Theor Appl Genet (2009) 119:459–470
DOI 10.1007/s00122-009-1052-z

ORIGINAL PAPER

Population structure and association mapping on chromosome 7 using a diverse panel of Chinese germplasm of rice (*Oryza sativa* L.)

Weiwei Wen · Hanwei Mei · Fangjun Feng · Sibin Yu · Zhicheng Huang · Jinhong Wu · Liang Chen · Xiaoyan Xu · Lijun Luo

Received: 9 December 2008 / Accepted: 21 April 2009 / Published online: 21 May 2009

Abstract The majority of 170 rice accessions used in this study were diverse landraces or varieties from a putative mini-core collection of Chinese germplasm along with some widely used parental lines in genetic analysis or breeding (a few from abroad). The population was genotyped using 84 SSR or InDel markers on chromosome 7 and 48 markers on other chromosomes. The phenotyping of heading date, plant height and panicle length were carried out in different locations for 2 years. Based on morphological characterization, distance-based clustering and model-based estimation of marker data, the population showed a predominant structure with two subpopulations in correspondence with *indica* and *japonica* subspecies. The estimation of linkage disequilibrium in 2 Mb windows varied along chromosome 7 and showed parallel changes with inter-subspecies differentiation of marker loci (*Fst*). Based on the mixed linear model considering population structure and family relatedness [i.e. the ($Q + K$) model], one to three associated markers ($P \leq 0.0001$) per trait per experiment were scanned out on rice chromosome 7. Most significant loci were repeated for the data from both field experiments while two loci were associated with two or three traits. Marker-based allelic effects were shown in a couple of associated markers as examples. The application of association results in breeding program was also discussed.

Introduction

As an important stable crop all over the world and the model species of cereal plants in genomic researches, rice has great merits as the target crop of association mapping based on the public data of genome sequences together with a wealth of genetic diversity in natural populations of germplasm. In past 10 years, core collections were developed in China for several crops including rice, maize, soybean and wheat, based on geological distribution, characteristic data and molecular markers (Li et al. 2002, 2008; Hao et al. 2008). Wang et al. (2008) analyzed the population structure and linkage disequilibrium (LD) of a mini-core set of maize inbred lines in China. In rice, the core collection consisted of about 5% samples from more than 50,000 accessions in the whole set of germplasm. Furthermore, a mini-core collection of rice germplasm was developed which had 0.5% accessions from the whole set and hosted about 66% genetic diversity (Li et al., personal communication). Xu et al. (2004) suggested the potential use of 32 samples from 236 US rice accessions for developing core collection and providing sufficient genetic diversity. Association mapping based on core collections of germplasm provides a valuable alternative to gene mapping approaches based on bi-parental populations (Holland 2007; Hasan et al. 2008).

Communicated by J. Yu.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-009-1052-z) contains supplementary material, which is available to authorized users.

W. Wen · H. Mei · S. Yu · Z. Huang · L. Luo
Huazhong Agricultural University, 430070 Wuhan, China

W. Wen · H. Mei (✉) · F. Feng · Z. Huang · J. Wu · L. Chen · X. Xu · L. Luo (✉)
Shanghai Agrobiological Gene Center, 201106 Shanghai, China
e-mail: hmei@sagc.org.cn

L. Luo
e-mail: lijun@sagc.org.cn

Association mapping, or LD mapping, has been practiced in a number of plant species (Thornsberry et al. 2001; Kraakman et al. 2004; Aranzana et al. 2005; Mazzucato et al. 2008; Zhu et al. 2008). Association mapping has the potential of simultaneous discovery of gene loci responsible for multiple traits with no need to develop permanent segregating populations. A higher proportion of polymorphic molecular markers could provide better genome coverage than any bi-parental population. As association mapping exploits the historical recombination events that have occurred during establishment of the sample population, higher mapping resolution could be obtained than that possible in small bi-parental experimental crosses (Flint-Garcia et al. 2005). This strategy has an attractive advantage in the ability to detect the comparative effects of multiple alleles at each genetic locus that exists in crop germplasm.

For many important crops, the complex breeding histories have resulted in complicated population structures within the germplasm (Flint-Garcia et al. 2003). A drawback of association mapping in rice is the possibility of false-positive results due to high levels of population structure caused by the differentiation of subspecies, along with diverse levels of pedigree relations in modern varieties. Population structure can be quantified based on Bayesian method which assigns individuals to subpopulations (Q matrix) using unlinked markers (Pritchard et al. 2000b). Relationships between individuals can be detected by marker-based estimation of the probability of identity by descent between individuals (Yu et al. 2006).

Zhang et al. (2005) successfully conducted whole-genome association analysis between microsatellite markers and multiple agronomic traits using discriminant analysis (DA) in 218 inbred lines of rice. Recently, Iwata et al. (2007) associated RFLP markers with the width and length of milled rice grains in a set of 332 rice germplasm using Bayesian method. Agrama et al. (2007) used the mixed linear model (MLM) method to disclose the associations between 123 SSR markers and yield components in rice cultivars.

In this study, the authors analyzed the genetic structure in a set of Chinese rice germplasm, including landraces and varieties from a putative mini-core collection, together with some parental lines widely used in genetic experiments (a few from abroad). Associated markers for three agronomic traits were scanned along rice chromosome 7 as an example. This chromosome was selected as the first one to be investigated mostly based on random draw. The authors are also interested in the phenomenon observed in two reciprocal introgression line (IL) populations between *indica* and *japonica* varieties. Along the long arm of chromosome 7, proportion of introgression segments is much higher from *indica* variety as a donor to *japonica* variety as a receptor than that in the reciprocal direction (unpublished data). The variation of LD over genomic regions had been widely reported in several crops (Remington et al. 2001; Caldwell et al. 2006; Hyten et al. 2007; Mather et al. 2007; Somers et al. 2007; Wang et al. 2008). It was observed to have relationship with the variation of recombination rate (Mather et al. 2007) and regarded as an indicator of genomic regions under selection pressure (Somers et al. 2007). So the authors tried to detect the variations in LD, marker polymorphism and subspecific differentiation along chromosome 7. The relationships among them and their influence on association mapping effects were analyzed and discussed based on the recent results.

Materials and methods

Plant materials

A diverse panel of 170 rice accessions was used in this study. Most accessions are landraces from a putative mini-core collection of rice germplasm in China, together with rice varieties widely used as parents of genetic populations, including the two sequenced varieties *Nipponbare* and *93-11*. A minority of these accessions were landraces and cultivars from other countries, such as Bangladesh (1), Brazil (1), Cote d'Ivoire (2), Cuba (1), Guiana (1), India (1), Japan (2), Nigeria (2), South Korea (1), Sri Lanka (1) and USA (2) (Supplementary materials, Table S1).

Field experiments and phenotyping

The field experiments were conducted twice in the experimental farms of Shanghai Academy of Agricultural Sciences (SAAS) at Chonggu and Baihe in Qingpu District, Shanghai in the summer seasons of 2006 and 2007, respectively. All materials were sown in nurseries in late May, and transplanted into four-row plots with a space of 25 × 20 cm in the field 25 days after sowing.

The Cheng's index for *indica–japonica* differentiation was investigated after heading in 2006, based on the observations of six morphological traits differentiated between the *indica* and *japonica* subspecies in rice, including the glume pubescence, glume color at heading time, length of rachis between the first and second panicle nodes, leaf pubescence, length/width ratio of grains and phenol reactions. Each trait was evaluated and given scores within 0–4 while *indica* varieties have lower values and *japonica* varieties vice versa (Cheng 1985).

Three agronomic traits, heading date (HD), plant height (PH) and panicle length (PL), were surveyed in both 2006 (HD06, PH06, PL06) and 2007 experiments (HD07, PH07, PL07). HD was recorded as the duration from seed sowing to heading of about 10% plants in each plot. Three plants were randomly selected from the central area of each plot for evaluating PH and PL. PH was measured in centimeters from the soil surface to the tip of the tallest panicle (excluding awns). The PL was measured in centimeters from the panicle neck to the panicle tip (excluding awns).

Genotyping

Genomic DNA was extracted from single plant of each accession by the CTAB method (Colosi and Schaal 1993). PCR amplifications were conducted following a popular protocol (Chen et al. 1997) with different annealing temperatures if specified for certain primers. The amplification products were separated by 6% polyacrylamide gel electrophoresis (PAGE). We used 126 SSR and 6 InDel markers distributed throughout the 12 chromosomes. Among them, 84 molecular markers were selected from chromosome 7 while other 48 markers were from the rest ones. The primer sequence and chromosomal position of each marker were obtained from GRAMENE and NCBI databases (http://www.gramene.org; http://www.ncbi.nlm.nih.gov/). Detailed information of the markers was shown in supplementary materials (Table S2).

Data analysis

Phenotypic data

Nested ANOVA was conducted using S-Plus for Windows V6.1 (Insightful Corporation 2001) to evaluate the variance between two subspecies (Subsp), accessions included in each subspecies (Acc(Subsp)) and between the experiments in 2 years (Exp). The model for HD, e.g., was input as a formula (HD $\sim$ Subsp/Acc + Exp). *F* test for Subsp is based on the ratio between $\mathrm{MS}_{\mathrm{Subsp}}$ and $\mathrm{MS}_{\mathrm{Acc(Subsp)}}$. *F* values for Exp and Acc(Subsp) were calculated in proportion to mean square of residuals ($\mathrm{MS}_{\mathrm{Error}}$). Expected variances of Subsp, Acc, Exp and residual ($\sigma^2_{\mathrm{Subsp}}$, σ^2_{Acc}, σ^2_{Exp} and $\sigma^2_{\mathrm{Error}}$) were estimated based on the formula of $(\mathrm{MS} - m \times \mathrm{MS}_{\mathrm{Error}})/n$, where MS is the mean square of each source and *m* is the number of replication. The *n* value is the accession number for Exp, and replication number for Acc. As the accession numbers in two subspecies are not balanced, the *n* value for Subsp was calculated using the formula: $(1/(r-1)) \times ((N_{\mathrm{ind}} + N_{\mathrm{jap}}) - ((N_{\mathrm{ind}})^2 + (N_{\mathrm{jap}})^2)/(N_{\mathrm{ind}} + N_{\mathrm{jap}}))$, where *r* is the number of subgroups ($r = 2$), N_{ind} and N_{jap} are accession numbers in *indica* and *japonica* subgroup, respectively ($N_{\mathrm{ind}} = 104$; $N_{\mathrm{jap}} = 66$; *n* of Subsp $= 80.75294$). Broad sense heritability (H^2_{B}, %) was estimated from $(\sigma^2_{\mathrm{Subsp}} + \sigma^2_{\mathrm{Acc}})/(\sigma^2_{\mathrm{Subsp}} + \sigma^2_{\mathrm{Acc}} + \sigma^2_{\mathrm{Exp}} + \sigma^2_{\mathrm{Error}}) \times 100$. The variation explained by the population structure (i.e. among *indica* and *japonica* subgroups) was estimated from $\sigma^2_{\mathrm{Subsp}}/(\sigma^2_{\mathrm{Subsp}} + \sigma^2_{\mathrm{Acc}} + \sigma^2_{\mathrm{Exp}} + \sigma^2_{\mathrm{Error}}) \times 100$. Phenotypic correlation coefficients were also estimated.

Genotypic data

For each marker locus, alleles were called as A1, A2, to A(*n*) in ascending order of the amplified fragment size. We used PowerMarker (Liu and Muse 2005) to calculate observed heterozygosity, allele number and allele frequency. Although rice varieties were normally pure lines, rare cases of heterozygosity (less than 0.3%) were still observed for molecular markers. Heterozygous data points and null alleles were treated as missing data. Rare alleles and haplotypes, with frequency of less than 5% in the panel or selected groups of rice accessions, were omitted in further analysis. The rest alleles were regarded as effective alleles from that the polymorphic allele numbers were given by subtracting one.

Population structure

The genetic structure of all 170 samples was investigated with the model-based method implemented in STRUCTURE (Pritchard et al. 2000a). Fifty-two unlinked or loosely linked marker loci (i.e. 4 on chromosome 7 together with 48 on other chromosomes; mostly with physical distance larger than 1 Mb) were used to analyze the population structure. Hypotheses were tested for number of subpopulations ranging from $k = 1$ to $k = 10$. For each *k* value, six runs were performed using the admixture model and correlated allele frequencies (Falush et al. 2003). The burn-in length and iterations were all set to 500,000. In the model-based method, each accession was estimated to have memberships in multiple subgroups, shown as the membership coefficients (*Q* values). As a population with two subgroups (i.e. *indica* and *japonica*) was adopted after above procedure, the Q_{indica} value (abbreviated as *Q* value) was used in further analysis, having $(Q_{\mathrm{indica}} + Q_{\mathrm{japonica}}) = 1$.

According to the same set of marker data, a hierarchical cluster was built based on the genetic similarity of the accessions, using the NTSYS-pc software (UPGMA) (Sneath and Sokal 1973). The software SPAGeDi (Hardy and Vekemans 2002) was used to estimate the kinship coefficients (Loiselle et al. 1995) based on all 132 marker loci. All negative kinship values were set to zero. *Fst* value of each marker locus was estimated to show the differentiation between *indica* and *japonica* subspecies (Weir and Cockerham 1984).

Linkage disequilibrium

Linkage disequilibrium extents were estimated between markers on chromosome 7 within a scale of 2 Mb windows using the software package TASSEL2.0 (Bradbury et al. 2007), for the whole population and two subpopulations as identified by the model-based partition. Squared allele frequency correlations (r^2) were calculated after modifying the option setting for loci with multiple alleles.

Association analysis

Association between traits and markers was calculated using a MLM function based on ($Q + K$) method in TASSEL2.0 (Yu et al. 2006). The significant marker-trait associations were declared by $P \leq 0.0001$ with relative magnitude represented by the R^2 value as the portion of variation explained by the marker. Least significant differences (LSDs) were calculated to test the differences among the accession groups defined by alleles or their haplotypes of associated markers.

Results

Summary and ANOVA of phenotypic data

Large variations were observed among rice accessions for three traits, i.e. HD, PH and PL (Table 1). For example, the earliest accession took about 45 days to get heading while the latest line spent about 115 days. The maximum values were larger than the minimum values by 2–3 folds in all three traits. Highly significant correlations were observed among three agronomic traits. The correlation coefficients were 0.4982** between HD and PH, 0.3456** between HD and PL, and 0.5447** between PH and PL.

Nested ANOVA showed highly significant variation between *indica* and *japonica* subspecies in PL, but not significant in HD and PH. There were significant variation between experiments in 2006 and in 2007 in HD ($P = 0.0112$) and highly significant variations in PH ($P \leq 0.0001$), PL ($P = 0.0049$). The variations among accessions in subspecies are highly significant in all traits ($P \leq 0.0001$) (Table 2). Broad sense heritabilities of HD, PH and PL were 96.18, 94.82 and 89.40%, respectively.

So these traits were basic agronomic characters with large variances among germplasm, high heritability and correlations with each other, partial influence from experimental environments (years and locations), but less differentiation between *indica* and *japonica* subspecies. In this study, they served more as examples to check the effect of association mapping than as target traits.

Model-based and distance-based analyses of the population structure

The model-based method was performed using the data of 52 polymorphic loci to recognize the genetic structure among all the 170 samples. Six times of independent calculations were done for each k value from $k = 1$ to $k = 10$. The posterior probability (ln $P(D)$) increased sharply from $k = 1$ to $k = 2$, but slowly after $k = 2$ (Supplementary materials, Fig. S1). The results of six parallel calculations converged with each other only for $k = 1$, 2, and 4. But either for $k = 3$ or $k = 4$, the clustering structures among rice accessions varied among parallel calculations. In this case, the population structure with two subpopulations (i.e. $k = 2$) was chosen for the samples in this study.

The population structure based on the Q values with $k = 2$ coincided very well with the UPGMA tree from distance-based analysis (Fig. 1). A batch of 66 accessions was grouped into one clade shown in the upper part of the dendrogram while the other group had 104 accessions located in the lower part.

Two subpopulations parallel to *indica* and *japonica* subspecies

The rice accessions in this study were empirically classified into two groups (Supplementary materials, Table S1). The

Table 1 Descriptive statistics of three agronomical characters observed in 170 rice accessions in two field experiments (2006 and 2007, Shanghai, China)

Traits	Experiments	Subpopulations	Mean ± SD	Range
Heading date (HD, days)	2006	*indica*	79.2 ± 15.0	44.0–116.0
		japonica	82.7 ± 15.8	46.0–113.0
	2007	*indica*	79.6 ± 13.1	60.0–115.0
		japonica	85.0 ± 16.0	43.0–115.0
Plant height (PH, cm)	2006	*indica*	108.5 ± 28.2	57.0–167.7
		japonica	109.4 ± 26.3	55.0–170.0
	2007	*indica*	111.4 ± 28.5	64.8–190.0
		japonica	115.4 ± 26.6	66.0–177.3
Panicle length (PL, cm)	2006	*indica*	23.3 ± 2.8	16.8–31.0
		japonica	21.4 ± 3.9	12.3–32.8
	2007	*indica*	22.9 ± 2.3	16.0–26.9
		japonica	21.3 ± 3.6	14.0–32.8

Table 2 ANOVA results of heading date (HD), plant height (PH) and panicle length (PL) based on fixed effect model

Traits	Effects[a]	*df*	MS	*F* value	Pr(*F*)	H^2_B (%)[b]	Subsp (%)[c]
HD (day)	Subsp	1	1527.412	3.76[d]	0.0541	96.18	4.74
	Exp	1	97.509	6.59	0.0112		
	Acc(Subsp)	168	405.998	27.43	<0.0001		
	Residuals	154	14.802				
PH (cm)	Subsp	1	1969.460	0.55[d]	0.4607	94.82	1.83
	Exp	1	3037.550	66.01	<0.0001		
	Acc(Subsp)	158	3601.421	78.27	<0.0001		
	Residuals	682	46.014				
PL (cm)	Subsp	1	871.232	19.55[d]	<0.0001	89.40	41.72
	Exp	1	21.008	7.96	0.0049		
	Acc(Subsp)	166	44.561	16.89	<0.0001		
	Residuals	767	2.638				

[a] Subsp, Acc, Exp represent subspecies, accessions, and experiments, respectively

[b] H_B (%) represents the broad sense heritability, i.e. the percentage of the sum of variance of Subsp and Acc in total variance

[c] Subsp (%) represents the percentage of the variance of Subsp in total variance

[d] For Subsp, *F* value = $MS_{Subsp}/MS_{Acc(Subsp)}$

Cheng's indices (CI) varied from 3 to 23 among rice accessions with two densely distributed ranges of CI = 5–10 and 16–22 (Fig. 2). The *indica* group had 104 accessions with CI = 3–14, and the *japonica* group had 66 accessions with CI = 15–23.

The *Q* values of accessions in *indica* group varied from 0.820 to 0.999, but leaving a single accession "Qingke" as an exception ($Q = 0.693$). The *Q* values had a range of 0.001–0.268 for accessions in *japonica* group. So, it is obvious that classification of *indica* and *japonica* subspecies in rice based on the Cheng's index was highly corresponding with the population structure based on molecular markers. Arrayed by both Cheng's index and *Q* values, the accessions can be brought into two largely isolated clusters (Fig. 2). The phenotypic variation explained by the population structure, i.e. the percentage of variance between two subspecies (Subsp in Table 2) in total variance, was estimated to be 4.74, 1.83 and 41.72% in HD, PH and PL, respectively.

Variation of LD level in 2 Mb windows along rice chromosome 7

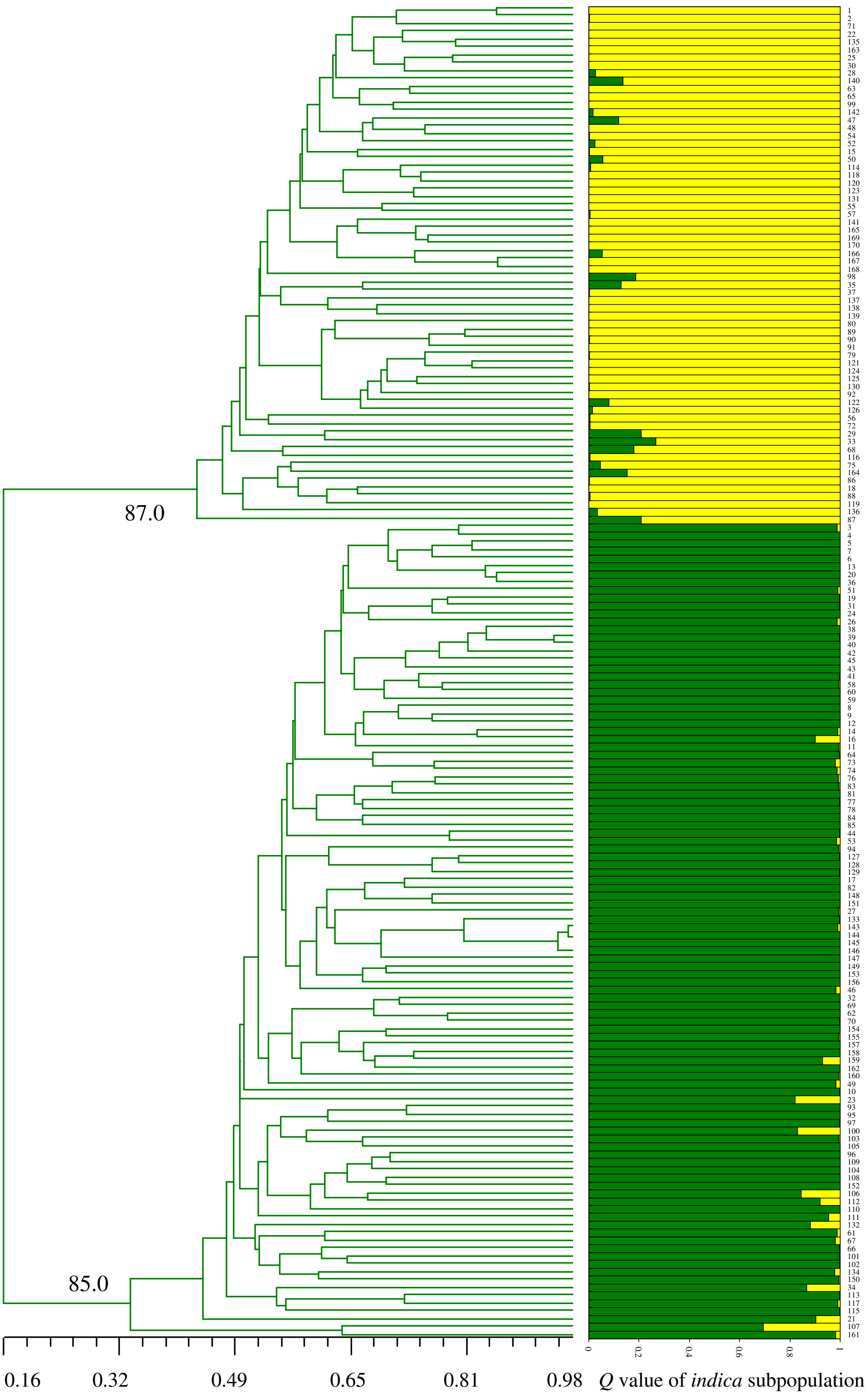
Along chromosome 7, the average LD (r^2) was estimated among markers in 2 Mb windows. The LD level in whole population had an uneven distribution with peaks on the chromosomal regions of 18–21, 21–24, and larger than 28 Mb (Fig. 3). Variance of LD was observed in both subpopulations along the same chromosome, but with much smaller ranges.

Association between agronomical traits and molecular markers on chromosome 7

For HD06, marker loci RM5672, RM3767 and RM1306 were significant, and among them RM3767 had the largest effect ($R^2 = 22.95\%$). Three markers, RM180, RM3767 and RM1306, were associated with HD07 having the largest effect from RM1306 ($R^2 = 20.12\%$). Two markers (RM3767 and RM1306) were commonly detected in two experiments while the third ones (RM5672 for HD06, RM180 for HD07) were adjacent marker loci. For PH06 and PH07, RM560 and RM1306 were significant markers in common. RM180 showed genetic association with PH06 only. RM1306 had the largest effects on both PH06 and PH07 explaining about 20% of phenotypic variation. The marker loci significantly associated with PL06 were RM1306 and RM22166. They individually explained 12.6 and 20.28% of the total phenotypic variations. For PL07, we identified only one significantly associated marker RM1306 ($R^2 = 19.39\%$) (Table 3; Fig. 4).

The effective polymorphic allele number and *Fst* values of 84 marker loci on chromosome 7 were integrated into Fig. 4. The effective polymorphic allele number ranged from 0 to 6. Two marker loci (RM2420 and RM2789) had no effective polymorphic allele after removing the rare alleles and no association results as well. Based on the Kendall's τ values (Supplementary materials, Table S3), the correlations between allele numbers and the association results (log *P*) were significant in HD06 ($r = -0.205*$),

Fig. 1 UPGMA clustering and population structure of 170 rice accessions based on 52 SSR or InDel loci on 12 chromosomes. *Left* The accessions were clustered into two subgroups by UPGMA. *Horizontal axis* represents the coefficient of similarity. *Labeled numbers* represent the bootstrap values of *indica* and *japonica* subgroups. *Right* Estimated population structure; each individual is represented by a *horizontal bar* broken into two colored segments, with lengths in proportion to *Q* values of the *indica* subpopulation (in *green color/dark gray*, serving as the scale of the *horizontal axis*) and *japonica* subpopulation (in *yellow color/light gray*), respectively



HD07 ($r = -0.259^{**}$) and PL07 ($r = -0.197^{*}$), but not significant in other three cases. The level of genetic differentiation of marker loci (i.e. *Fst*) had no significant correlation with the log *P* values of marker loci in all traits.

Identification of the allelic effects for highly significant associated markers

The rice accessions were grouped according to the alleles of associated markers with lowest log *P* values. Means of

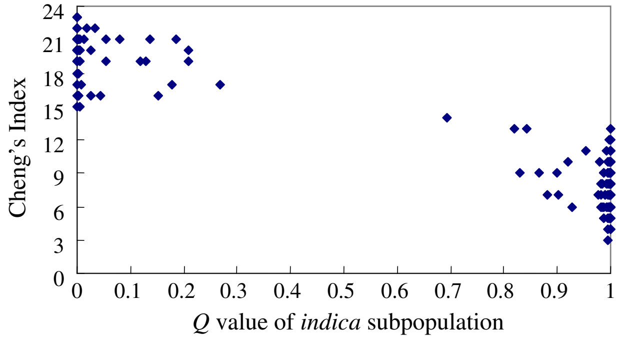


Fig. 2 The Cheng's indices correspond to the Q values estimated in the population structure analysis. The rice accessions fall apart into two subgroups including *japonica* varieties at the *top left* and *indica* varieties at the *bottom right*

groups and results of multiple comparisons based on $LSD_{0.05}$ and $LSD_{0.01}$ were listed in Table 4.

Marker locus RM1306, with five alleles, was significantly associated with all three traits. Accessions with allele 1 (A1) had the shortest growth period, PH and PL. Allele 2 (A2) had the second earliest HD, but had the highest mean values in PH and PL. Accessions with other three alleles (A3, A4 and A5) had later HDs than A1 and A2, but means of PH and PL between A1 and A2. For RM3767, group means of HD06 and HD07 had an order of A2 > A3 > A1 ≈ A4. Among three alleles of RM560, group means of PH were ordered as A2 > A3 > A1, while only the mean of A1 group was highly significant from two other allelic groups. RM22166 was detected as associated marker of PL in 2006 only. A4 had the lowest PL that was significantly different from top two alleles (A3 and A6).

Taking HD07 as an example, we compared the effects of several haplotypes, i.e. allele combinations, at two associated loci RM3767 (M1) and RM1306 (M2). A total of seven haplotypes were identified with the frequency of more than 5%, which were listed in descending order as M1A2:M2A2, M1A2:M2A4, M1A4:M2A4, M1A3:M2A2, M1A3:M2A4, M1A1:M2A2 and M1A1:M2A1 (Table 5). The samples with haplotype M1A1:M2A1 ($n = 11$) had the earliest HD that was highly significant from all other haplotype groups. Samples with haplotype M1A1:M2A2 ($n = 44$) had the second earliest HD and was significantly different from the haplotype of M1A2:M2A2 that had latest HD.

Table 3 Association of molecular markers on rice chromosome 7 with heading date (HD), plant height (PH) and panicle length (PL) observed in the 2006 and 2007 experiments

Traits	Experiments	Loci	Positions (Mb)	log *P*	R^2
HD06	2006	RM5672	6.379	−4.98	0.1851
		RM3767	9.037	−7.89	0.2295
		RM1306	28.946	−4.34	0.1747
HD07	2007	RM180	5.734	−4.11	0.1191
		RM3767	9.037	−5.57	0.1476
		RM1306	28.946	−6.07	0.2012
PH06	2006	RM180	5.734	−4.59	0.0811
		RM560	19.582	−6.33	0.0948
		RM1306	28.946	−8.05	0.1866
PH07	2007	RM560	19.582	−5.10	0.1610
		RM1306	28.946	−7.77	0.2237
PL06	2006	RM1306	28.946	−4.84	0.1260
		RM22166	29.416	−4.81	0.2028
PL07	2007	RM1306	28.946	−7.02	0.1939

Cursory categorization of accessions bearing different marker alleles

The rice accessions can be roughly categorized according to their origin places, belonging to *indica* or *japonica* subspecies, landraces or modern cultivars and so on. The accessions bearing different marker alleles had uneven distribution among such kind of categories.

For marker RM3767, all the accessions with A1 belong to the *indica* subpopulation. The 80% accessions from north China, Japan and South Korea had A4 with *japonica* accessions as the majority (89.7%). A large proportion of accessions with A2 also belonged to *japonica* subpopulation (69%), but 85% of them were from south and east China.

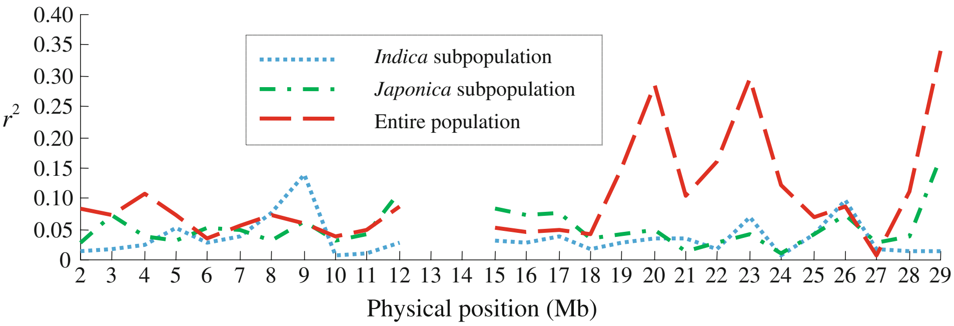


Fig. 3 Plot of linkage disequilibrium extent (r^2) against the physical positions along chromosome 7. *The lines* indicate the average r^2 values for 2 Mb windows in the entire population and two subpopulations

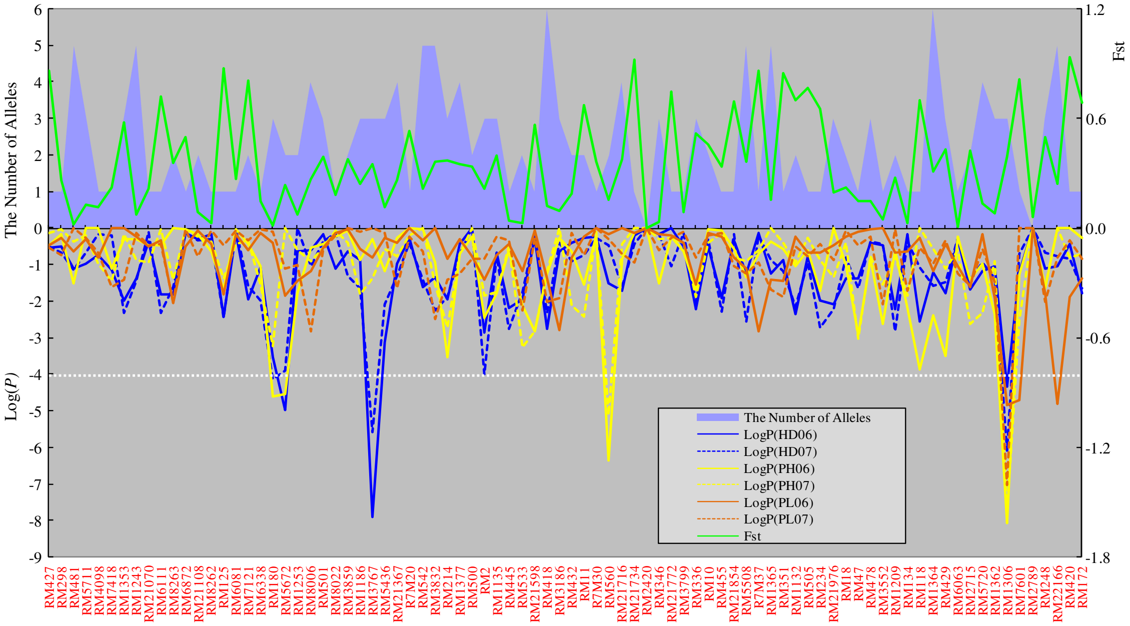


Fig. 4 Scanning the associations (in log *P*) of heading date (HD), plant height (PH) and panicle length (PL) with 84 marker loci on chromosome 7 in rice. Eighty-four marker loci are ordered on the *horizontal axis* according to their physical positions on the chromosome (see the list in Supplementary materials, Table S2). *The vertical scale* above *x*-axis represents the effective polymorphic allele number and *Fst* values of each marker loci. Effective polymorphic allele number = (allele number – rare allele number – 1)

For marker RM560, allele A1 had a predominant frequency in the whole set of accessions, and 68% samples among them were *indica* accessions. In the accessions with A2, 70% samples belong to the *indica* subpopulation and within which most accessions are landraces from Sichuan, Yunnan and Guizhou, as neighbor provinces in south-west China. All accessions with A3 belong to *japonica* subpopulation.

All 13 accessions with A1 of RM1306 came from south China and 10 of them were cultivars from Hunan province. For all 81 accessions with A2, 94% of them belonged to the *indica* subpopulation. All the 29 accessions with A4 were classified into the *japonica* subpopulation.

Discussion

Chinese rice germplasm had a predominant population structure containing two subpopulations in correspondence with *indica* and *japonica* subspecies

The differentiation of two subspecies, *indica* and *japonica*, in Asian cultivated rice (*Oryza sativa* L.) was widely accepted for a long history (Kato et al. 1928). More complicated intra-species classification was later suggested by several authors. Briefly, the third subspecies, *javanica*, or interpreted as one or more eco-groups (*aus*, *javanica*, or *tropical japonica*, etc.) was added to embody a small proportion of germplasm from south Asia. Those germplasm were less differentiated, and had large genetic distances from both *indica* and *japonica* subspecies (reviewed by Wang et al. 1998).

Some previous studies on genetic structure of rice (*O. sativa*) detected more than two groups (Garris et al. 2005; Agrama et al. 2007). Garris et al. (2005) gave out a population structure with five groups, *aromatic*, *aus*, *indica*, *temperate japonica* and *tropical japonica*, in a set of 234 rice accessions collected from all over the world. In our study, the model-based estimation from 52 SSR or InDel markers suggested a structure with two subpopulations in 170 rice accessions. Most accessions (97.6%) were assigned into the corresponding subgroups with the membership (*Q*) larger than 0.80. The two subgroup divisions were obtained with coincidence based on either *Q* values or UPGMA clustering, and were corresponding with dividing of subspecies according to empirical category and the observation of referential morphological traits (i.e. Cheng's index). Only four accessions were located near the intermediate part between *indica* and *japonica* subpopulations in UPGMA clustering. They are *Kasalath* (161) from India, *Qingke* (107) and *Sanbangqishiluo* (87) from Yunnan province and *Qingsiai16B* (21) from Guangdong province, China. This phenomenon has been observed and discussed in our previous report (Mei et al. 2007).

Table 4 Multiple comparison among means of accessions grouped by the alleles of associated markers of heading date (HD), plant height (PH) and panicle length (PL)

Traits	Markers	Alleles	*N*	Means[a]	$P \leq 0.05$	$P \leq 0.01$
HD06	RM3767	A2	31	93.81	a	A
		A3	34	83.56	b	AB
		A1	61	75.03	c	B
		A4	27	74.52	c	B
HD07	RM3767	A2	32	91.53	a	A
		A3	37	84.68	b	B
		A4	29	77.93	bc	BC
		A1	65	75.83	c	C
HD06	RM1306	A3	8	89.00	a	A
		A5	10	84.70	a	A
		A4	26	84.54	a	A
		A2	74	81.92	a	A
		A1	13	61.54	b	B
HD07	RM1306	A5	10	88.50	a	A
		A3	8	88.25	a	A
		A4	29	86.76	a	A
		A2	80	82.20	a	A
		A1	13	61.85	b	B
PH06	RM560	A2	20	129.25	a	A
		A3	24	122.34	a	A
		A1	100	100.15	b	B
PH07	RM560	A2	15	135.40	a	A
		A3	20	127.65	a	A
		A1	98	105.27	b	B
PH06	RM1306	A2	70	118.14	a	A
		A5	9	112.00	a	A
		A3	7	110.76	a	A
		A4	28	106.18	a	A
		A1	13	71.15	b	B
PH07	RM1306	A2	63	122.01	a	A
		A5	7	113.71	a	A
		A4	27	112.54	a	A
		A3	5	110.60	ab	AB
		A1	13	77.99	b	B
PL06	RM22166	A3	15	25.07	a	A
		A6	14	24.06	a	A
		A2	19	23.23	a	AB
		A1	14	22.87	ab	AB
		A5	22	22.42	ab	AB
		A4	10	19.18	b	B
PL06	RM1306	A2	76	24.13	a	A
		A5	10	21.87	ab	AB
		A4	28	21.52	b	B
		A3	7	21.00	b	B
		A1	13	20.43	b	B
PL07	RM1306	A2	77	23.83	a	A
		A5	10	22.06	ab	AB
		A3	6	21.19	ab	AB
		A4	28	20.52	b	B
		A1	13	20.19	b	B

[a] Means followed by different letters were significantly different by the LSD test at the level $P \leq 0.05$ (in lowercases) and $P \leq 0.01$ (in uppercases)

Table 5 Multiple comparison among means of heading date in 2007 (HD07) of accessions grouped by the combinations of alleles on RM3767 (M1) and RM 1306 (M2) loci

Haplotypes	*N*	Means[a]	$P \leq 0.05$	$P \leq 0.01$
M1A2:M2A2	10	94.00	a	A
M1A2:M2A4	9	88.33	ab	AB
M1A4:M2A4	12	86.58	ab	AB
M1A3:M2A2	20	86.05	ab	AB
M1A3:M2A4	8	85.25	ab	AB
M1A1:M2A2	45	79.18	b	B
M1A1:M2A1	11	61.64	c	C

[a] Means followed by different letters were significantly different by the LSD test at the level $P \leq 0.05$ (in lowercases) and $P \leq 0.01$ (in uppercases)

LD variation in parallel with the level of genetic differentiation (*Fst*) between *indica* and *japonica* subspecies

Variation of LD level among genomic regions was observed in this study (Fig. 3) and in many other researches (Remington et al. 2001; Caldwell et al. 2006; Hyten et al. 2007; Mather et al. 2007; Somers et al. 2007; Wang et al. 2008). Mather et al. (2007) observed the correlation between estimated recombination rate and the extent of LD across the genome. Somers et al. (2007) suggested that the changes in LD along chromosomes are indicative of areas of genome that are under selection pressure.

In this study, the LD extent in 2 Mb windows varied along rice chromosome 7 in the entire population, but hardly significant in two subpopulations. According to the allelic frequency in *indica* and *japonica* accessions, *Fst* values of marker loci were calculated to show the inter-subspecies differentiation. The chromosomal regions with high LD extent in entire population were strictly parallel to the positions with high level of *Fst* values (Figs. 3, 4). In our previous study, we developed two reciprocal IL populations between *indica* and *japonica* lines, the proportion of introgression segments from *indica* donor to *japonica* background was much higher than that in the reciprocal direction on the long arm of rice chromosome 7 (unpublished data). This phenomenon was not properly explained until now, but showed an unknown linkage to the result in this study. Based on above observations, the change of LD extent against marker loci along rice chromosome 7 might reflect to the varied level of *indica*/*japonica* differentiation of different genomic regions.

Before getting more convincing conclusion, further studies should be done to eliminate two weaknesses in this experiment. First, the changes of LD extents along other chromosomes should be detected to show the panorama of whole genome. Secondly, validation based on more and different types of molecular markers seems necessary to avoid the probably biased estimation based on SSR and InDel markers.

Associated marker loci and multiple allelic effects of each locus in germplasm provided sufficient resources for rice breeding

Although only the mapping results of three agronomic traits on single chromosome were shown in this paper, the MLM association analysis based on $(Q + K)$ model seemed to be efficient and reliable. First, two to three associated marker loci were detected for each trait in single chromosome that was repeatable in most cases for the data collected from two independent experiments in different locations and years (Table 3). Secondly, the mapping results in this study were coincident with many other reports. For example, the significant marker loci associated with HD in this study were located near several HD quantitative trait locus (QTL) on chromosome 7 as reported previously (Lin et al. 1998; Li et al. 2003; Xue et al. 2008). Marker loci RM180, significant for PH, probably linked to the chromosomal regions where *Ghd7* has been detected. The QTL *Ghd7* has major effects on an array of traits in rice, including number of grains per panicle, PH and HD (Xue et al. 2008). Two other significant associated markers to PH in both experiments (RM560 and RM1306) located within the regions of *ph*-7 (Lu et al. 1997) and *ph*7 for PH (Yan et al. 1998). Marker loci RM1306 and RM22166 were associated to PL and located nearby at the distal end of chromosome 7 where QTLs for PL had been detected in several populations (Xu et al. 2001; Jiang et al. 2004). The information not only provided partial evidence of the consistency between association analysis and linkage mapping, but also encouraged the authors to take further study on the allelic diversity of cloned genes (like *Ghd7*) or candidate genes of major QTLs using the diverse panel of rice germplasm.

The estimation of allelic effects showed diverse patterns for associated markers in our present study (Table 4). For example, SSR marker RM1306 was detected in all six cases of three traits by two field experiments, showing strong pleiotropic effects on HD, PH and PL. This result is in agreement with the fact that three measured traits are highly correlated with each other. Among five alleles, A1 had negative effects on all three traits and was found in 13 *indica* accessions from south China (including 10 accessions from Hunan province). A2 was carried by 81 accessions (94% are *indica*) and contributed to the second shortest HD but highest group means of PH and PL. A3–A5 had similar patterns of effects (i.e. high HD, and medium PH and PL). A4 and A5 appeared in *japonica* accessions only (but one exception). This kind of information is useful in the breeding program based on germplasm with distant genetic relations, by which the genetic diversity of modern varieties can be broadened.

Before using associated markers in marker-assisted breeding, careful interpretation and case-by-case validation seemed to be necessary to confirm the consistency among accessions and to estimate unbiased expectations of genetic gain (Breseghello and Sorrells 2006). Association analysis based on higher density of molecular markers, larger size of populations and repeated experiments under multiple environments will significantly reduce the bias between marker allele estimates and the real effects of target gene alleles.

Acknowledgments The authors are grateful to Dr. Zichao Li for providing us the rice accessions in the putative mini-core collection of Chinese rice germplasm. Two anonymous reviewers gave us many critical comments, based on that the manuscript was improved in several issues from its early version. This study was jointly supported by a grant from the Ministry of Agriculture of China (948-2006-R1), a grant from National Key Basic Research Program of China (973 Plan), a grant from the National Natural Science Foundation of China (30830071) and a grant from Shanghai Municipal Commission of Science and Technology.

References

Agrama HA, Eizenga GC, Yan W (2007) Association mapping of yield and its components in rice cultivars. Mol Breed 19:341–356

Aranzana MJ, Kim S, Zhao K, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang C, Toomajian C, Traw B, Zheng H, Bergelson J, Dean C, Marjoram P, Nordborg M (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet 1:531–539

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:165–1177

Caldwell KS, Russell J, Langridge P, Powell W (2006) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. Genetics 172:557–567

Chen X, Temnykh S, Xu Y, Cho XG, McCouch SR (1997) Development of a microsatellite framework map providing genome-wide coverage in rice (*Oryza sativa* L.). Theor Appl Genet 95:553–567

Cheng KS (1985) A statistical evaluation of the classification of rice cultivars into *hsien* and *keng* subspecies. Rice Newsl 2:46–48

Colosi JC, Schaal BA (1993) Tissue grinding with ball bearings and vortex mixer for DNA extraction. Nucleic Acids Res 21:1051–1052

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Ann Rev Plant Biol 54:357–374

Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44:1054–1064

Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. Genetics 169:1631–1638

Hao CY, Dong YC, Wang LF, Zhang HN, Ge HM, Jia JZ, Zhang XY (2008) Genetic diversity and construction of core collection in Chinese wheat genetic resources. Chin Sci Bull 53:1518–1526

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyze spatial genetic structure at the individual or population levels. Mol Ecol Notes 2:618–620

Hasan M, Friedt W, Pons-Kühnemann J, Freitag NM, Link K, Snowdon RJ (2008) Association of gene-linked SSR markers to seed glucosinolate content in oilseed rape (*Brassica napus* ssp. *napus*). Theor Appl Genet 116:1035–1049

Holland JB (2007) Genetic architecture of complex traits in plants. Curr Opin Plant Biol 10:156–161

Hyten DL, Choi IY, Song Q, Shoemaker RC, Nelson RL, Costa JM, Specht JE, Cregan PB (2007) Highly variable patterns of linkage disequilibrium in multiple soybean populations. Genetics 175:1937–1944

Insightful Corporation (2001) S-Plus 6 for windows, guide to statistics, vol 1. Insightful Corporation, Seattle

Iwata H, Uga Y, Yoshioka Y, Ebana K, Hayashi T (2007) Bayesian association mapping of multiple quantitative trait loci and its application to the analysis of genetic variation among *Oryza sativa* L. germplasms. Theor Appl Genet 114:1437–1449

Jiang GH, Xu CG, Li XH, He YQ (2004) Characterization of the genetic basis for yield and its component traits of rice revealed by doubled haploid population. Acta Genet Sin 31:63–72

Kato S, Kosaka H, Hara S (1928) On the affinity of rice varieties as shown by fertility of hybrid plants. Bull Sci Fac Agric Kyushu Univ 3:132

Kraakman AT, Niks RE, Van den Berg PM, Stam P, Van Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168:435–446

Li ZC, Zhang HL, Zeng YW, Yang ZY, Shen SQ, Sun CQ, Wang XK (2002) Studies on sampling schemes for the establishment of core collection of rice landraces in Yunnan, China. Genet Resour Crop Evol 49:67–74

Li ZK, Yu SB, Lafitte HR, Huang N, Courthouse B, Hittalmani S, Vijayakumar CH, Liu GF, Wang GC, Shashidhar HE, Zhuang JY, Zheng KL, Singh VP, Sidhu JS, Srivantaneeyakul S, Khush GS (2003) QTL environment interactions in rice. I. Heading date and plant height. Theor Appl Genet 108:141–153

Li YH, Guan RX, Liu ZX, Ma YS, Wang LX, Li LH, Lin FY, Luan WJ, Chen PY, Yan Z, Guan Y, Zhu L, Ning XC, Smulders MJ, Li W, Piao RH, Cui YH, Yu ZM, Guan M, Chang RZ, Hou AF, Shi AN, Zhang B, Zhu SL, Qiu LJ (2008) Genetic structure and diversity of cultivated soybean (*Glycine max* (L.) Merr.) landraces in China. Theor Appl Genet 117:857–871

Lin SY, Sasaki T, Yano M (1998) Mapping quantitative trait loci controlling seed dormancy and heading date in rice. *Oryza sativa* L., using backcross inbred lines. Theor Appl Genet 96:997–1003

Liu K, Muse M (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129

Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). Am J Bot 82:1420–1425

Lu CF, Shen LH, Tan ZB, Xu YB, He P, Chen Y, Zhu LH (1997) Comparative mapping of QTLs for agronomic traits of rice across environments by using a doubled-haploid population. Theor Appl Genet 94:145–154

Mather KA, Caicedo AL, Polato NR, Olsen KM, McCouch S, Purugganan MD (2007) The extent of linkage disequilibrium in rice (*Oryza sativa* L.). Genetics 177:2223–2232

Mazzucato A, Papa R, Bitocchi E, Mosconi P, Nanni L, Negri V, Picarella ME, Siligato F, Soressi GP, Tiranti B, Veronesi F (2008) Genetic diversity, structure and marker-trait associations in a collection of Italian tomato (*Solanum lycopersicum* L.) landraces. Theor Appl Genet 116:657–669

Mei HW, Feng FJ, Lu BR, Wen WW, Paterson AH, Cai XX, Chen L, Feltus FA, Xu XY, Wu JH, Yu XQ, Chen HW, Li Y, Luo LJ (2007) Experimental validation of inter-subspecific genetic diversity in rice represented by the differences between the DNA sequences of ‘Nipponbare’ and ‘93-11’. Chin Sci Bull 52:1327–1337

Pritchard JK, Stephen M, Donnelly P (2000a) Inference on population structure using multilocus genotype data. Genetics 155:945–959

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. Am J Hum Genet 67:170–181

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci 20:11479–11484

Sneath PHA, Sokal RR (1973) Numerical taxonomy. Freeman, San Francisco

Somers DJ, Banks T, DePauw R, Fox S, Clarke J, Pozniak C, McCartney C (2007) Genome-wide linkage disequilibrium analysis in bread wheat and durum wheat. Genome 50:557–567

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) *Dwarf8* polymorphisms associate with variation in flowering time. Nat Genet 28:86–289

Wang XK, Li RH, Sun CQ, Li ZC, Cai HW, Sun XL (1998) Identification and classification of subspecies of Asian cultivated rice and their hybrids. Chin Sci Bull 43:1864–1872

Wang RH, Yu YT, Zhao JR, Shi YS, Song YC, Wang TY, Li Y (2008) Population structure and linkage disequilibrium of a mini core set of maize inbred lines in China. Theor Appl Genet 117:1141–1153

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. Evolution 38:1358–1370

Xu JL, Xue QZ, Luo LJ, Li ZK (2001) QTL dissection of panicle number per plant and spikelet number per panicle in rice (*Oryza sativa* L.). Acta Genet Sinica 28:752–759

Xu Y, Beachell H, McCouch SR (2004) A marker-based approach to broadening the genetic base of rice in the USA. Crop Sci 44:1947–1959

Xue WY, Xing YZ, Weng XY, Zhao Y, Tang WJ, Wang L, Zhou HJ, Yu SB, Xu CG, Li XH, Zhang QF (2008) Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. Nat Genet 40:761–767

Yan JQ, Zhu J, He C, Benmoussa M, Wu P (1998) Molecular dissection of developmental behavior of plant height in rice (*Oryza sativa* L.). Genetics 150:1257–1265

Yu JM, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Zhang N, Xu Y, Akash M, McCouch S, Oard JH (2005) Identification of candidate markers associated with agronomic traits in rice using discriminant analysis. Theor Appl Genet 110:721–729

Zhu CS, Gore M, Buckler ES, Yu JM (2008) Status and prospects of association mapping in plants. Plant Genome 1:5–20