-marker analysis, both the recombination fraction and the effect of QTLs are confounded and therefore the QTL effect is always underestimated.

The interval mapping method (IM) is based on a maximum likelihood parameter estimation that provides a likelihood ratio test for QTL detection (Lander & Botstein, 1989; Hackett, 1997). A major problem with IM is that the estimates of QTL locations and effects can be biased when QTLs are linked (Knott & Haley, 1992; Martinez & Curnow, 1992). Additionally, the estimate of QTL location has a rather wide confidence interval, and it is not efficient to use only two markers at a time for mapping

* Corresponding author. Institute of Crop Science and The National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, No. 12 Zhongguancun South Street, Beijing 100081, China. Tel.: +86 10 6891 8563. Fax: +86 10 6897 5212. e-mail: wanjm@caas.net.cn

analysis. Some of these problems are overcome by composite interval mapping (CIM) (Zeng, 1994; Jansen, 1994), which combines interval mapping with multiple marker regression analysis and therefore increases the precision of interval mapping. More recently, Bayesian models have been widely studied in QTL mapping (Satagopan *et al*., 1996; Sen & Churchill, 2001; Xu, 2003; Bogdan *et al*., 2004; Beaumont & Rannala, 2004). However, these methods have not yet been widely used.

Concerning the mapping populations required for QTL detection, such as F_2, backcross (BC), doubled haploids (DH), and recombination inbred lines (RIL), they can be classified into two categories: temporary populations and permanent populations. In a temporary population such as F_2 or BC, each individual in the population will segregate after self-pollination. In contrast, in a permanent population such as DH and RIL, each individual in the population is genetically homozygous, and the genetic construction will not change through self-pollination. Thus, in permanent populations the phenotypic value of complex quantitative traits can be measured repeatedly through a replicated experiment design, and the same genotype can be tested under different environments, allowing the study of genotype × environment interaction. Therefore, with permanent populations the random environmental errors can be better controlled and the precision of QTL mapping can be improved. Currently, IM and CIM are the two QTL mapping methods that are commonly used in these cases.

QTLs identified from those mapping populations described above normally have 10 cM or even wider confidence intervals (Alpert & Tanksley, 1996). Recently, permanent populations consisting of series of chromosome segment substitution (CSS) lines (also called introgression lines) have been used for gene fine mapping (Eshed & Zamir, 1995; Tanksley & Nelson, 1996; Nadeau *et al*., 2000; Wissuwa *et al*., 2002; Kubo *et al*., 2002; Belknap, 2003; Cowley *et al*., 2003; Wan *et al*., 2004, 2005, 2006). In the idealized case that each CSS line has a single segment from the donor parent, the standard analysis of variance (ANOVA), followed by multiple mean comparison between each line and the background parent, can be readily used to test whether the segment in the tested CSS line carries QTLs controlling the trait of interest (Belknap, 2003). Unfortunately, it will take much labour and time to develop a population consisting of idealized CSS lines. Usually in a preliminary CSS population each line carries a few segments from the donor parent. Due to high-intensity selection in the process of generating CSS lines, the gene and marker frequencies with CSS lines do not follow the same path as in a standard mapping population such as F_2, BC, DH or RIL. Thus the methods previously described are not suitable here, and to our knowledge no QTL mapping methods with such a population have been formally reported.

The main objective of this study was to identify QTLs for grain length in a non-idealized CSS population in rice (*Oryza sativa* L.) by means of a likelihood ratio test based on stepwise regression. Since the chromosome segments are represented by individual markers, in this study chromosome segments and markers will be used without any distinction.

2. Materials and methods

(i) *One CSS population in rice consisting of 65 non-idealized lines*

We illustrate the use of the proposed QTL method in a population consisting of 65 rice CSS lines and 82 chromosome segments. The two parents are the *japonica* rice variety Asominori (the background parent, denoted as P_1) and the *indica* rice variety IR24 (the donor parent, denoted as P_2) (for details about the development of this population, see Tsunematsu *et al*., 1996; Kubo *et al*., 1999, 2002; Wan *et al*., 2004). Each CSS line in the population contains 1 to 10 segments from the donor parent IR24. On average, each substitution segment exists in 3·7 CSS lines, and each CSS line carries 4·6 segments from the donor parent (Fig. 1).

The two parents and 65 CSS lines were grown in eight environments (denoted as E1 to E8; four locations in two years). Each entry plot contained 10 rows, and each row had 10 individual plants. A randomized complete block design with two replications was used in each environment. At maturity, each entry was harvested in bulk. After drying, grains were stored at room temperature for 3 months, and then the milled rice was used for measuring grain length (Wan *et al*., 2006). This trait will be used to show the outcomes from the proposed QTL mapping method.

One may want to use the standard *t*-test to show whether there is a significant difference between one CSS line, say CSSL1, and the background parent Asominori (Fig. 1). In this case, no significant difference implies no evidence that there are QTLs on the four markers M1, M2, M3 and M22. However, if there is a significant difference between CSSL1 and Asominori, it cannot be determined whether there is only one QTL on one of the four segments or multiple QTLs on segments M1, M2, M3 and M22, as the effects of the four segments are confounded.

(ii) *Inclusion of the background parent and exclusion of the donor parent*

For n idealized CSS lines, the correlation coefficient between any pair of markers is $r = -\frac{1}{n-1}$. If the background parent (P_1) is included, the coefficient becomes

https://doi.org/10.1017/S0016672306008408 Published online by Cambridge University Press

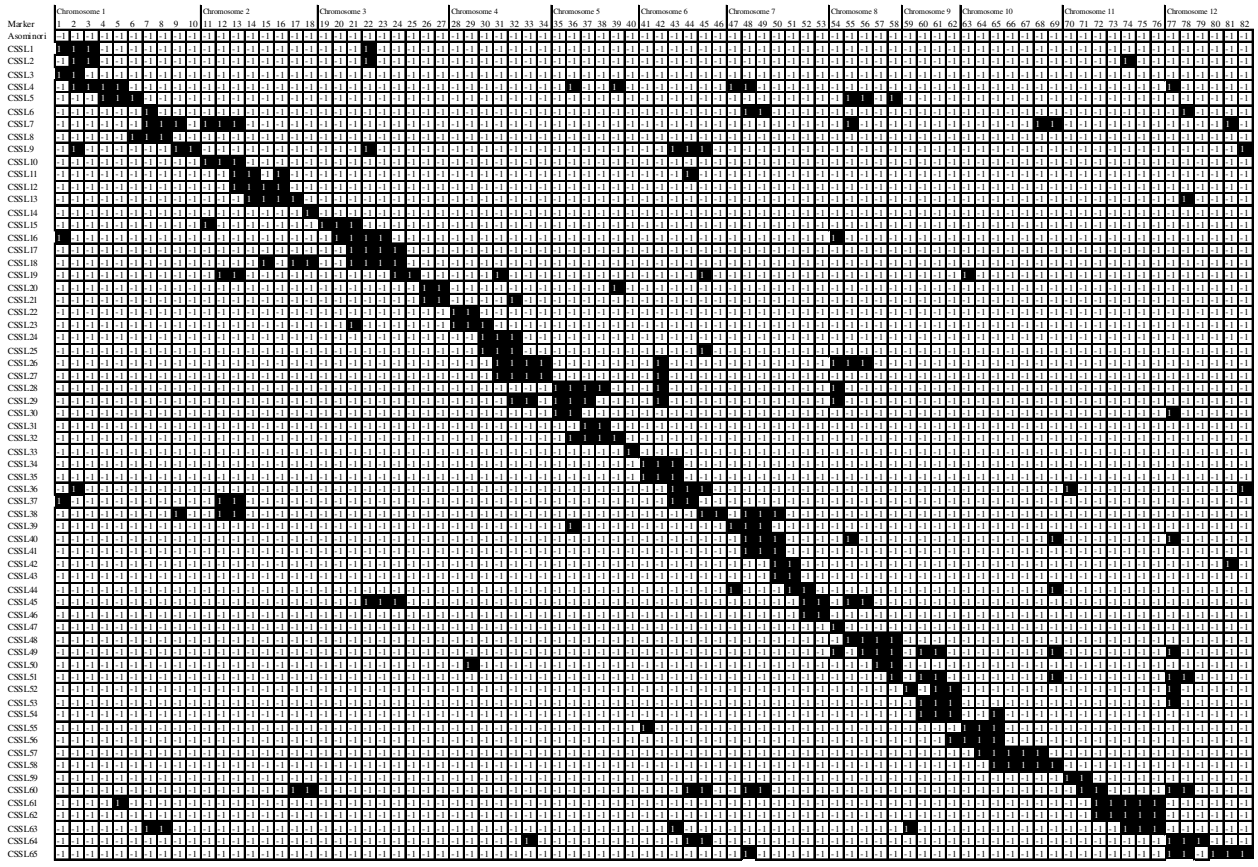

Fig. 1. A mapping population consisting of 65 non-idealized CSS lines and 82 chromosome segments (markers) derived from the two rice parents, *japonica* Asominori and *indica* IR24, where Asominori is the background (recurrent) parent (chromosome segments represented by -1) and IR24 is the donor parent (chromosome segments represented by 1).

$r = -\frac{1}{n}$, which is close to 0 when n is high. If the donor parent (P_2) is included, the coefficient becomes $r = \frac{n-3}{2(n-1)}$, which is close to 0·5. When both parents are included, the correlation coefficient between any pair of markers can be found to be $r = \frac{n-2}{2n}$. For non-idealized CSS lines, the inclusion of the donor parent results in a higher multicollinearity among markers. Therefore, in the sense of low correlation and low multicollinearity when using linear models, the background parent should be included in QTL mapping but the donor parent should not.

(iii) *The multiple linear model for CSS lines*

Suppose two parents, P_1 (the background parent) and P_2 (the donor parent), differ in t markers. The marker type is designated by -1 in P_1 and 1 in P_2. A total of n CSS lines were derived from the two parents through advanced backcrossing and marker-assisted selection (Eshed & Zamir, 1995; Tanksley & Nelson, 1996; Kubo *et al*., 2002), in which most of the chromosome segments in these CSS lines were from the background parent P_1. The phenotypic values for a quantitative trait of interest can be assessed and measured in a replicated field experiment where the average performance of the ith CSS line, y_i, is represented by the following linear model:

$$y_i = b_0 + \sum_{j=1}^{t} b_j x_{ij} + e_i \quad (1)$$

where $i = 0$ (for the background parent), 1, 2, …, n, b_0 is the intercept, b_j ($j = 1, \ldots, t$) is the partial regression coefficient of phenotype on the jth marker, which represents QTL additive effect on each segment, x_{ij} is the indicator variable for the jth marker in the ith CSS line, which is equal to -1 if the marker type is the same as in P_1 and 1 if the marker type is the same as in P_2, and e_i is the random experimental error following a normal distribution.

(iv) *Assessing a marker's multicollinearity*

When using model (1) on the dataset as shown in Fig. 1 an obvious problem occurs, i.e. the multicollinearity (Myers, 1990) among markers. For example, the correlation coefficients between segments of markers such as M14 and M16, M26 and M27, M66 and M67, and M75 and M76 are all equal to 1. In QTL mapping, however, the goals are to

https://doi.org/10.1017/S0016672306008408 Published online by Cambridge University Press

understand how various markers affect the phenotype of a trait of interest, and how well the markers predict the phenotypic performance. Therefore, multicollinearity among markers must be considered.

The level of multicollinearity can be assessed by the variable inflation factor and the condition number (Myers, 1990; Wang & Chow, 1994). Here we consider the condition number, which is defined as $k=\lambda_{\max}/\lambda_{\min}$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum eigenvalues of the correlation matrix between markers, respectively. When there is no collinearity at all, the correlation matrix in model (1) is an identity matrix, and therefore the eigenvalues and condition number will be equal to 1. As collinearity increases, some eigenvalues will be greater than 1 and some smaller than 1, and the condition number will increase. There is no theoretical way to find an appropriate threshold value for judging whether multicollinearity is high. An informal rule of thumb is that if the condition number is more than 100, multicollinearity is of concern, and if it is greater than 1000, multicollinearity is a very serious concern (Wang & Chow, 1994). In this study, 1000 was used as the threshold value for the condition number in QTL mapping with CSS lines.

A sequential process for decreasing multicollinearity was proposed. If marker pairs have a perfect correlation, meaning that these pairs of markers display no differences among all lines, one of them can be randomly deleted. If the correlation is high, but not perfect, the marker present in more lines is deleted, so that the mapping population will gradually approach an idealized one.

(v) *A likelihood ratio test combined with stepwise regression*

The most direct way to find the estimates of parameters in model (1) is to use stepwise regression. As usual, the largest P value for entering variables was set at 0·05, and smallest P value for removing variables was set at 0·10. If we assume we need to test whether there is a QTL in the chromosome segment represented by the jth marker, the observation values in model (1) can be adjusted by

$$\Delta y_i = y_i - \sum_{k \neq j} b_k x_{ik}.$$

When scanning for QTLs along the chromosomes, parameters in the above equation are estimated from stepwise regression, and will not change. The deviation thus calculated contains the QTL information on the current chromosome segment, and at the same time the effects from other QTLs have been controlled.

Suppose that the QTL has two alleles q and Q, where q is located in the background parent and Q in the donor parent. Rearrange Δy_i and let $i=0, 1, 2, \ldots, n_1$ represent the CSS lines having P_1 marker type, and $i=n_1+1, n_1+2, \ldots, n$ refer to the CSS lines having P_2 marker type. Thus Δy_i follows the distribution $N(\mu_1,\sigma_A^2)$ for $i=0, 1, 2, \ldots, n_1$ and the distribution $N(\mu_2,\sigma_A^2)$ for $i=n_1+1, n_1+2, \ldots, n$, where $N(\mu_1,\sigma_A^2)$ and $N(\mu_2,\sigma_A^2)$ represent the normal distributions of the two QTL genotypes qq and QQ, respectively. The existence of QTLs in the current chromosome segment can be tested by the following hypotheses: H_0: $\mu_1=\mu_2$ versus H_A: $\mu_1 \neq \mu_2$. Under the null hypothesis, H_0, all Δy_i follow the same normal distribution denoted by $N(\mu_0,\sigma_0^2)$. The mean and variance of this distribution can be estimated as

$$\mu_0 = \frac{1}{n+1}\sum_{i=0}^{n} \Delta y_i \text{ and } \sigma_0^2 = \frac{1}{n+1}\sum_{i=0}^{n} (\Delta y_i - \mu_0)^2.$$

The log-likelihood function under the null hypothesis H_0 can be calculated as

$$L_0 = \sum_{i=0}^{n} \ln f(\Delta y_i; \mu_0, \sigma_0^2)$$

where $f(\Delta y_i;\mu_0,\sigma_0^2)$ is the density function for normal distribution $N(\mu_0,\sigma_0^2)$. The log-likelihood function under the alternative hypothesis H_A is

$$L_A = \sum_{i=0}^{n_1} \ln f(\Delta y_i; \mu_1, \sigma_A^2) + \sum_{i=n_1+1}^{n} \ln f(\Delta y_i; \mu_2, \sigma_A^2)$$

where

$$\mu_1 = \frac{1}{n_1+1}\sum_{i=0}^{n_1} \Delta y_i, \quad \mu_2 = \frac{1}{n-n_1}\sum_{i=n_1+1}^{n} \Delta y_i,$$

and

$$\sigma_A^2 = \frac{1}{n+1}\left[\sum_{i=0}^{n_1} (\Delta y_j - \mu_1)^2 + \sum_{i=n_1+1}^{n} (\Delta y_j - \mu_2)^2\right].$$

Therefore, the likelihood ratio test can be built from the two likelihoods under the two hypotheses, and the LOD score in the current chromosome segment can be found. This process is called RSTEP-LRT hereafter. It can be proved that RSTEP-LRT is equivalent to the standard t-test for idealized CSS lines.

(vi) *Putative QTLs in a simulation study*

We considered 10 QTLs with different additive effects (Table 1), which have been used in Zeng (1994) to investigate the power of composite interval mapping in backcross populations. During simulation, the genotypic value of each CSS line was calculated from its genotype (or marker type in our case) and the putative QTL effects, on the basis of which the genotypic variance σ_g^2 can be calculated. According to

https://doi.org/10.1017/S0016672306008408 Published online by Cambridge University Press

Table 1. *Ten putative QTLs and percentage of genetic and phenotypic variances explained by each QTL*

QTL	QTL1	QTL2	QTL3	QTL4	QTL5	QTL6	QTL7	QTL8	QTL9	QTL10
Location	M3	M8	M11	M17	M21	M26	M31	M38	M41	M44
ADD[a]	0·42	0·75	0·58	1·02	−1·23	−1·26	−0·46	1·61	0·88	0·74
PGVE[b]	1·76	5·62	3·36	10·39	24·38	10·74	3·41	25·88	7·73	10·42
PPVE[c]	1·41	4·49	2·69	8·31	19·50	8·59	2·73	20·71	6·19	8·33

[a] Additive effect. A positive effect indicates the allele in IR24 increases the trait of interest, while a negative effect indicates the allele in IR24 decreases the trait value.
[b] Percentage of genetic variance explained by an individual QTL in the CSS population derived from the two rice parents, *japonica* Asominori and *indica* IR24.
[c] Percentage of phenotypic variance explained by an individual QTL when the heritability in the broad sense was 0·8.

the definition of the heritability in the broad sense (H), the error variance can be estimated from $\sigma_\varepsilon^2 = \frac{1-H}{H}\sigma_g^2$. The heritability in the broad sense was set at $H = 0{\cdot}8$ in the simulation study. Thus, a random error from the normal distribution $N(0,\sigma_\varepsilon^2)$ was added to the genotypic value to obtain the phenotypic value for each CSS line.

The genetic variance of a QTL is calculated from its additive effect a and the allele frequency p in the population, i.e. $4p(1-p)a^2$. Therefore, the percentages of genetic and phenotypic variances explained by each QTL can be estimated (Table 1). The two percentages of a QTL depend on its additive effect and the allele frequency. For example, QTL5 has a smaller effect than QTL6 in absolute value, i.e. 1·23 versus 1·26, but explains a much larger genetic variation than QTL6, i.e. 24·38 % versus 10·74 % (Table 1).

3. Results

(i) *The sequential process for decreasing multicollinearity*

The following sequential process was used to remove multicollinearity among markers in the CSS population shown in Fig. 1. First, duplicate markers with perfect correlation were identified and deleted. Marker pairs M14 and M16, M26 and M27, M66 and M67, and M75 and M76 have a perfect correlation ($r = 1$ in Table 2), meaning that these pairs of markers display no differences among all 66 lines (including the background parent Asominori). In this case, one of them was randomly deleted. Here, the deleted markers were M16, M27, M67 and M76 (Table 2).

Second, when the correlation was high but not perfect, the marker present in more lines was deleted. For example in step 5 (Table 1), M60 and M61 had a high correlation ($r = 0{\cdot}8872$). M60 was present in four lines and M61 in five lines, so M61 was deleted. This process was repeated until the threshold value of 1000 for the condition number is achieved. For the non-idealized CSS population used in this study (Fig. 1), the condition number becomes smaller than 1000 after 27 markers have been deleted (Table 2). At this point the highest correlation coefficient is 0·6508, which is lower than the empirical threshold value of 0·70 for the correlation coefficient reflecting multicollinearity (Myers, 1990). In comparison, 34 markers have to be deleted to reduce the condition number to this point if the donor parent were also included, which showed the negative effect of including the donor parent in QTL mapping either.

(ii) *Power analysis of the RSTEP-LRT mapping method*

Four duplicate markers M16, M27, M67 and M76 were first deleted (Fig. 1, Table 2). Then the genotypic values of the 65 CSS lines plus the recurrent parent Asominori were used for mapping QTLs using the 10 putative QTLs in Table 1 (Fig. 2). The LOD score is used in QTL mapping to declare the existence of a QTL at a testing position. A higher LOD score means higher detection power. For single-marker analysis the average LOD score across the 100 simulations on each chromosome segment was below 3·10 (Fig. 2*A*), which means a low power would be observed. Under the LOD threshold 2·5, the power of single-marker analysis was 0·74 for the largest QTL on M38, and 0·56 for the second largest QTL on M21. The power was rather low for other QTLs. In addition, the false QTL on M37 had a power of 0·33 (Fig. 2*B*). These results indicate the inappropriateness of using single-marker analysis for a non-idealized CSS population.

The proposed RSTEP-LRT mapping method can significantly increase the LOD score and improve the mapping power (Fig. 2*C*, *D*). The average LOD scores across 100 simulations on chromosome segments where putative QTLs were located were over 3·0 for all QTLs except QTL1, QTL3 and QTL7 (each explains less than 3 % of the phenotypic variance) (Fig. 2*C*). Under the LOD threshold 2·5, the power for RSTEP-LRT to identify the two largest QTLs (i.e. QTL8 on M38, QTL5 on M21) was near 1·00, and the power to identify the smallest QTL (QTL1 at M3) was 0·25. Some false QTLs may occur,

https://doi.org/10.1017/S0016672306008408 Published online by Cambridge University Press

Table 2. *Marker deletion process for decreasing multicollinearity in the CSS population derived from the two rice parents,* japonica *Asominori and* indica *IR24*

Step	Condition no.	Two markers with the highest correlation				Correlation coefficient (r)	Marker deleted
		First marker	Lines	Second marker	Lines		
1	Infinity	M14	3	M16	3	1	M16
2	Infinity	M26	2	M27	2	1	M27
3	Infinity	M66	2	M67	2	1	M67
4	Infinity	M75	3	M76	3	1	M76
5	Infinity	M60	4	M61	5	0·8872	M61
6	Infinity	M7	4	M8	3	0·8591	M7
7	Infinity	M37	4	M38	3	0·8591	M37
8	Infinity	M74	4	M75	3	0·8591	M74
9	Infinity	M48	8	M49	6	0·8515	M48
10	Infinity	M12	5	M13	7	0·8312	M13
11	Infinity	M4	2	M5	3	0·8101	M5
12	Infinity	M28	2	M29	3	0·8101	M29
13	Infinity	M52	3	M53	2	0·8101	M52
14	Infinity	M66	2	M68	3	0·8101	M68
15	Infinity	M72	3	M73	2	0·8101	M72
16	Infinity	M73	2	M75	3	0·8101	M75
17	Infinity	M57	3	M58	5	0·7622	M58
18	Infinity	M64	3	M65	5	0·7622	M65
19	Infinity	M22	7	M23	4	0·7374	M22
20	Infinity	M23	4	M24	4	0·7339	M24
21	6021	M31	5	M32	6	0·7062	M32
22	1819	M55	6	M56	5	0·7062	M55
23	1766	M19	1	M20	2	0·7016	M20
24	1725	M33	4	M34	2	0·6960	M33
25	1394	M2	6	M3	3	0·6901	M2
26	1340	M35	3	M36	6	0·6901	M36
27	1293	M14	3	M15	3	0·6508	M15
	758						

especially for M73, where the probability of a false positive was 0·12.

Under a reduced multicollinearity among markers, the RSTEP-LRT mapping method can further increase the LOD score and improve the mapping power (Fig. 2*E*, *F*). The average LOD scores across 100 simulations on chromosome segments where putative QTLs were located were over 3·0 for all QTLs except the smallest QTL, i.e. QTL1 (explains 1·41% of the phenotypic variance) (Fig. 2*E*). Under the LOD threshold 2·5, the power for RSTEP-LRT to identify the first six largest QTLs was above 0·80 (i.e. 0·91 for a QTL on M17, 1·00 for QTLs on M21, M26 and M38, and 0·82 for QTLs on M41 and M44). The power to identify the smallest QTL (QTL1 at M3) was 0·37, and the probability of a false QTL on M73 was reduced from 0·12 to 0·02.

(iii) *Threshold LOD score of QTL mapping with CSS lines*

A problem common to QTL mapping methods is the difficulty of determining the threshold LOD score, or likelihood ratio (LR) test (Lander & Botstein, 1989; Zeng, 1994; Churchill & Doerge, 1994). A LOD threshold between 2 and 3 was proposed in Lander & Botstein (1989) to ensure an overall false positive rate of 5%. A permutation test was recommended by Churchill & Doerge (1994) to determine an appropriate threshold for the experimental data at hand. When permutation tests are applied in RSTEP-LRT (Fig. 3), little difference can be identified in the LOD score distribution across the eight environments and two QTL mapping models (Fig. 3*A*, *B*). The probability that the LOD score is over 2·5 was lower than 0·05 in any environment (Fig. 3).

For a specific mapping population, higher threshold values of LODs result in a lower power, and therefore QTLs with small effects cannot be readily identified. Lower thresholds result in higher powers but also suffer from higher false positive rates. As pointed out by Dudley (1993), the appropriate significance level to use depends on the purpose of the mapping experiment. If QTL mapping is performed with the eventual goal of cloning QTLs (Frary *et al*., 2000) or introgressing a few QTLs with large effects, a stringent threshold of LOD such as 3·00 should be used for QTL mapping with a CSS

https://doi.org/10.1017/S0016672306008408 Published online by Cambridge University Press

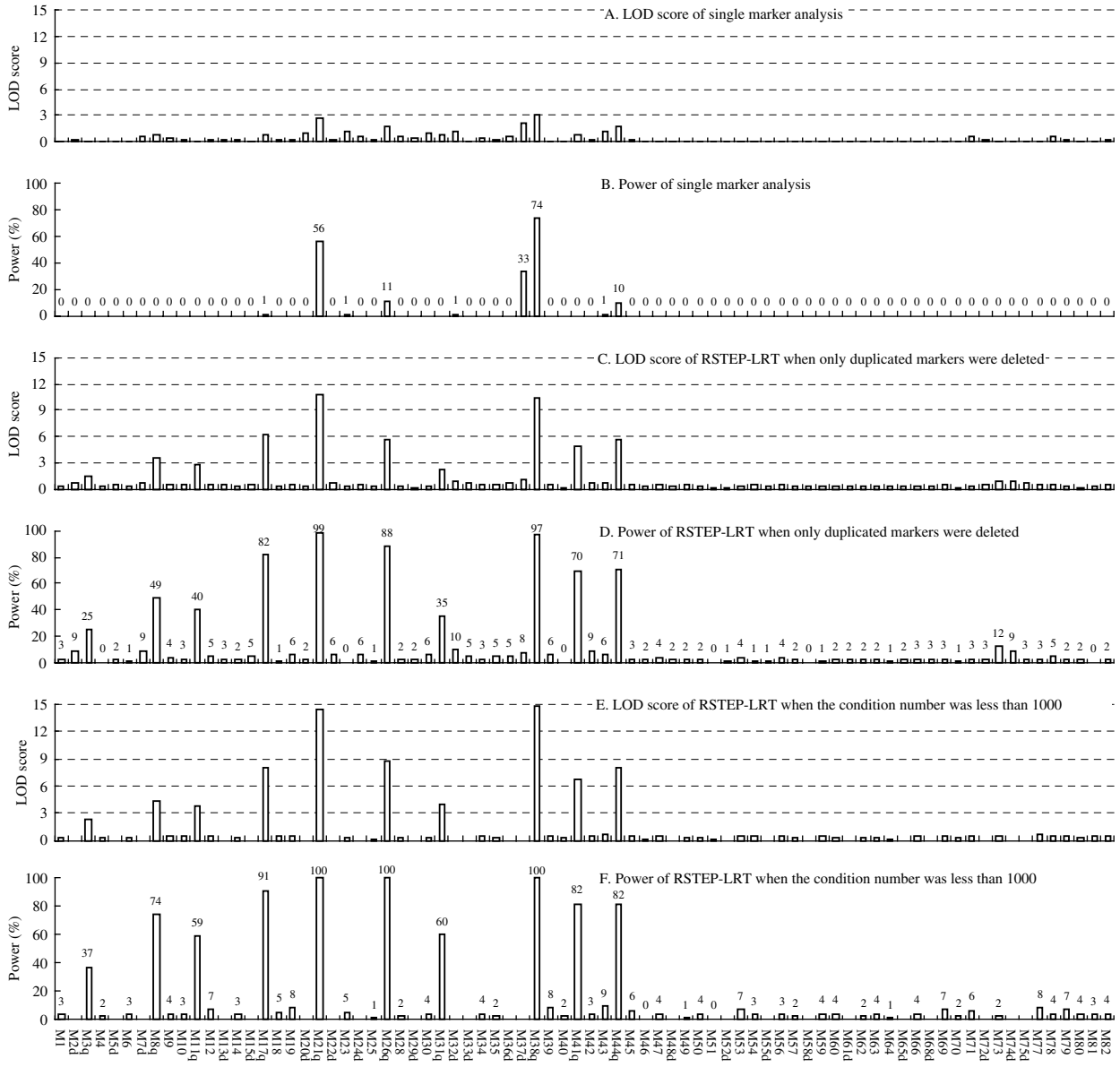

Fig. 2. LOD score and power of the QTL mapping method RSTEP-LRT from a simulation study. The LOD threshold of 2·5 was used to declare the presence of a QTL. Markers with 'd' were deleted to reduce the multicollinearity, and markers with 'q' were associated with putative QTLs.

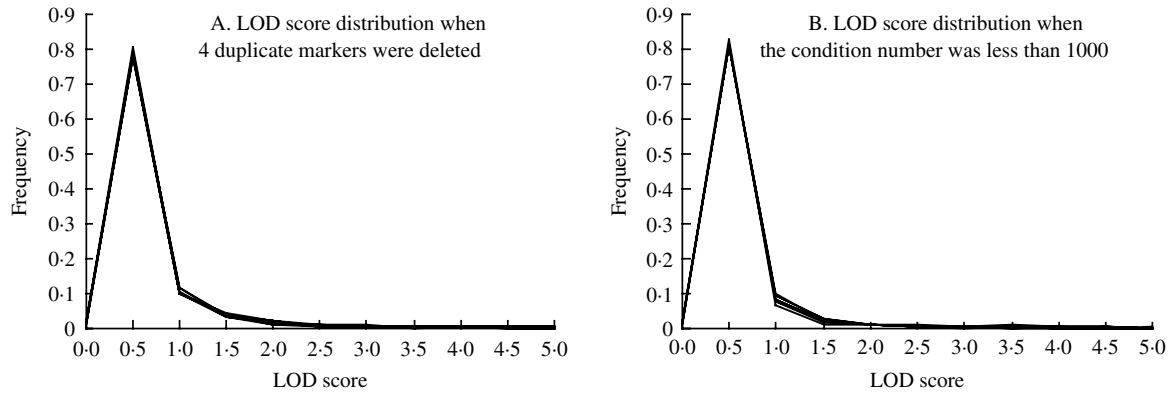

Fig. 3. Frequency distributions of LOD score in eight environments calculated from permutation tests.

https://doi.org/10.1017/S0016672306008408 Published online by Cambridge University Press

population. But if the goal is to exploit QTL information in marker-assisted selection for a complex trait, a less stringent LOD threshold such as 2·0 may be appropriate, as the false positive will have very limited influence on the results from marker-assisted selection. When we conducted the QTL mapping for grain length, the threshold LOD of 2·5 was applied.

(iv) *QTLs for grain length from the CSS population derived from* japonica *Asominori and* indica *IR24*

The analysis of variance (ANOVA) shows that there is a significant difference in grain length among the background parent Asominori and the 65 lines ($P<0{\cdot}01$). The estimate of the genetic variance among the 66 lines was 0·0339, and the heritability in the broad sense was estimated as 0·85. The environmental effects were also significant ($P<0{\cdot}0001$), but the genotype by environment interaction was insignificant ($P=0{\cdot}30$).

When a LOD threshold of 2·5 was used and the condition number among markers was less than 1000, a total of 18 chromosome segments demonstrated the existence of QTLs for grain length. These QTLs are distributed on eight of the 12 rice chromosomes. The three QTLs on M3, M23 and M34 were detected in all eight environments (Table 3). The QTL on M23 has the highest LOD score and explains the largest variation in all environments. The LOD score of M23 ranged from 14·41 in E6 to 21·17 in E2, and the percentage of variance explained by M23 from 28·91 in E6 to 47·60 in E1. The allele from IR24 at this locus constantly increases the grain length in all environments, which indicates this allele has a high stability. The QTL on M23 has been confirmed and fine mapped in Wan *et al*. (2006).

For the QTL on M3, the LOD score ranged from 2·95 in E8 to 12·91 in E3, and the percentage of variance explained from 3·40 in E8 to 14·49 in E3. For the QTL on M34, the LOD score ranged from 7·26 in E1 to 15·62 in E4, and the percentage of variance explained from 8·22 in E2 to 20·33 in E5. The two alleles from IR24 at both loci constantly decrease the grain length in all environments.

QTLs on other chromosome segments were detected in some environments but not others. These QTLs, which have smaller effects than the three QTLs on M3, M23 and M34, can be called minor QTLs. Interestingly, some minor QTLs also have high stability, i.e. constantly increasing or decreasing the grain length. For example, QTLs on M10 and M18 have negative effects, and QTLs on M51 and M80 have positive effects in all environments (Table 3). These minor QTLs should also be considered in gene pyramiding to maximize the genetic gain.

Asominori is a short grain variety, while IR24 is a long grain variety (Wan *et al*., 2006). The QTL on M23 with the largest effect is the major reason why IR24 has a long grain. However, some QTLs have negative effects on grain length, which means the short-grain parent Asominori also has alleles that could increase grain length, such as those on M3 and M34. This explains the transgressive segregation in grain length in the 65 CSS lines and recombination inbred lines derived from Asominori and IR24.

4. Discussion

CSS lines have the potential for QTL fine mapping and map-based cloning (Frary *et al*., 2000; Wan *et al*., 2006). But CSS lines with more than one chromosome substitution segment make it impossible to locate a QTL on a single chromosome segment through the comparison of the trait performance between one CSS line and the background parent. We present a likelihood ratio test for QTL detection using CSS lines. Three steps are needed in the analysis: (1) detecting multicollinearity and deleting redundant markers; (2) performing marker selection using stepwise regression; and (3) conducting a likelihood ratio test to declare statistical significance for each marker. Advantages of the proposed method over individual marker analysis were demonstrated using both simulated data and data from 65 CSS lines of rice (Table 3, Fig. 2).

Multicollinearity occurs when using regression analysis of trait performance on chromosome segments. One option for removing multicollinearity is to delete redundant markers. In this study, we propose deleting the most correlated markers to decrease the multicollinearity among markers. The decrease in multicollinearity increases the mapping power but has one disadvantage, i.e. the QTLs on deleted markers cannot be identified. But the correlation between a deleted marker and a retained marker showing evidence of QTLs can be used as the basis for a conjecture about whether the deleted marker is associated with a QTL. For example, one QTL on M26 was identified in environments E1, E2, E4 and E6 (Table 3). M27 is a duplicate marker of M26, and was not included in QTL mapping. Strictly speaking, if a QTL was associated with M27 but not with M26, the QTL would be mapped on M26. Without additional information it is impossible to know whether the QTL is on M26, on M27, or on both markers. However, if two additional CSS lines can be derived, one with M26 and the other with M27, this problem may be solved.

CIM (Zeng, 1994) was previously used due to the lack of a suitable mapping method with non-idealized CSS lines (Wan *et al*., 2004, 2005). Using the same dataset, CIM identified the two major stable QTLs for grain length associated with M3 and M24 (see table 4 in Wan *et al*., 2005), but missed the major stable QTL associated with M34 (Table 3). As pointed out before,

Table 3. *QTLs for grain length in the CSS population derived from the two rice parents,* japonica *Asominori and* indica *IR24*

Chromosome		1		2		3			4		6			7		11	12		
Donor segment		M3	M10	M17	M18	M23	M25	M26	M30	M34	M42	M45	M46	M50	M51	M73	M78	M80	M82
LOD[a]	E1	**7·08**	1·64	0·04	2·38	**18·76**	**3·08**	**3·30**	0·03	**7·26**	1·61	0·04	0·33	0·05	0·18	1·98	**2·68**	0·20	0·05
	E2	**8·00**	0·38	2·49	0·11	**21·17**	**9·90**	**3·09**	0·01	**8·69**	0·00	**6·67**	**6·58**	0·06	0·65	2·08	0·09	0·44	0·00
	E3	**12·91**	2·23	**3·74**	0·48	**20·08**	**6·90**	0·53	0·00	**12·80**	0·05	**6·62**	**7·67**	**4·20**	0·32	0·56	0·26	0·62	0·04
	E4	**7·66**	0·14	0·00	**5·97**	**29·37**	**4·66**	**2·83**	**4·28**	**15·62**	**2·90**	0·22	0·07	0·04	0·64	**3·48**	**4·46**	**5·42**	**11·59**
	E5	**4·90**	1·24	0·06	**3·87**	**15·26**	0·08	2·03	2·24	**9·75**	0·72	0·55	0·04	0·00	0·69	1·25	**4·25**	**6·71**	**7·30**
	E6	**6·58**	**6·66**	0·02	**2·63**	**14·41**	1·48	**3·24**	0·95	**7·65**	0·02	0·27	0·12	0·06	0·61	2·37	1·79	0·63	1·15
	E7	**6·26**	**4·72**	0·02	2·15	**14·64**	0·02	0·74	0·25	**10·20**	0·38	0·65	0·06	0·00	**2·95**	0·38	1·64	0·98	0·37
	E8	**2·95**	**2·89**	0·35	0·30	**19·41**	**7·18**	0·95	0·00	**8·28**	0·02	**2·71**	1·36	0·19	**4·04**	2·21	0·09	0·29	0·34
ADD[b]	E1	**−0·16**	−0·12	−0·01	−0·09	**0·29**	**0·17**	**−0·13**	0·01	**−0·20**	0·05	0·01	0·05	−0·01	0·02	−0·10	**−0·07**	0·04	−0·01
	E2	**−0·13**	−0·04	−0·06	−0·01	**0·24**	**0·25**	**−0·09**	0·00	**−0·16**	0·00	**−0·07**	**0·19**	0·01	0·03	−0·07	0·01	0·04	0·00
	E3	**−0·17**	−0·10	**−0·08**	−0·03	**0·22**	**0·19**	−0·03	0·00	**−0·21**	−0·01	**−0·08**	**0·20**	**0·07**	0·02	−0·04	−0·01	0·05	−0·01
	E4	**−0·11**	−0·02	0·00	**−0·09**	**0·29**	**0·14**	**−0·07**	**−0·08**	**−0·22**	**0·04**	0·01	−0·01	−0·01	0·03	**−0·08**	−0·06	0·15	−0·15
	E5	**−0·15**	−0·12	−0·02	**−0·13**	**0·28**	−0·03	−0·11	−0·10	**−0·27**	−0·04	−0·03	0·02	0·00	0·05	−0·09	**−0·10**	**0·31**	**−0·19**
	E6	**−0·15**	**−0·25**	−0·01	**−0·09**	**0·22**	0·11	**−0·12**	−0·05	**−0·20**	−0·01	−0·02	0·03	−0·01	0·04	−0·10	**−0·05**	**0·07**	**−0·05**
	E7	**−0·15**	**−0·22**	−0·01	−0·08	**0·24**	0·01	−0·06	−0·03	**−0·25**	−0·02	−0·03	0·02	0·00	**0·10**	−0·04	−0·05	0·09	−0·03
	E8	**−0·09**	**−0·14**	0·03	−0·03	**0·27**	**0·25**	−0·06	0·00	**−0·19**	−0·01	**−0·06**	0·10	−0·02	**0·10**	−0·09	0·01	0·04	−0·03
PVE[c]	E1	**11·28**	2·14	0·05	3·10	**47·60**	**4·21**	**4·62**	0·03	**11·79**	2·13	0·05	0·42	0·05	0·22	2·66	**3·70**	0·25	0·07
	E2	**7·34**	0·26	1·69	0·07	**33·19**	**9·78**	**2·37**	0·01	**8·22**	0·00	**5·34**	**5·67**	0·04	0·46	1·54	0·05	0·30	0·00
	E3	**14·49**	1·67	**3·00**	0·34	**30·42**	**6·21**	0·38	0·00	**14·48**	0·03	**5·91**	**6·83**	**3·42**	0·22	0·40	0·19	0·43	0·03
	E4	**4·64**	0·06	0·00	**3·28**	**43·96**	**2·52**	**1·43**	**2·27**	**12·95**	**1·46**	0·10	0·03	0·02	0·30	**1·81**	**2·35**	**3·01**	**8·16**
	E5	**9·71**	2·16	0·10	**7·42**	**43·42**	0·12	3·64	4·03	**20·33**	1·00	0·73	0·07	0·00	1·17	2·16	**8·25**	**14·22**	**15·71**
	E6	**9·69**	**9·59**	0·02	**3·28**	**28·91**	1·82	**4·23**	1·14	**11·39**	0·03	0·31	0·14	0·07	0·72	3·00	2·10	0·75	1·30
	E7	**10·67**	**7·60**	0·03	3·07	**34·41**	0·02	1·04	0·34	**18·74**	0·44	0·83	0·08	0·00	**4·42**	0·52	2·35	1·36	0·51
	E8	**3·40**	**3·25**	0·33	0·29	**42·61**	**9·69**	1·02	0·00	**11·61**	0·02	**3·06**	1·47	0·20	**4·85**	2·49	0·09	0·31	0·36

[a] LOD score, in bold type when over 2·5.
[b] Additive effect, in bold type when the corresponding LOD score is over 2·5.
[c] Percentage of phenotypic variance explained by the QTL, in bold type when the corresponding LOD score is over 2·5.

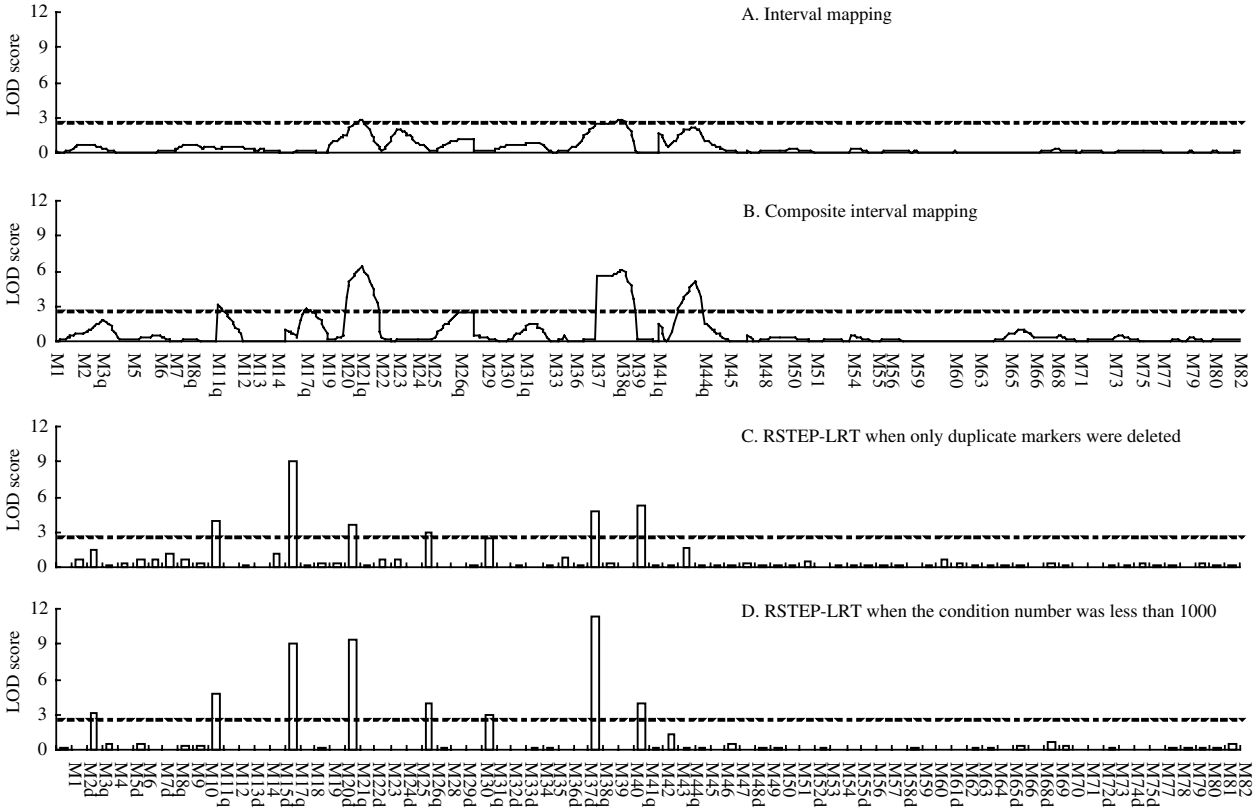

Fig. 4. LOD scores from QTL mapping methods – IM (*A*), CIM (*B*) and RSTEP-LRT (*C*, *D*) – using one simulation dataset. For clarity, not all markers are shown on the *x*-axis in (*B*). In (*D*), markers with 'd' were deleted to reduce the multicollinearity, and markers with 'q' were associated with putative QTLs.

traditional IM and CIM are not suitable for CSS populations. To further illustrate this point, we applied different mapping methods on a simulated phenotypic dataset (Fig. 4), where the putative QTLs were the same as in Table 1. IM and CIM were implemented by Cartograher 2·5 (Wang *et al*., 2005), and RSTEP-LRT was implemented by an in-house program called MappingQTL (written in Fortran 90/95, freely available). Under the LOD threshold of 2·5, IM found two QTLs around M21 and M38 (Fig. 4*A*), and CIM found 5 QTLs around M11, M17, M21, M38 and M44 (Fig. 4*B*). For a CSS population, the objective of QTL mapping is to identify the introgressed segments having QTLs for the trait of interest. It does not make much sense to map any QTL between two segments where IM or CIM was used. Six QTLs were detected using RSTEP-LRT after four duplicate markers were deleted (Fig. 4*C*, Table 2). When the condition number was reduced to 758 (Table 2), RSTEP-LRT identified eight QTLs (Fig. 4*D*). Additionally, it was clearly shown in Fig. 4*D* that markers associated with putative QTLs have significantly higher LOD scores but those not associated with QTLs have rather lower LOD scores, which indicates that using RSTEP-LRT will increase the mapping power but is less likely to result in more false positives.

It is not our intent to encourage the use of non-idealized CSS lines in QTL mapping. On the contrary, we prefer the idealized CSS lines, where simple standard methods such as ANOVA and the *t*-test can be readily applied to map the additive QTLs. One can also study the dominance effects and levels of epistasis by crossing one or more CSS lines with the background parent, or crossing two or more idealized CSS lines. Regarding gene pyramiding using marker-assisted selection, non-idealized CSS lines may have some advantages over idealized CSS lines if the mapped QTLs have been confirmed in other mapping populations, such as the secondary mapping population between a CSS line and the recurrent parent (Wan *et al*., 2006). Based on QTL mapping, for example, we may need to pyramid four donor chromosome segments in one genotype. If one CSS line has two segments and one has the other two, a single cross between the two lines will produce the required genotype. On the contrary, if each line has only one distinct donor segment, a double cross is needed to pyramid the four segments. Obviously, a single cross is more feasible and has a higher frequency of the target genotype than a double cross.

In the example CSS population, QTL mapping using the proposed RSTEP-LRT method identified a total of 18 segments on eight of the 12 rice

https://doi.org/10.1017/S0016672306008408 Published online by Cambridge University Press

chromosomes that harboured QTLs affecting grain length under the LOD threshold of 2·5. Three major stable QTLs detected in all eight environments were located on chromosome segments represented by M3, M23 and M34. The allele from IR24 on M23 constantly increases the grain length in all environments, but the alleles from IR24 on M3 and M34 constantly decrease the grain length (Table 3). Some minor QTLs were not detected in all environments, but they could increase or decrease the grain length constantly (Table 3). These QTLs are useful in breeding once the major QTLs have been fixed.

Due to the lack of significant genotype by environment interaction on grain length in this study, we conduct QTL mapping in each environment separately. But it would be interesting to include the QTL by environment interaction in model (1) for traits demonstrating significant genotype by environment interactions. This is being investigated for traits in rice such as grain width and protein contents where significant genotype by environment interaction has been detected.

The authors wish to thank Professor A. Yoshimura, Kyushu University, Japan for providing the CSS lines, and two anonymous reviewers for their useful comments on an earlier version of the manuscript. This work was supported by the Generation and HarvestPlus Challenge Programs of the Consultative Group for International Agricultural Research (CGIAR) (http://www.generationcp.org), and the National 973 Projects of China (No. 2006CB101700).

References

Alpert, K. B. & Tanksley, S. D. (1996). High-resolution mapping and isolation of a yeast artificial chromosome contig containing *fw2.2*: a major fruit weight quantitative trait locus in tomato. *Proceedings of National Academy of Sciences of the USA* **93**, 15503–15507.

Beaumont, M. A. & Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Review Genetics* **5**, 251–261.

Belknap, J. K. (2003). Chromosome substitution strains: some quantitative considerations for genome scans and fine mapping. *Mammalian Genome* **14**, 723–732.

Bogdan, M., Ghosh, J. K. & Doerge, R. W. (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* **167**, 989–999.

Churchill, G. A. & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Cowley, A. W. Jr, Roman, R. J. & Jacob, H. J. (2003). Application of chromosome substitution techniques in gene-function discovery. *Journal of Physiology* (*London*) **554**, 46–55.

Dudley, J. W. (1993). Molecular markers in plant improvement: manipulation of genes affecting quantitative traits. *Crop Science* **33**, 660–668.

Eshed, Y. & Zamir, D. (1995). An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. *Genetics* **141**, 1147–1162.

Frary, An., Nesbitt, T. C., Frary, Am., Grandillo, S., Knaap, E. V. D., Cong, B., Liu, J. P., Meller, J., Elber, R., Alpert, K. B. & Tanksley, S. D. (2000). *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. *Science* **289**, 85–88.

Hackett, C. A. (1997). Model diagnostics for fitting QTL models to trait and marker data by interval mapping. *Heredity* **79**, 319–328.

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

Jansen, R. C. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**, 1447–1455.

Knott, S. A. & Haley, C. S. (1992). Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genetical Research* **60**, 139–151.

Kubo, T., Nakamura, K. & Yoshimura, A. (1999). Development of a series of Indica chromosome segment substitution lines in Japonica background of rice. *Rice Genetics Newsletter* **16**, 104–106.

Kubo, T., Aida, Y., Nakamura, K., Tsunematsu, H., Doi, K. & Yoshimura, A. (2002). Reciprocal chromosome segment substitution series derived from Japonica and Indica cross of rice (*Oryza sativa* L.). *Breeding Science* **52**: 319–325.

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

Martinez, O. & Curnow, R. N. (1992). Estimating the locations and the sizes of the effects of quantitative loci using flanking markers. *Theoretical and Applied Genetics* **85**, 480–488.

Myers, R. H. (1990). *Classical and Modern Regression with Applications*, 2nd edn. Pacific Grove, CA: Duxbury Thomson Learning.

Nadeau, J. H., Singer, J. B., Martin, A. & Lander, E. S. (2000). Analysis complex genetics traits with chromosome substitution strains. *Nature Genetics* **24**, 221–225.

Satagopan, J. M., Yandell, B. S., Newton, M. A. & Osborn, T. C. (1996). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **144**, 805–816.

Sen, S. & Churchill, G. A. (2001). A statistical framework for quantitative trait mapping. *Genetics* **159**, 371–387.

Tanksley, S. D. & Nelson, J. C. (1996). Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theoretical and Applied Genetics* **92**, 191–203.

Tsunematsu, H., Yoshimura, A., Harushima, Y., Nagamura, Y., Kurata, N., Yano, M., Sasaki, T. & Iwata, N. (1996). RFLP framework map using recombinant inbred lines in rice. *Breeding Science* **46**, 279–284.

Wan, X.-Y., Wan, J.-M., Su, C.-C., Wang, C.-M., Shen, W.-B., Li, J.-M., Wang, H.-L., Jiang, L., Liu, S.-J., Chen, L.-M., Yasui, H. & Yoshimura, A. (2004). QTL detection for eating quality of cooked rice in a population of chromosome segment substitution lines. *Theoretical and Applied Genetics* **110**, 71–79.

Wan, X.-Y., Wan, J.-M., Weng, J.-F., Jiang, L., Bi, J.-C., Wang, C.-M. & Zhai, H.-Q. (2005). Stability of QTLs for rice grain dimension and endosperm chalkiness characteristics across eight environments. *Theoretical and Applied Genetics* **110**, 1334–1346.

Wan, X.-Y., Wan, J.-M., Jiang, L., Wang, J.-K., Zhai, H.-Q., Weng, J.-F., Wang, H.-L., Lei, C.-H., Wang, J.-L., Zhang, X., Cheng, Z.-J. & Guo, X.-P. (2006). QTL analysis for rice grain length and fine mapping of an identified QTL with stable and major effects. *Theoretical and Applied Genetics* **112**, 1258–1270.

Wang, S. G. & Chow, S. C. (1994). *Advanced Linear Models*. New York: Marcel Dekker.

Wang, S., Basten, C. J. & Zeng, Z.-B. (2005). *Windows QTL Cartographer 2.5*. Raleigh, NC: Department of Statistics, North Carolina State University.

Whittaker, J. C., Thompson, R. & Visscher, P. M. (1996). On the mapping of QTL by regression of phenotype on marker-type. *Heredity* **77**, 23–32.

Wissuwa, M., Wegner, J., Ae, N. & Yano, M. (2002). Substitution mapping of Pup1: a major QTL increasing phosphorus uptake of rice from phosphorous-deficient soil. *Theoretical and Applied Genetics* **105**, 890–897.

Xu, S. (2003). Estimating polygenic effects using markers of entire genome. *Genetics* **163**, 789–801.

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