

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/299414537>

Genome-wide association study of grain yield and related traits using a collection of advanced indica rice breeding lines...

Article in *Field Crops Research* · March 2016

DOI: 10.1016/j.fcr.2016.03.008

CITATIONS

0

READS

158

7 authors, including:



Shanshan Liang

International Rice Research Institute

12 PUBLICATIONS 198 CITATIONS

[SEE PROFILE](#)



Xiangqian Zhao

Zhejiang Academy of Agricultural Sciences

23 PUBLICATIONS 364 CITATIONS

[SEE PROFILE](#)



Meixue Zhou

University of Tasmania

198 PUBLICATIONS 3,475 CITATIONS

[SEE PROFILE](#)



Guoyou Ye

International Rice Research Institute

91 PUBLICATIONS 1,237 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Allelic mining [View project](#)



Preharvest sprouting in wheat [View project](#)



Contents lists available at ScienceDirect

Field Crops Research

journal homepage: www.elsevier.com/locate/fcr



Genome-wide association study of grain yield and related traits using a collection of advanced *indica* rice breeding lines for irrigated ecosystems

Shanshan Liang^{a,b}, Liwen Wu^{c,d}, Guangjun Ren^e, Xiangqian Zhao^b, Meixue Zhou^a, David McNeil^a, Guoyou Ye^{b,*}

^a University of Tasmania, Private Bag 98, Hobart, Tasmania 7001, Australia

^b Genetics and Biotechnology Division, International Rice Research Institute (IRRI), Los Baños, Laguna, Philippines

^c State Key Laboratory of Rice Biology/Chinese National Center for Rice Improvement, China National Rice Research Institute, Hangzhou 310006, China

^d Agricultural Genomics Institute in Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China

^e Institute of Crop Science, Sichuan Academy of Agricultural Sciences, Chengdu 610066, Sichuan, China

ARTICLE INFO

Article history:

Received 3 February 2016

Received in revised form 17 March 2016

Accepted 19 March 2016

Available online xxx

Keywords:

Rice

Yield-related traits

QTL

Genome-wide association studies (GWAS)

Genotyping-by-sequencing

ABSTRACT

Genome-wide association study (GWAS) is an effective approach for the identification of marker-trait associations (MTA) using more diverse germplasm not constrained by specific crossing. This study used a collection of 327 advanced *indica* breeding lines and varieties genotyped using a genotyping-by-sequencing method and phenotyped in eight environments to identify MTA for grain yield (GY) and 10 related traits. The current panel of *indica* breeding lines were divided into two subpopulations by three analytical methods including STRUCTURE, principal component analysis (PCA) and neighbour joining (NJ) tree. STRUCTURE and PCA gave the same classification of genotypes while the NJ tree results are slight difference in assigning genotypes to subpopulation from STRCUTURE and PCA. Wide variation was observed for all the 11 traits in the whole panel and the two subpopulations inferred by STRUCTURE analysis. Linkage disequilibrium (LD) analysis showed that LD decay varied across the chromosomes and average pairwise squared correlation coefficient (r^2) dropped to half of its maximum value ($r^2 = 0.25$) within a physical distance of 200 kb in the whole population. A MLM model controlling both population structure and cryptic relatedness was chosen to identify MTA for all the 11 tested traits. A total of 452 MTA that were delineated into 43 QTL were identified for all traits but PB, SB, and SR with 39 QTL being not reported before. Three QTL on chromosome 6, 9 and 12 were identified for GY but only in DS2. The numbers of QTL identified for the remaining traits varied from two to 26. Most of the detected QTL were found in only one environment. Four QTL were located in the regions containing genes/QTL previously identified for other related traits. The effects of identified QTL were relatively small with the highest percentage of phenotypic variance explained by a single QTL being 9.6%. The identified QTL are directly relevant to and can be more effectively used in breeding programs.

© 2016 Published by Elsevier B.V.

Abbreviations: AM, association mapping; DTF, days to flowering; GN, grain number per panicle; GBS, genotyping-by-sequencing; GEI, genotype-by-environment interaction; GLM, general liner model; GS, genomic selection; GWAS, genome-wide association studies; GY, grain yield; LD, linkage disequilibrium; MAF, minor allele frequency; MAS, marker-assisted selection; MET, multi-environment trials; MLM, mixed linear model; MSD, mean squared difference; MTA, marker-trait associations; NJ, neighbour joining; PB, primary branches per panicle; pH, plant height; PN, panicle number per plant; QEI, QTL-by-environment interaction; SN, spikelet number per panicle; SR, seed setting rate; SB, secondary branches per panicle; TGW, thousand grain weight; TN, tiller number per plant.

* Corresponding author.

E-mail address: g.ye@irri.org (G. Ye).

<http://dx.doi.org/10.1016/j.fcr.2016.03.008>

0378-4290/© 2016 Published by Elsevier B.V.

1. Introduction

Rice (*Oryza sativa* L.) is extensively cultivated on every continent in more than 100 countries and consumed by more than half of the world's population (Juliano, 1993). As world population grows, the global rice demand is estimated to rise from 439 million tons (milled rice) in 2010–496 million tons in 2020 and further increase to 555 million tons in 2035 (GRISP, 2010). Thus a significant increase in rice production is needed for the future of food security.

The Green Revolution in the 1960s greatly increased rice production with at least 50% of the increase being due to adoption of new cultivars. However, increased production potential of modern

rice cultivars has stagnated (Nguyen and Ferrero, 2006) partially due to its narrow genetic base resulting from narrow breeding populations and the population bottleneck that occurred during domestication. How to break the yield barrier is still the major challenge for rice breeders (GRiSP, 2010; Ye et al., 2013). On the other hand, as a model crop species for plant molecular biology and genomics, rice has more accumulated molecular and genomic information than most other crops. The utilization of this genetic information offers the rice breeding community a range of modern tools and methods for addressing this challenge. An integrated strategy was proposed by the International Rice Research Institute (IRRI) to increase breeding efficiency by effectively and efficiently utilizing well proven conventional breeding methods, new techniques and methods enabled by modern molecular biology and genomics and advanced methods in experimental design and data analysis (GRiSP, 2010; Ye et al., 2013). This strategy utilizes (marker-assisted) recurrent selection to quickly pyramid the major genes/QTL in the first few selection cycles and maintain genetic variation contributed by many minor genes to be explored in later cycles, explores genomic selection (GS) for reducing the breeding cycles and the costs of phenotyping and adopts advanced experimental design and data analysis methods to improve heritability (Ye et al., 2013).

To implement this strategy, 392 advanced breeding lines and cultivars from many breeding programs in different countries representing the genetic diversity of breeding gene pools for irrigated ecosystems were collected to be used as part of the base breeding population. This population has been phenotyped via multi-environment trials (METs) in South-east Asia and genotyped using markers for well-characterized genes/QTL and genome-wide markers to obtain essential information for designing more efficient mating and selection schemes of this general breeding strategy. Phenotypic analysis and basic genotypic analysis were used to remove erroneous lines and lines unadapted to the targeted tropical and subtropical environments and to investigate genotype-by-environment interaction (GEI) for GY (Liang et al., 2015). Thirty-nine cloned or fine-mapped genes/QTL for GY and related traits were tested for their usefulness through association analysis (Liang et al., 2016). Significant gene-by-environment interaction was found for all of the genes/QTL and GY could not be well predicted using the markers significantly associated with the measured traits. More marker-trait associations (MTA) need to be identified for grain yield improvement.

Considering that the majority of the well characterized genes/QTL were identified using biparental populations derived from contrasting parents, it is not unexpected that their transferability to breeding populations is low. An alternative method for the detection of MTA is association mapping (AM). AM identifies MTA caused by linkage disequilibrium (LD), which is the non-random association of alleles at separate loci. AM utilizes ancestral recombination events to identify MTA and provides comparatively higher mapping resolution than the biparental linkage analysis (Zhu et al., 2008). AM has been successfully employed in rice to identify MTA for a range of traits, including yield and yield components (Agrama et al., 2007), cold tolerance (Cui et al., 2013), grain quality, flowering time (Ordóñez et al., 2010) and harvest index (Li et al., 2012). Many associated markers found in these studies were in regions where QTL have previously been identified. The development of several next generation sequencing platforms makes GWAS more and more popular. Using shallow sequencing, a total of 80 MTA for 14 agronomic traits were identified, explaining an average of about 36% of the phenotypic variance (Huang et al., 2010). Among these loci, six were close to known genes *OsC1* (chromosome 6), *ALK* (chromosome 6), *Rc* (chromosome 6), *qSW5* (chromosome 5) and *GS3* (chromosome 3), respectively. The same method was applied to a larger collection of 950 worldwide rice varieties and

identified a total of 32 new loci for flowering time and 10 grain-related traits (Huang et al., 2012). Zhao et al. (2011) performed GWAS using a global collection of 413 rice accessions genotyped using a high-quality custom-designed 44,100 SNP array and phenotyped for 34 traits including morphological, developmental and agronomic traits over two consecutive crop seasons. Dozens of common variants influencing numerous complex traits were identified.

AM is usually conducted using panels of diverse germplasm. While maximizing the genetic diversity, which is beneficial for identifying novel QTL and candidate genes that underlie traits of interest, these panels may not be adapted to identifying relevant variations directly useful in breeding programs (Bordes et al., 2014). Any identified QTL should be validated in a breeding population before they can be used in selection. AM in locally-adapted breeding material should be of more practical benefit to breeders (Bernardo, 2008), since identified beneficial alleles could be incorporated into the creation of new cultivars with limited deleterious effect (Pauli et al., 2014). Empirical studies in wheat and barley demonstrated that GWAS using elite breeding populations is an effective strategy for integrating new genomic technologies into the development of superior cultivars (Bordes et al., 2014; Mohammadi et al., 2015; Pauli et al., 2014; Pozniak et al., 2012). Recently, GWAS performed using a population of elite *indica* rice breeding lines from IRRI's irrigated breeding program genotyped with GBS identified 52 QTL for 11 agronomic traits, with some QTL being co-localized with QTL identified before (Begum et al., 2015). However, the genome-wide reduction of genetic diversity caused by intensive selection during breeding could reduce the efficiency of AM for some traits. A population mixing lines from several breeding programs, each led by independent breeders with their own germplasm, could increase the level of diversity for an efficient use of AM (Bordes et al., 2014).

The main objective of this study was to identify MTA/QTL for GY and 10 related traits using the base breeding population assembled for future *indica* rice breeding for irrigated ecosystem at IRRI. More than 321k markers were generated by the newly developed genotyping-by-sequencing (GBS) method and 76k high quality markers with good genome coverage were used. Population structure of the panel was studied using a model-based Bayesian cluster analysis implemented in STRUCTURE (Falush et al., 2003; Pritchard et al., 2000a), principal component analysis (PCA) and nearest neighbour joining (NJ) methods (Saitou and Nei, 1987). LD patterns of the whole populations and the inferred subpopulations were investigated. To reduce false positives caused by population structure and unequal relatedness among genotypes (lines) four statistical models were compared using the QQ plot and mean squared difference (MSD) between expected and observed *p* values from analysis proposed by Stich et al. (2008). The PK model was chosen to detect MTA for all the trait environment combinations. A total of 452 MTA that were delineated into 43 QTL were identified for GY and seven related traits.

2. Materials and methods

2.1. Plant materials and phenotyping

Three hundred and ninety two rice varieties or advanced lines developed for the irrigated lowland ecosystem were used in this study to achieve a large amount of genetic diversity. The majority of the lines were from IRRI (225). A good number of lines were also from PhilRice (38), CIAT (19), China (14) and Vietnam (10) (Supplementary Table S1). Phenotyping was conducted in Jiangxi (JX) and Sichuan (SC) in China and IRRI headquarters (Los Baños, Philippines). The experiments in JX and SC were conducted in one crop season in 2012. At IRRI, the experiments were carried

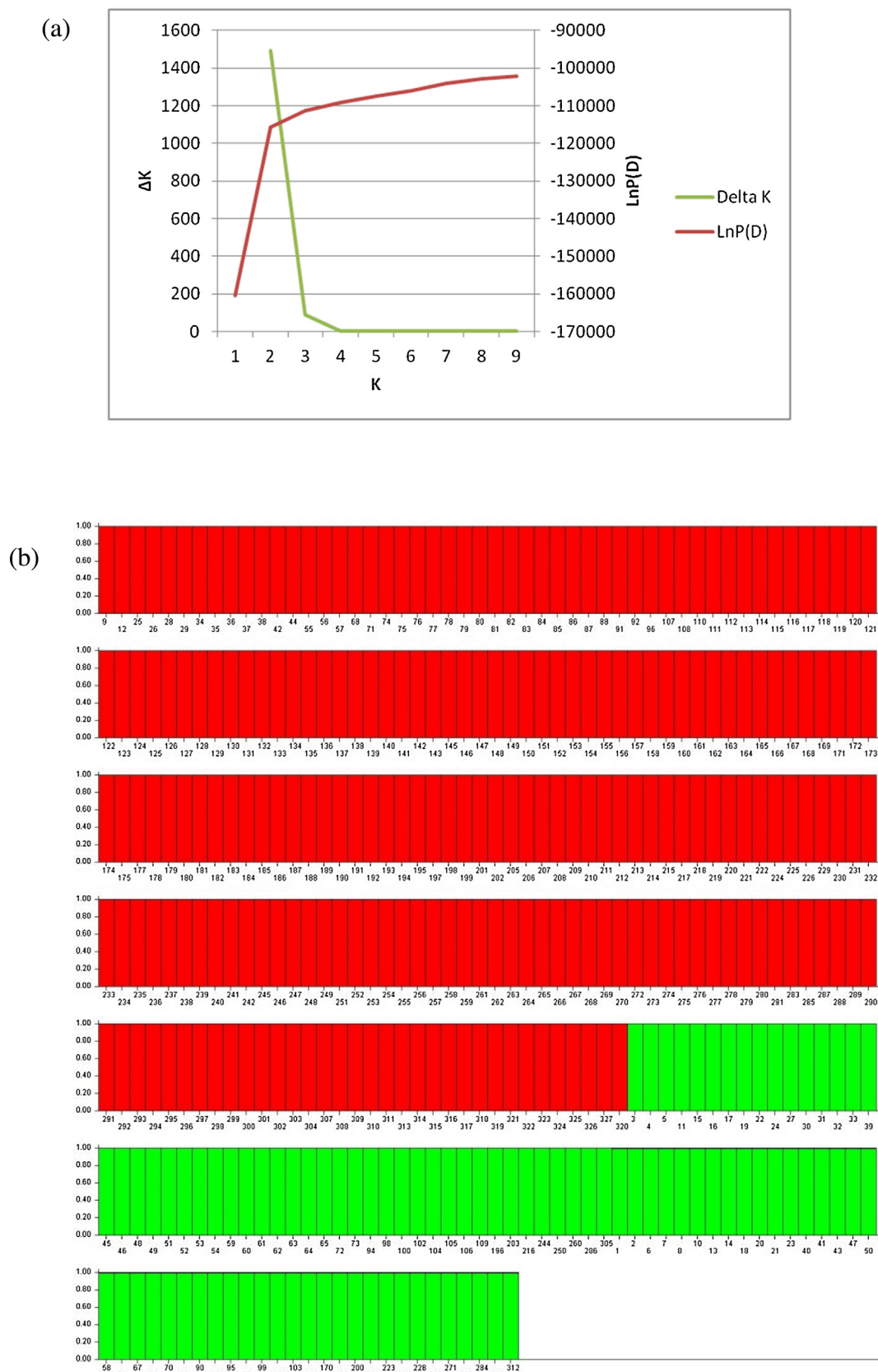


Fig. 1. Population structure analysis using STRUCTURE based on 1072 SNPs. (a) ΔK and LnP(D) as a function of the number of subpopulations (k), (b) population sub-structuring.

out in both dry season (DS) and wet season (WS) of 2012. Three nitrogen fertilizer application rates, no nitrogen, low (90 kg ha^{-1}) and high (180 kg ha^{-1}), were used to create three artificial environments in the DS, designated as DS1, DS2 and DS3, respectively. Similarly, three nitrogen fertilizer application rates, no nitrogen, low (45 kg ha^{-1}) and high (90 kg ha^{-1}), were used to create three artificial environments in the WS, designated as WS1, WS2 and WS3, respectively. GY and 10 related traits including GN, PN, TGW, SN, SR, PB, SB, TN, DTF and pH were measured. The traits measured included GY, grain number per panicle (GN), panicle number (PN), thousand grain weight (TGW), spikelets number per panicle (SN), seed setting rate (SR), number of primary branches per panicle (PB), number of secondary branches per panicle (SB), tiller number per plant (TN), days to flowering (DTF) and plant height (pH). All 11 traits were measured in the experiments conducted at IRRI except for DS2 with only DTF, pH, PN, GY and TGW being measured. GY, DTF, GN, pH, PN, TGW and SR were measured in SC, while GY, DTF, GN, pH and SR were measured in JX. Detailed trial description and trait measurement were given in Liang et al. (2015).

2.2. Genotyping and GBS data analysis

Young leaf tissue of a single plant was collected from each line. DNA extraction was performed using the commercial Qiagen DNeasy 96 Plant kit (cat. no. 69181) following the manufacturer's protocol. Library preparation, sequencing, SNP discovery and calling were performed using 384-plex GBS by the Cornell University using the protocol described by Elshire et al. (2011). Imputation of missing values in the unfiltered SNP data was performed using the software TASSEL 3.0 with default settings. The algorithm works by dividing the SNPs into small SNP windows, then identifying the most similar inbred lines within each window to fill the missing data. The algorithm takes advantage of small identical by decent (IBD) regions shared between pairs of inbred lines in the collection. If the window from the closest neighbour has more than 5% difference from the line being imputed, the data point is left as missing (Romay et al., 2013). To further improve data quality, other checks were carried out. SNPs with minor allele frequency (MAF) <5% or heterozygosity rate >5% were discarded after imputation.

2.3. Statistical analysis

The number of lines used in population structure detection, LD analysis and association analysis was 327 (Table S1), since 19 lines confirmed to be tropical japonica, lines with low genotyping quality and a few lines detected as outliers using STRUCTURE based on 50 evenly distributed SSR markers were removed.

2.3.1. Population structure

A subset of 1072 SNPs markers evenly distributed on the 12 chromosomes were chosen to detect population structure of the association panel using a model-based Bayesian clustering analysis method implemented in the software STRUCTURE 2.3.4 (Pritchard et al., 2000a). An admixture model with correlated allele frequencies was used. The number of groups (k) varied from 1 to 10, with burn in 5000 followed by 100,000 MCMC (Markov Chain Monte Carlo) iterations. Fifteen independent runs were conducted for each k. The optimal number of subpopulations was determined based on change of the $\text{LnP}(D)$ and the Delta k method proposed by Evanno et al. (2005). PCA was performed with all 76,452 SNPs using GAPIT R package (Lipka et al., 2012). The first three principal components were used to visualize the dispersion of the association panel in a graph. PCA was also analyzed in the subpopulations derived from the STRUCTURE analysis. The pairwise similarity coefficients were calculated for all pairwise combinations of the 327 lines according to the method developed by Nei and Li (1979) in the software

TASSEL v.5.2.11 (Bradbury et al., 2007). The NJ tree based on this genetic similarity matrix was constructed accordingly. To display the familial relationship, a heatmap of the kinship matrix was built up in R (R Core Team, 2015).

2.3.2. Linkage disequilibrium analysis

Genome-wide LD analysis was performed for the whole panel and the two subpopulations. LD was measured using squared allele frequency correlations (r^2) between marker pairs, which was calculated using software TASSEL5.2.11 by setting the sliding window size at 50. The loci was considered to be in significant LD when $p < 0.001$. The pattern and distribution of intra-chromosomal LD within the whole population and each of the two subpopulation was visualized and studied from LD plots generated for the whole genome and each of the chromosomes using R (R Core Team, 2015). To investigate the average LD decay in the whole genome significant intra-chromosomal r^2 values were plotted against the distance (bp) between markers. The LD decay was measured as the physical distance (kb) at which r^2 dropped to half of its maximum value (Huang et al., 2012).

2.3.3. Association analysis

Genotypic best linear unbiased estimations (BLUEs) obtained from the single-site phenotypic analysis (Liang et al., 2015) was used as phenotypes for association analysis. Kinship matrix (K) which measured the genetic similarity between individuals as the proportion of times a given pair of individuals had the same genotype across all SNPs (IBS values) was calculated using TASSEL v.5.2.11 (Bradbury et al., 2007). The following models implemented in TASSEL were tested to choose the best model: (i) Q model, a general linear model (GLM) including population structure coefficients derived from STRUCTURE (Q matrix) as covariates; (ii) K model, a mixed linear model (MLM) using K to account for unequal genetic relatedness among genotypes; (iii) QK model, a MLM including both Q and K correcting for population structure and genetic relatedness; (iv) PK model, a MLM including both the first three significant PCs of a PCA of all SNPs as covariates and K. For model comparison, the MSD between the expected and observed p values was computed for each model. The model with the smallest MSD was selected as the most appropriate model (Stich and Melchinger, 2009). For each trait, positive false discovery rate (pFDR; q value) was calculated using the QVALUE in R (Storey and Tibshirani, 2003; Storey, 2002). A q value cut off of 0.1 was used for declaring significant associations. Significant markers with r^2 higher than 0.1–0.2 (varying among chromosomes) were delineated into a single QTL.

3. Results

3.1. Population structure and genetic relatedness between lines

Analysis using a model based approach implemented in STRUCTURE resulted in two optimal subpopulations based on the change of delta k (Fig. 1a). $\text{LnP}(D)$ value was an increasing function of k for all the values observed. However, the biggest increase of $\text{LnP}(D)$ was observed when k increased from one to two (Fig. 1a). Therefore, the association panel was grouped into two subpopulations with 234 and 93 lines, respectively (Fig. 1b). Table S1 listed the lines in each subpopulation and summarizes the number of lines from different regions/programs of the two subpopulations. All the 93 lines in the subpopulation 2 were from IRRI's breeding programs. All the lines from Pakistan (2), India (3), Africa Rice (4), Bangladesh (6), Indonesia (6), Vietnam (10), China (14), CIAT (19) and PhilRice (38) fell into subpopulation 1 along with 132 IRRI lines.

The PCA analysis showed that the first three PCs explained 15.3%, 3.9% and 2.6% of the total variance, respectively. The panel was grouped into two subpopulations (Fig. 2a) according to the first

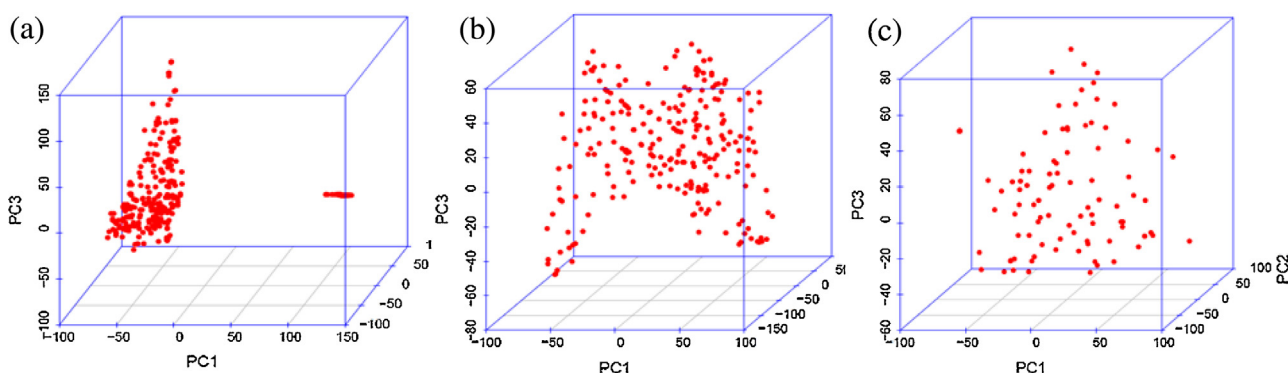


Fig. 2. Population structure detection using a principal component analysis (PCA). (a) The whole panel of 327 lines, with two subpopulations clearly visible (b) Subpopulation 1 of 234 lines. (c) Subpopulation 2 of 93 lines.

Table 1

Chromosome (Chr) size and SNP markers per Chr in the whole population and the two subpopulations inferred by STRUCTURE.

Chr	Size (Mb)	Whole population		Subpopulation 1		Subpopulation 2	
		SNP	Kb/SNP	SNP	Kb/SNP	SNP	Kb/SNP
1	43.3	8610	5.0	5357	8.1	3252	13.3
2	35.9	8035	4.5	5342	6.7	2693	13.3
3	36.4	5041	7.2	3905	9.3	1136	32.0
4	35.5	8664	4.1	6862	5.2	1802	19.7
5	30.0	5977	5.0	4199	7.1	1778	16.8
6	31.1	5629	5.5	4091	7.6	1538	20.2
7	29.6	5207	5.7	3875	7.7	1331	22.3
8	28.4	7000	4.1	5094	5.6	1906	14.9
9	22.6	4908	4.7	3640	6.3	1267	17.8
10	22.8	5104	4.5	3346	6.9	1758	13.0
11	29.0	6663	4.4	4666	6.2	1997	14.5
12	27.4	5614	4.9	3811	7.2	1803	15.2
Total	372.1	76452	4.9	54188	6.9	22261	16.7

PC. The grouping of lines into two subpopulations was the same as the STRUCTURE analysis. PCA analysis conducted within each subpopulation found that the first three PCs explained 5.8% (7.4%), 4.0% (5.7%) and 2.6% (5.2%) of the total variance, respectively for subpopulation 1 (subpopulation 2) (Fig. 2b and c). No obvious substructure was found within either of the subpopulations.

Two distinct subpopulations were also resulted from the NJ tree on the basis of the genetic similarity matrix (Fig. 3). There was very slight difference in assigning genotypes to subpopulation compared to STRCUTURE and PCA. Four lines in the subpopulation 1 derived from STRUCTURE and PCA analyses were assigned to the subpopulation 2 by the NJ tree.

The kinship coefficients for all the pairwise combinations among the 327 lines ranged from 0 to 2, with a mean of 0.58. There were about 0.18% distantly related genotype pairs with kinship being lower than 0.10 and 0.46% highly related genotype pairs with kinship being higher than 1.50 (Fig. 4a). The majority of the genotype pairs had kinship similar to half-sibs (kinship = 0.5). Generally, the relationship within a breeding population was greater than among the breeding populations. The unequal familial relatedness among genotypes could be easily seen in the heatmap (Fig. 4b).

3.2. Linkage disequilibrium

The numbers of markers used were 54,188 and 22,261 for subpopulation 1 and subpopulation 2, respectively (Table 1). In the whole panel, the r^2 estimate was as high as 0.25 (Fig. 5a). The mean r^2 among the individual chromosome ranged from 0.21 to 0.34 (Fig. 5b). LD varied along individual chromosome with regions of high LD interspersing with regions of low LD. As expected, the r^2 value declined with the increasing physical distance between

markers. The average r^2 for the whole genome decreased to half its maximum value (0.25) within 200 kb in the whole panel (Fig. 5a). The LD decay was slower for the chromosomes 8, 12, 1 and 6. The average r^2 dropped to half of its maximum within 250 kb for chromosomes 8 and 12, and 210 kb for chromosomes 1 and 6 (Fig. 5b). LD decay was faster for chromosomes 2, 9 and 10. The average r^2 dropped to half of its maximum within 130 kb for chromosomes 2 and 9 and 150 kb for chromosome 10.

In subpopulation 1 the average r^2 for the whole genome also dropped to half of its maximum (0.36) within 190 kb while it dropped to half of its maximum value (0.65) within 350 Kb in the subpopulation 2 (Fig. 5a). Similar to the whole panel, the LD decay patterns differed among chromosomes within both of the two subpopulations as well (Fig. 5b). The maximum r^2 for individual chromosome varied from 0.30 to 0.45 in the subpopulation 1 and from 0.59 to 0.76 in the subpopulation 2 (Fig. 5b). For subpopulation 1, the LD decay was slower for the chromosomes 12, 3, 4 and 6. The mean r^2 dropped to half of its maximum within 260 kb for chromosome 12, 240 kb for chromosome 4, and 230 kb for chromosomes 4 and 6, respectively (Fig. 5b). LD decay was faster for chromosomes 9, 10, 11, 2 and 8 with the mean r^2 dropping to half of its maximum value within 100 kb, 130 kb, 130 kb, 140 kb and 140 kb, respectively. For the subpopulation 2, a slower LD decay was observed for chromosomes 1, 8, and 12, with r^2 dropping to half of its maximum value within 490 kb, 420 kb, and 400 kb, respectively (Fig. 5b). For chromosomes 9, 5, and 10, the mean r^2 dropped to half of its maximum within 170 kb, 180 kb and 260 kb, respectively.

3.3. Association analysis

3.3.1. Comparison of models

To choose the most appropriate model for association analysis for all tested traits, four models were tested to detect associations between SNP markers and selected traits from selected environments. Since significant population structure was present in our association panel, only four models considering population structure and/or relatedness among the individuals including Q, K, QK and PK models were tested. The four models were used to detect the associations between SNP markers and DTF in the environments DS2, SC, WS1 and WS2 and DTF, GY, PN and TGW in DS2. Model comparison was made using the QQ plot, which is based on the observed and expected p values. The ideal distribution of p values should follow a uniform distribution with less deviation from the expected p values (Kang et al., 2008). Fig. 6a gave the QQ plots for the associations between markers and DTF measured in DS2, SC, WS1 and WS2. The p values of the Q model significantly deviated from the $Y=X$ line for the four environments. The K, QK and PK models showed an improved fit for p values and the three models behaved very similarly in the four environments. Fig. 6b

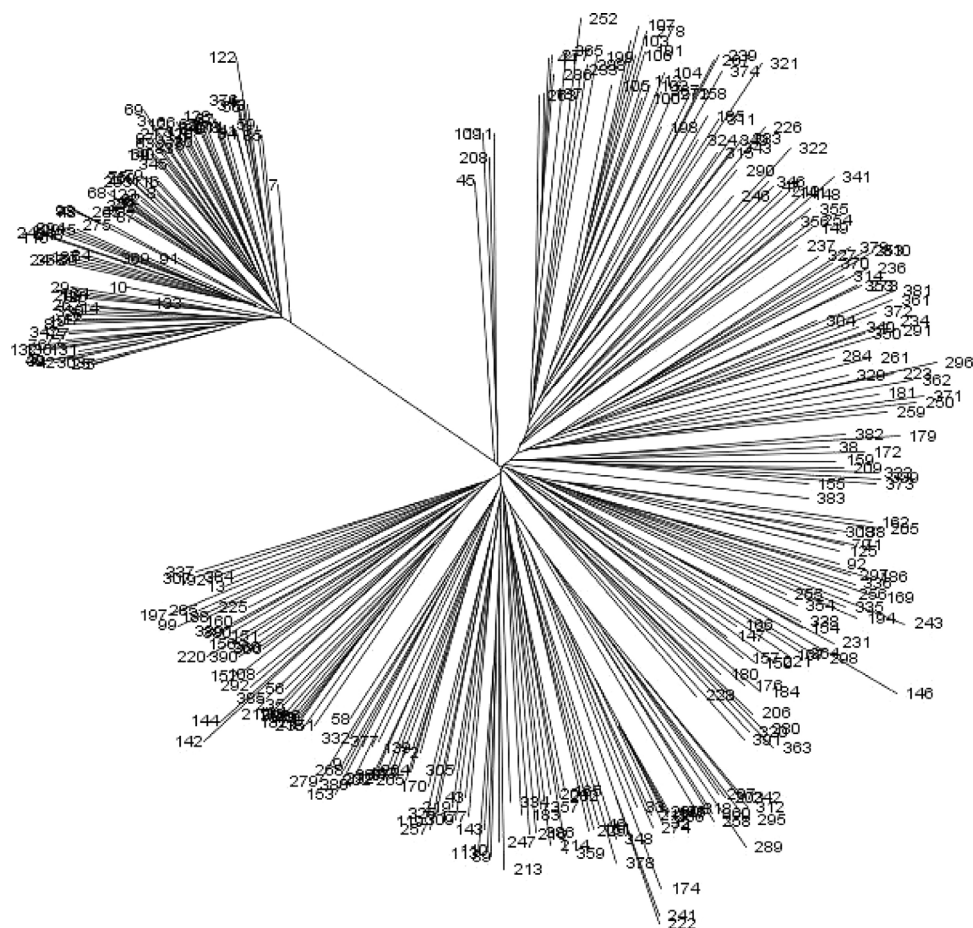


Fig. 3. Unrooted phylogenetic tree depicting the genetic relations among 327 indica lines based on Nei's genetic distance using 76,452 GBS-based SNPs (MAF >5%).

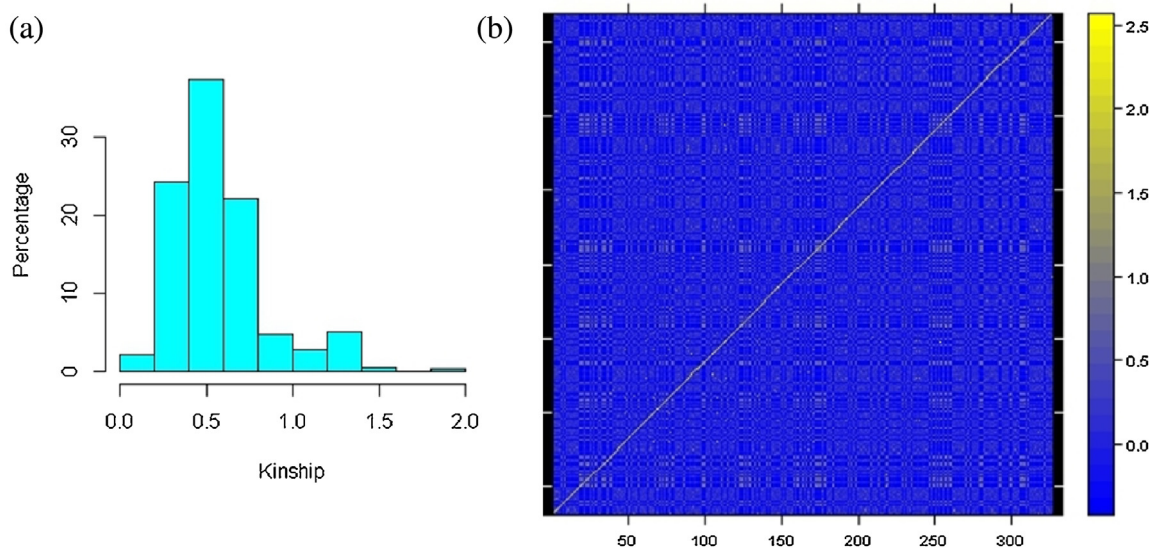


Fig. 4. Distribution (a) and heatmap (b) of relative kinship coefficient across the 327 lines on 76,452 SNPs.

gave the QQ plots for the associations between markers and DTF, GY, PN and TGW measured in DS2. The Q model showed a severe extent of inflation of p values for the four traits, especially for PN. For TGW, the p values of the Q model were close to the other three models, especially at the high value level. The K, QK and PK models behaved very similarly for the four traits in DS2. The result

suggested that MLM using either kinship alone or both kinship and population structure could sufficiently account for structure and cryptic familial relatedness among the individuals in the current association panel. Association analysis was then carried out for all the traits in all the testing environments using the K, QK and PK models. MSD indicated that the most appropriate model was the

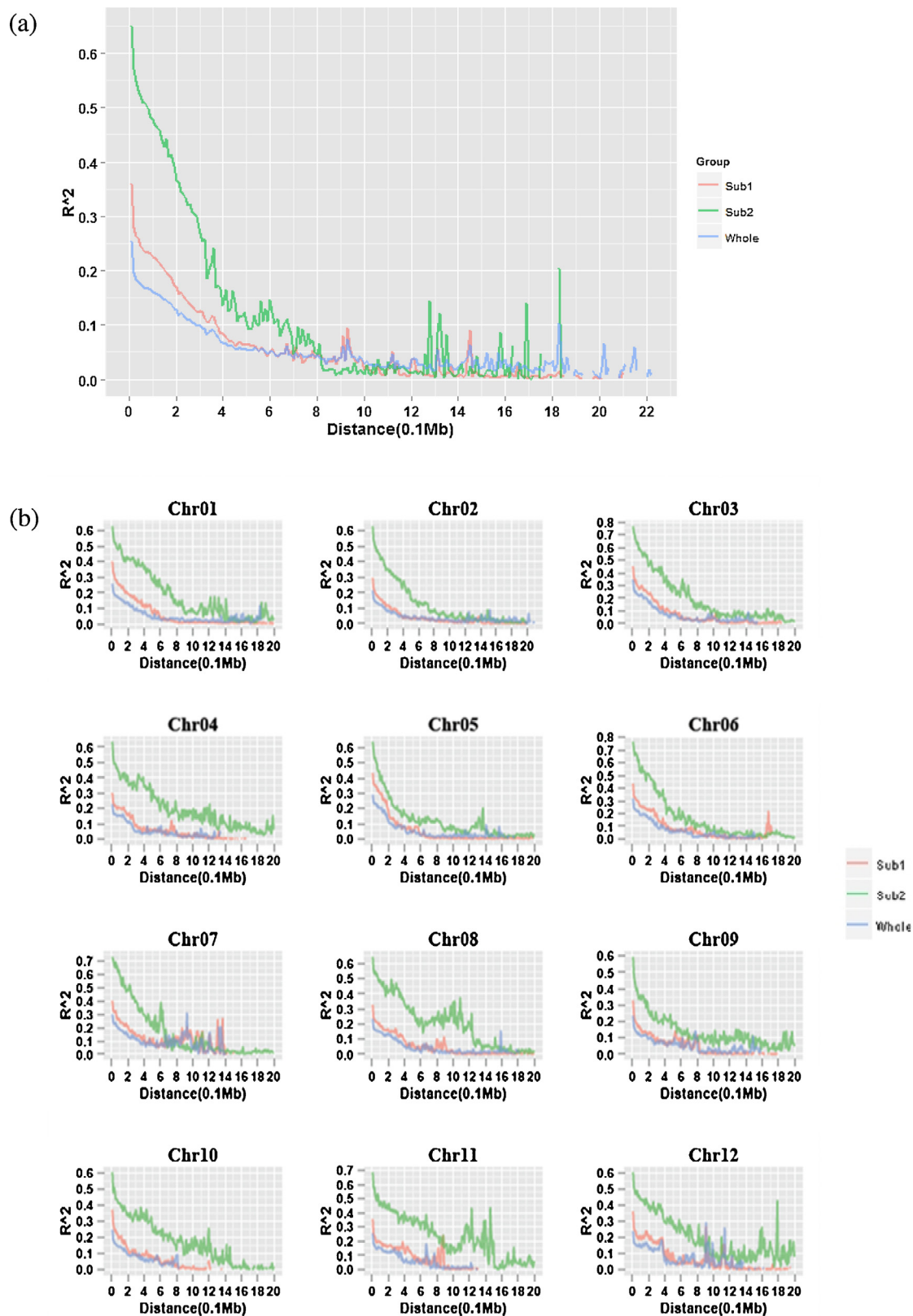


Fig. 5. LD plot for in the whole panel (whole) and the two subpopulations (sub1 and sub2) inferred by STRUCTURE analysis. (a) Whole genome and (b) individual chromosome.

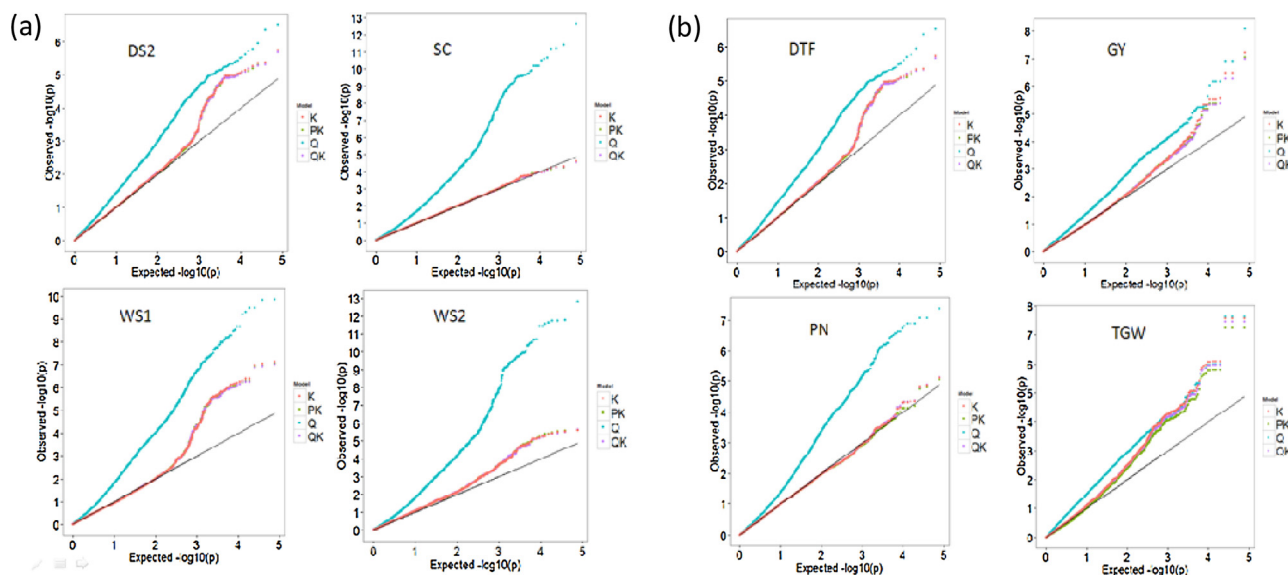


Fig. 6. Quantile-quantile (QQ) plots of Log p values. (a) Four models for GY in environments DS2, SC, WS1 and WS2; (b) Four traits including DTF, GY, PN and TGW in DS2.

PK model (Stich and Melchinger, 2009). Therefore only the results from the PK model were presented.

3.3.2. Marker-trait associations

Wide variation was observed for all the tested traits within the whole panel and the two subpopulations based on the average data across the eight testing environments (Fig. 7 and Table 2). Using the PK model, 452 significant MTA were identified for eight out of the 11 tested traits. The Manhattan plots and QQ plots for GY, DTF and PN in DS2 were shown in Fig. 8 to demonstrate the quality of our GWAS while those for other trait-environment combinations were given in Supplementary Fig. S1–S15. The details of MTA were given in Table S2. By delineating markers with r^2 higher than 0.1–0.2 (varying among chromosomes) into a single QTL, there were total of 43 QTL (Tables 3–5).

For GY, three QTL, $qGY-6$, $qGY-9$, and $qGY-12$ were identified but only in DS2 (Table 3). They were on chromosome 6, 9 and 12, and explained 6.7%, 8.6% and 6.3% of the phenotypic variation, respectively. There were 11 markers within these three QTL regions. The peak SNP for $qGY-6$ is a T/C polymorphism with the C allele increasing GY. The peak SNPs for $qGY-9$ and $qGY-12$ are G/A polymorphisms with the A allele increasing GY. A total of 19, 22 and 16 lines had the increasing alleles, respectively. Four lines, 259, 31, 331 and 330 carry the increasing alleles of all three QTL. Line 259 ranked first in DS2, and lines 31 and 331 ranked fifth and 23rd, respectively. The yield data for line 330 in DS2 was missing due to rat damage.

Fourteen markers displayed significant associations with GN. These markers were grouped into four QTL on chromosome 2, 4, 8 and 9, respectively (Table 3). $qGN-2$, $qGN-4$ and $qGN-8$ were identified in WS1 and explained 7.5%, 7.5% and 5.8% of the phenotypic variation, respectively. $qGN-9$ was detected in WS3 and accounted for 7.6% of the total phenotypic variation. The peak SNPs for the four QTL are T/C, A/G, C/A and A/G polymorphisms with the T, G, A and G alleles increasing GN, respectively. There were 264, 23, 33 and 19 lines carrying the GN-increasing alleles of the peak SNPs of the four QTL, respectively. No line carries the increasing alleles of the peak SNPs of all the four QTL. Four lines, 75, 124, 225 and 355 that carry the increasing alleles of three of the four QTL but $qGN-9$ ranked 22, 32, 35 and 100 in WS1, respectively.

Two QTL involving six markers were significantly associated with pH (Table 3). These two QTL were on chromosome 1 and 9 and explained 6.8% and 7.8% of the phenotypic variation, respectively.

The peak marker of $qpH-1$ is a G/T polymorphism with the G allele increasing pH while the marker of $qpH-9$ is a C/G polymorphism with the C allele being associated with increased pH.

Three markers on chromosome 9 were found to be associated with SN. These three markers were delineated into two QTL, $qSN-9-1$, and $qSN-9-2$ (Table 3). The QTL $qSN-9-1$ was identified in the average environment and explained 7.0% of the phenotypic variation while $qSN-9-2$ was detected in the average environment and WS3, explaining 7.9% and 8.0% of the phenotypic variation, respectively. The peak markers for the two QTL are both G/A polymorphisms with the A allele of the former and the G allele of latter increasing SN.

Twenty three markers on chromosome 11 were found to be significantly associated with TGW but only in SC (Table 3). These markers were delineated into two QTL explaining 5.9% and 7.5% of the total phenotypic variation, respectively. The peak marker of $qTGW-11-1$ is an A/G polymorphism with the A allele increasing TGW while the peak marker of $qTGW-11-2$ is a G/T polymorphism with the T allele increasing TGW. The number of lines carrying increasing allele was 195 and 21, respectively. No line carries both of the increasing alleles.

Twenty-six significantly associated markers in four QTL regions were found for TN (Table 3). These QTL were on chromosome 2, 8, 9 and 9. $qTN-2$ and $qTN-8$ were identified in WS3, explaining 8.9% and 6.3% of the phenotypic variation, respectively. $qTN-9-1$ and $qTN-9-2$ were detected in WS2, explaining 6.2% to 8.7% of the phenotypic variation, respectively. The peak marker of $qTN-2$ is a G/A polymorphism with the A allele increasing TN. The peak markers of $qTN-8$, $qTN-9-1$ and $qTN-9-2$ are all A/T polymorphisms with the T allele of the first two and the A allele of the last one increasing TN. The number of lines carrying the increasing allele was 58, 289, 299 and 19, respectively.

For DTF, nine QTL were identified (Table 4). There were 116 markers in the nine QTL regions. Three QTL were located on chromosome 3. $qDTF-3-1$ was associated with DTF in DS2, WS1, WS2, and WS3, explaining 9.4%, 8.1%, 7.2% and 7.1% of the phenotypic variation, respectively. The peak marker of $qDTF-3-1$ varied in the four environments. $qDTF-3-2$ was associated with DTF in WS1 and explained 4.9% of the phenotypic variation. The peak marker for $qDTF-3-2$ is a G/A polymorphism with the A allele being associated with earlier heading. $qDTF-3-3$ was

Table 2

Mean, standard deviation (SD) and range of yield and 10 related traits estimated in the whole population and the two subpopulations inferred by STRUCTURE.

Trait ^a	Whole population		Subpopulation 1		Subpopulation 2	
	Mean $\pm$ SD	Range	Mean $\pm$ SD	Range	Mean $\pm$ SD	Range
GY	687.6 $\pm$ 84.4	429.8–982.9	675.6 $\pm$ 83.7	429.8–982.9	717.5 $\pm$ 79.0	587.0–962.3
DTF	92 $\pm$ 4.9	79–104	91 $\pm$ 0.04	79–103	94 $\pm$ 4.0	85–104
GN	103.7 $\pm$ 13.6	76–147	104.7 $\pm$ 14.9	76–147	101.3 $\pm$ 9.5	82–125
PB	10.1 $\pm$ 0.9	7–13	10.0 $\pm$ 0.9	7–12	10.5 $\pm$ 0.8	9–13
pH	100.6 $\pm$ 7.2	75.2–124.6	99.7 $\pm$ 7.6	75.2–124	102.7 $\pm$ 5.5	85.8–115.0
PN	9.1 $\pm$ 1.7	6–20	8.8 $\pm$ 1.6	6–14	9.6 $\pm$ 1.9	7–20
SB	24.4 $\pm$ 3.6	15–39	24.5 $\pm$ 3.8	15–39	24.1 $\pm$ 3.0	18–34
SN	124.9 $\pm$ 16.9	88–200	125.3 $\pm$ 18.7	88–200	23.7 $\pm$ 11.4	98–151
SR	82.0 $\pm$ 3.8	71.23–91.02	82.4 $\pm$ 3.6	72.74–89.20	80.9 $\pm$ 34.0	71.23–91.02
TGW	26.69 $\pm$ 2.25	18.21–36.43	26.66 $\pm$ 2.40	18.21–36.43	26.79 $\pm$ 1.8	22.41–32.50
TN	8.7 $\pm$ 1.3	5–16	8.6 $\pm$ 1.4	5–16	8.8 $\pm$ 1.1	7–14

^a See abbreviations list for trait descriptors.**Table 3**

QTL for GY, GN, pH, SN, TGW, and TN identified by GWAS with q value less than 0.1.

Trait ^a	Env ^b	Marker	Chr ^c	Pos ^d	Interval(NM ^e)	QTL	q	R ^{2f}	Major ^g	Minor ^h	Effect ⁱ
GY	DS2	S6_22641374	6	22641374	22.64(3)	<i>qGY-6</i>	0.051	0.067	T	C	–295.01
GY	DS2	S9_22230978	9	22230978	22.19–22.23(9)	<i>qGY-9</i>	0.014	0.086	G	A	–335.09
GY	DS2	S12_16504535	12	16504535	16.50(1)	<i>qGY-12</i>	0.081	0.063	G	A	–320.08
GN	WS1	S2_6491433	2	6491433	6.44–6.56(9)	<i>qGN-2</i>	0.016	0.075	T	C	15.87
GN	WS1	S4_33386030	4	33386030	33.35–33.39(2)	<i>qGN-4</i>	0.016	0.075	A	G	–29.06
GN	WS1	S8_26716249	8	26716249	26.72(1)	<i>qGN-8</i>	0.093	0.058	C	A	–16.24
GN	WS3	S9_22503856	9	22503856	22.50(2)	<i>qGN-9</i>	0.045	0.076	A	G	–30.44
pH	DS1	S1_42469942	1	42469942	42.47–42.62(4)	<i>qpH-1</i>	0.080	0.078	G	T	4.53
pH	DS1	S9_6666230	9	6666230	6.67–6.71(2)	<i>qpH-9</i>	0.087	0.068	C	G	5.44
SN	Ave	S9_19599366	9	19599366	19.60(1)	<i>qSN-9-1</i>	0.044	0.070	G	A	–14.27
SN	WS3	S9_22503856	9	22503856	22.50(2)	<i>qSN-9-2</i>	0.025	0.080	A	G	–32.66
SN	Ave	S9_22503857	9	22503857	22.50(2)	<i>qSN-9-2</i>	0.023	0.079	A	G	–23.97
TGW	SC	S11_20183569	11	20183569	20.00–20.33(22)	<i>qTGW-11-1</i>	0.031	0.075	A	G	2.29
TGW	SC	S11_26273587	11	26273587	26.27(1)	<i>qTGW-11-2</i>	0.095	0.059	G	T	–2.68
TN	WS3	S2_31613390	2	31613390	30.98–31.75(7)	<i>qTN-2</i>	0.013	0.089	G	A	–1.50
TN	WS3	S8_26164899	8	26164899	26.16(1)	<i>qTN-8</i>	0.083	0.063	A	T	–1.31
TN	WS2	S9_6918094	9	6918094	6.92(2)	<i>qTN-9-1</i>	0.044	0.062	A	T	–1.23
TN	WS2	S9_12223796	9	12223796	12.21–12.27(16)	<i>qTN-9-2</i>	0.007	0.087	T	A	–1.77

^a See abbreviation list for trait descriptors.^b Env, environment in which the QTL was detected (see abbreviation list for environment descriptor).^c Chr, Chromosome.^d Pos, physical position (in bp).^e NM, number of significant SNP markers within the QTL interval.^f R², variance explained by the marker.^g Major, major allele.^h Minor, minor allele.ⁱ Effect, effect of the major allele of the most significant marker within the QTL interval.**Table 4**

QTL for DTF identified by GWAS with q value less than 0.1.

Env ^a	Marker	Chr ^b	Pos ^c	Interval(NM ^d)	QTL	q	R ^{2e}	Major ^f	Minor ^g	Effect ^h
WS1	S1_22687304	1	22687304	22.69(1)	<i>qDTF-1</i>	0.058	0.052	C	G	–3.38
DS2	S3_550877	3	550877	0.26–1.40(204)	<i>qDTF-3-1</i>	0.050	0.072	T	G	–5.71
WS3	S3_885458	3	885458	0.26–1.40(204)	<i>qDTF-3-1</i>	0.014	0.081	G	T	5.39
WS2	S3_1166378	3	1166378	0.26–1.40(204)	<i>qDTF-3-1</i>	0.051	0.071	C	T	–3.72
WS1	S3_1166378	3	1166378	0.26–1.40(204)	<i>qDTF-3-1</i>	0.003	0.094	C	T	–5.81
WS1	S3_1682760	3	1682760	1.68(1)	<i>qDTF-3-2</i>	0.071	0.049	G	A	3.55
WS2	S3_34390019	3	34390019	34.39(1)	<i>qDTF-3-3</i>	0.058	0.060	T	C	4.47
DS2	S3_34390019	3	34390019	34.39(1)	<i>qDTF-3-3</i>	0.091	0.053	T	C	4.82
WS3	S4_31516736	4	31516736	31.52(2)	<i>qDTF-4</i>	0.013	0.085	T	A	–4.85
WS2	S4_31516736	4	31516736	31.52(2)	<i>qDTF-4</i>	0.058	0.062	T	A	–3.80
WS1	S4_31516736	4	31516736	31.52(2)	<i>qDTF-4</i>	0.011	0.066	T	A	–4.00
WS2	S5_20558247	5	20558247	20.56(2)	<i>qDTF-5</i>	0.062	0.058	T	A	–3.79
WS1	S5_20558247	5	20558247	20.56(2)	<i>qDTF-5</i>	0.006	0.077	T	A	–4.81
WS1	S7_20665062	7	20665062	18.77–20.66(2)	<i>qDTF-7</i>	0.033	0.057	C	T	3.07
WS1	S9_7844994	9	7844994	7.84(1)	<i>qDTF-9</i>	0.052	0.051	C	T	5.54
JX	S10_2734797	10	2734797	2.73–2.93(8)	<i>qDTF-10</i>	0.067	0.083	T	C	6.14

^a Env, environment in which the QTL detected (see abbreviation list for environment descriptor).^b Chr, Chromosome.^c Pos, physical position (in bp).^d NM, number of significant SNP markers within the QTL interval.^e R², variance explained by the marker.^f Major, major allele.^g Minor, minor allele.^h Effect, effect of the major allele of the most significant marker within the QTL interval.

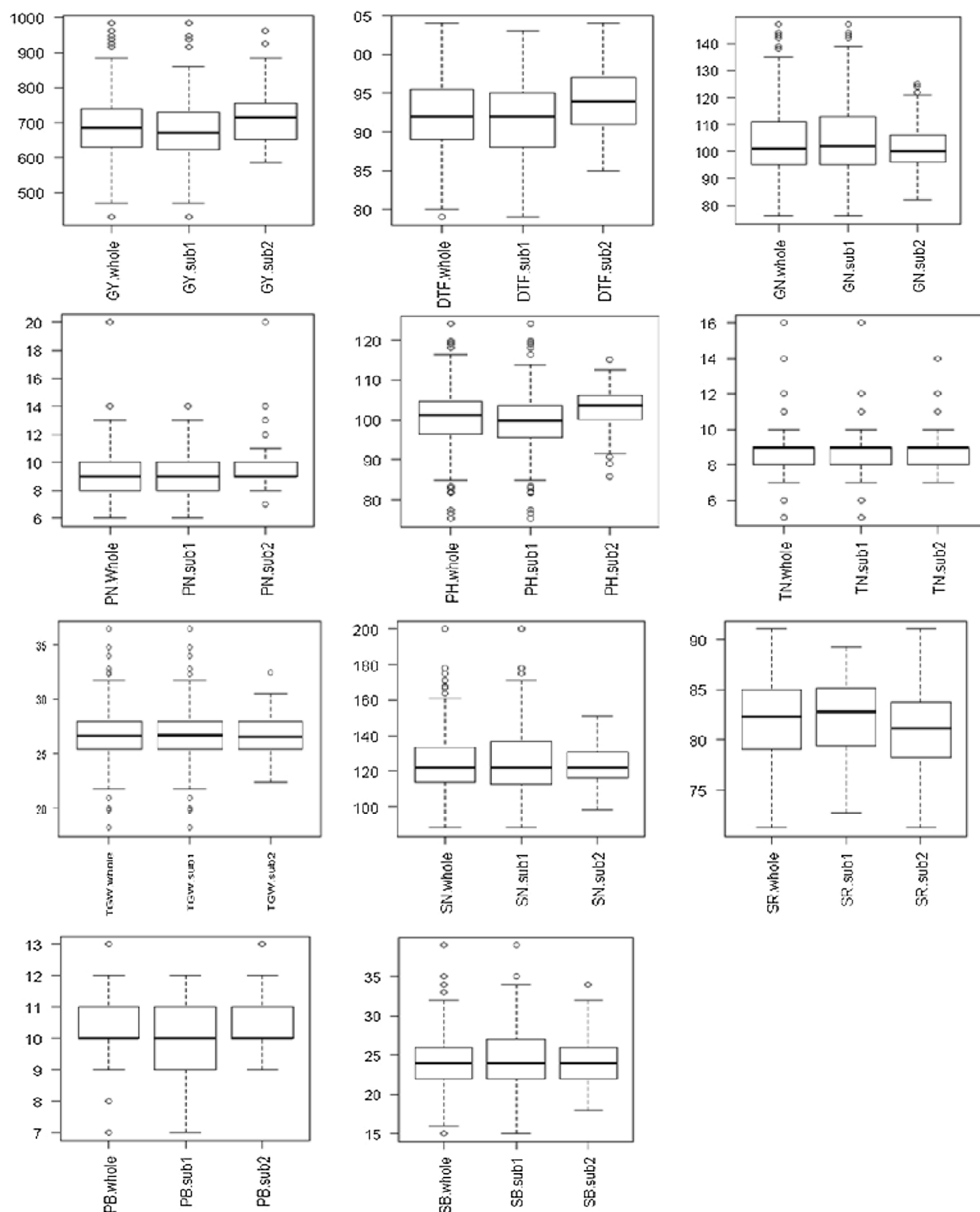


Fig. 7. Box plots showing the distribution of average BLUEs (y axis) of GY, DTF, pH, GN, PN, PB, SB, SN, SR, TGW and TN within whole panel (whole), subpopulation 1 (sub1) and subpopulation 2 (sub2) inferred by STRUCTURE.

detected in DS2 and WS2 and explained 6.0% and 5.3% of the phenotypic variation, respectively. The peak marker for *qDTF-3-3* is a T/C polymorphism with the C allele promoting heading. *qDTF-3-4* was associated with DTF in WS1, WS2, and WS3, explaining 6.6%, 6.2% and 8.5% of the phenotypic variation, respectively. The

peak marker of *qDTF-3-4* is a T/A polymorphism with the T allele promoting heading. *qDTF-3-5* was identified in WS1 and WS2, explaining 7.7% and 5.8% of the phenotypic variation, respectively. The peak marker of *qDTF-3-5* is a T/A polymorphism with the T allele promoting heading. *qDTF-3-7* and *qDTF-3-9* were identified

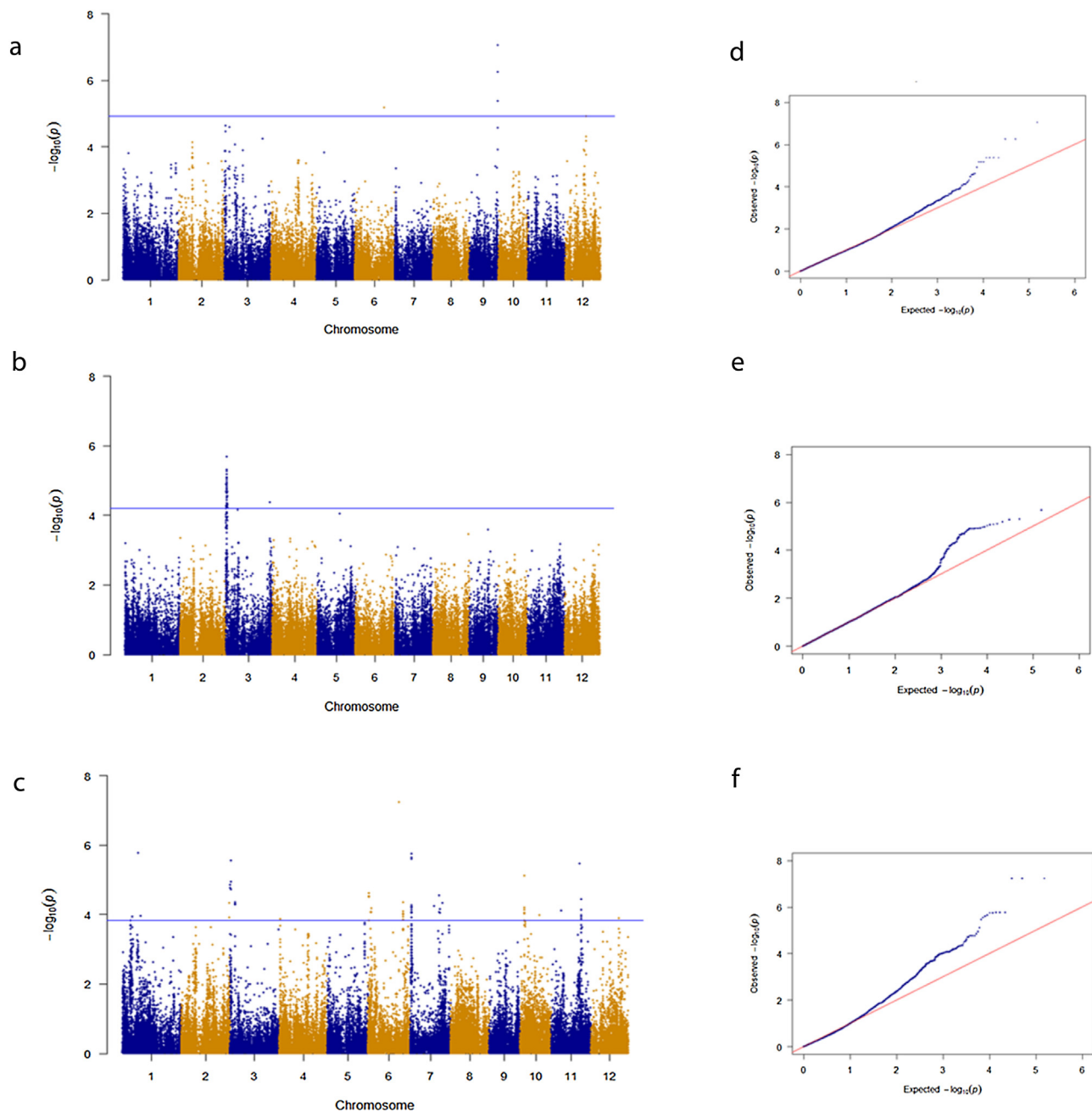


Fig. 8. Manhattan plots and QQ plots resulting from the GWAS for GY (a, d), DTF (b, e), and PN (c, f) in DS2.

in WS1 and explained 5.7% and 5.1% of the phenotypic variation, respectively. The peak markers of *qDTF-3-7* and *qDTF-3-9* are C/T polymorphism with the T allele promoting heading. In JX, one QTL, *qDTF-3-10*, was identified to be associated with DTF and explained 8.3% of the phenotypic variation. The peak SNP is a T/C polymorphism with the C allele promoting heading. Combinations of alleles promoting/delaying heading resulted much earlier/late heading date. For example, for the latest two lines (DTF=104) in WS1, six of the seven alleles of respective QTL were associated with late heading.

For PN, 26 QTL containing 134 markers were identified on chromosome 1, 2, 3, 4, 6, 7, 9, 10 and 11 (Table 5). The percentage of phenotypic variation explained ranged from 4.6% (*qPN1-1* in DS2) to 9.6% (*qPN-6-3* in DS2). A single QTL was identified on chromosome 9 (*qPN-9*) and 12 (*qPN-12*), respectively. *qPN-9* and *qPN-12* explained 8.0% and 4.9% of the phenotypic variation, respectively.

The peak marker for *qPN-9* is a T/A polymorphism with the A allele increasing PN. The peak SNP for *qPN-12* is an A/C polymorphism with the A allele increasing PN. Two QTL were identified on chromosome 2 (*qPN-2-1* and *qPN-2-2*), 4 (*qPN-4-1* and *qPN-4-2*) and 10 (*qPN-10-1* and *qPN-10-2*), respectively. *qPN-2-1* was identified in WS2 and WS3, accounting for 5.6% and 8.1% of the phenotypic variation, respectively. Different markers peaked in this QTL region in the two environments. The peak SNP is an A/T polymorphism with the T allele increasing PN in WS2, while is a G/A polymorphism with the A allele increasing PN in WS3. *qPN-2-2* was identified in DS2 and explained 5.3% of the phenotypic variation. The associated SNP is a C/T polymorphism with the T allele increasing PN. *qPN-4-1* was detected in DS2 and explained 4.7% of the phenotypic variation. The peak marker is an A/G polymorphism with the G allele increasing PN. *qPN-4-2* was detected in WS2 and explained 5.8% of the phenotypic variation. The peak marker is a G/T polymorphism with the G

Table 5
QTL for PN identified by GWAS with q value less than 0.1.

Env ^a	Marker	Chr ^b	Pos ^c	Interval(NM ^d)	QTL	q	R ² ^e	Major ^f	Minor ^g	Effect ^h
DS2	S1.7023757	1	7023757	6.19–7.02(2)	<i>qPN-1-1</i>	0.092	0.047	C	T	–3.13
DS2	S1.11793133	1	11793133	11.79(3)	<i>qPN-1-2</i>	0.017	0.074	C	T	–4.31
Ave	S1.11793133	1	11793133	11.79(3)	<i>qPN-1-2</i>	0.059	0.066	C	T	–1.77
DS2	S1.13102888	1	13102888	13.10(1)	<i>qPN-1-3</i>	0.091	0.048	A	C	–3.76
WS2	S1.18835949	1	18835949	18.84(1)	<i>qPN-1-4</i>	0.032	0.071	T	G	–1.59
WS2	S2.31105099	2	31105099	31.11–31.61(2)	<i>qPN-2-1</i>	0.097	0.056	A	T	–1.07
WS3	S2.31613390	2	31613390	31.11–31.61(2)	<i>qPN-2-1</i>	0.037	0.081	G	A	–1.57
DS2	S2.35210846	2	35210846	35.21–35.24(2)	<i>qPN-2-2</i>	0.085	0.053	C	T	–3.28
DS2	S3.519551	3	519551	0.34–0.52(11)	<i>qPN-3-1</i>	0.019	0.070	G	T	–4.16
DS2	S3.3849936	3	3849936	3.85(5)	<i>qPN-3-2</i>	0.085	0.053	G	C	–3.39
WS2	S3.20032981	3	20032981	20.03(2)	<i>qPN-3-3</i>	0.073	0.059	A	G	–1.45
DS2	S4.456956	4	456956	0.46–0.56(4)	<i>qPN-4-1</i>	0.095	0.047	A	G	–3.34
WS2	S4.18880870	4	18880870	18.88(1)	<i>qPN-4-2</i>	0.091	0.058	G	T	1.21
DS2	S6.658071	6	658071	0.47–0.66(4)	<i>qPN-6-1</i>	0.077	0.057	C	T	–3.80
DS2	S6.2174568	6	2174568	2.12–2.19(6)	<i>qPN-6-2</i>	0.085	0.051	T	C	–3.16
DS2	S6.22641374	6	22641374	22.64(3)	<i>qPN-6-3</i>	0.001	0.096	T	C	–4.90
Ave	S6.22641374	6	22641374	22.64(3)	<i>qPN-6-3</i>	0.013	0.080	T	C	–1.80
DS2	S6.26068703	6	26068703	26.07–26.16(15)	<i>qPN-6-4</i>	0.085	0.053	A	G	–4.04
DS2	S7.858133	7	858133	0.86–1.16(17)	<i>qPN-7-1</i>	0.017	0.073	A	G	–4.39
DS2	S7.17450613	7	17450613	17.45(1)	<i>qPN-7-2</i>	0.085	0.052	A	G	4.19
DS2	S7.21459811	7	21459811	21.41–22.14(4)	<i>qPN-7-3</i>	0.085	0.056	G	C	–3.89
DS2	S7.23435774	7	23435774	23.44(1)	<i>qPN-7-4</i>	0.085	0.055	A	T	–3.65
WS2	S9.12223796	9	12223796	12.21–12.26(16)	<i>qPN-9</i>	0.020	0.080	T	A	–1.69
DS2	S10.2888145	10	2888145	2.83–3.0(13)	<i>qPN-10-1</i>	0.045	0.064	T	C	–4.23
DS2	S10.14091112	10	14091112	14.09(1)	<i>qPN-10-2</i>	0.085	0.048	T	C	–3.07
DS2	S11.7252823	11	7252823	7.25(1)	<i>qPN-11-1</i>	0.085	0.050	C	T	–3.45
DS2	S11.20916392	11	20916392	20.92(1)	<i>qPN-11-2</i>	0.022	0.070	A	C	–4.62
DS2	S11.22242728	11	22242728	22.01–22.24(15)	<i>qPN-11-3</i>	0.085	0.054	G	C	–3.90
DS2	S12.20919811	12	20919811	20.92(1)	<i>qPN-12</i>	0.094	0.049	C	A	–3.187

^a Env, environment in which the QTL detected (see abbreviation list for environment descriptor).^b Chr, Chromosome.^c Pos, physical position (in bp).^d NM, number of significant SNP markers within the QTL interval.^e R², variance explained by the marker.^f Major, major allele.^g Minor, minor allele.^h Effect, effect of the major allele of the most significant marker within the QTL interval.

allele increasing PN. The two QTL on chromosome 10 were detected in DS2 and explained 6.4% and 4.8% of the phenotypic variation, respectively. The peak markers are T/C polymorphisms with the C allele increasing PN. Three QTL were detected on chromosomes 3 (*qPN-3-1*, *qPN-3-2* and *qPN-3-3*) and 11 (*qPN-11-1*, *qPN-11-2* and *qPN-11-3*). *qPN-3-1* and *qPN-3-2* explained 7.0% and 5.3% of the phenotypic variation in DS2, respectively. For *qPN-3-1*, the peak marker, is a G/T polymorphism with the T allele increasing PN. For *qPN-3-2*, the peak SNP is a G/C polymorphism with C allele increasing PN. *qPN-3-3* was detected in WS2 and explained 5.9% of the phenotypic variation. The peak marker is an A/G polymorphism with the G allele increasing PN. The three QTL on chromosome 11 were detected in DS2 and explained 5.0%, 7.0% and 5.4% of the phenotypic variation, respectively. The peak markers are C/T, A/C and G/C polymorphisms with PN increasing allele being T, C and C, respectively. Four QTL were identified on chromosome 1 (*qPN-1-1*, *qPN-1-2*, *qPN-1-3* and *qPN-1-4*), 6 (*qPN-6-1*, *qPN-6-2*, *qPN-6-3* and *qPN-6-4*) and 7 (*qPN-7-1*, *qPN-7-2*, *qPN-7-3* and *qPN-7-4*), respectively. *qPN-1-1*, *qPN-1-2* and *qPN-1-3* were identified in DS2 and explained 4.7%, 7.4% and 4.8% of the phenotypic variation, respectively. The three corresponding peak markers are all C/T polymorphisms with the T allele increasing PN. *qPN-1-4* was detected in WS2 and explained 7.1% of the phenotypic variation. The peak SNP is a T/G polymorphism with the G allele increasing PN. The four QTL on chromosome 6 identified in DS2 explained 5.1% – 9.6% of the phenotypic variation. *qPN-6-3* was also detected in the average environment and explained 8.0% of the phenotypic variation. The corresponding peak markers are all T/C polymorphisms with the C allele increasing PN. The peak marker of *qPN-6-4* is an A/G polymorphism with the G allele increasing PN. Four QTL on chro-

mosome 7 were identified in DS2 only, and explained 7.3%, 5.2%, 5.6% and 5.5% of the phenotypic variation, respectively. The peak marker, is A/G, A/G, G/C and A/T polymorphisms with the increasing allele being G, A, C and T, respectively.

Multi-trait QTL were detected in many chromosome regions. Their chromosome positions were shown with other trait-specific QTL in Fig. 9. Clusters of QTL were identified on chromosomes 2, 3, 6, 8, 9, 10 and 11. GY and PN had a common QTL on chromosome 6. GY also shared a QTL on chromosome 9 with GN and SN. Traits DTF and PN shared a common QTL on chromosome 3. DTF had a QTL in common with PN on chromosome 10. Traits PN and TN shared two common QTL on chromosomes 2 and 9. A QTL region on chromosome 8 was associated with GN and TN. There was a QTL on chromosome 9 for pH and TN.

4. Discussion

4.1. Population structure and unequal familial relationship were present

AM has emerged as a powerful approach for dissecting the genetic basis of complex traits in plants and GWAS is becoming more and more feasible with the advance of high throughput genotyping. However associations detected in AM are often spurious because population structure and unequal relatedness among individuals in a given panel are still remaining a confounding factor (Zhang et al., 2010). In order to reduce false positives, population structure and familial relatedness should be carefully taken into account in AM. The strong population structure and genetic relatedness within plant samples might be affected by many fac-

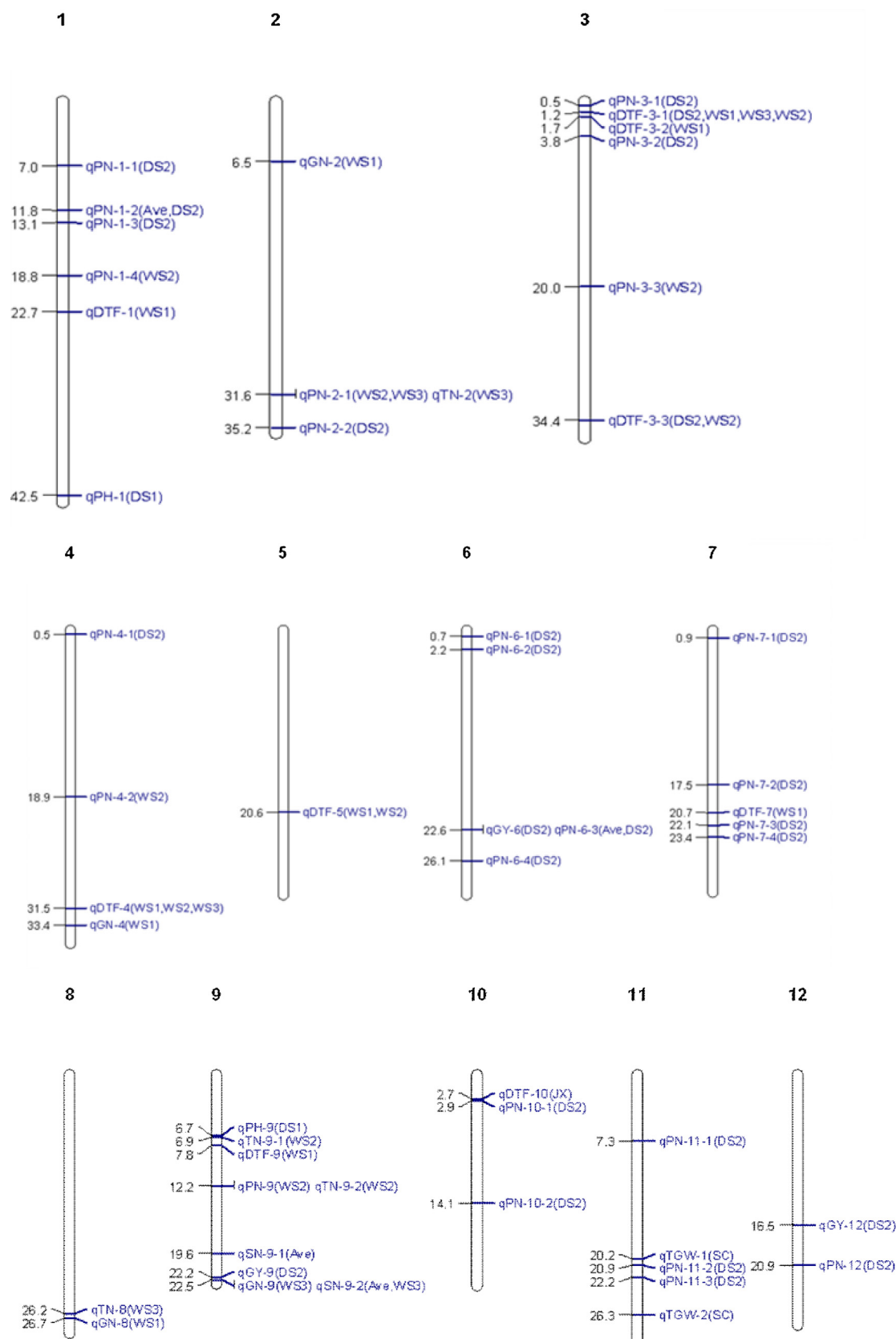


Fig. 9. Chromosomal regions of significant QTL identified. The environment in which the QTL was identified was given in the bracket; the number on the left of the chromosome indicated the physical position (Mb).

tors such as geographic origins, reproductive nature and artificial selection (Atwell et al., 2010). In the present study, the lines used were from different breeding programs in different parts of the world, with most from IRRI and PhilRice. Although the association panel was from one subspecies, *indica*, the subdivision within the subspecies was still detectable and the population structure still played an important role in QTL identification, as illustrated by the association results from different models with or without population structure coefficients as covariates (see next section). In our study, two consistent subpopulations were detected by all three of methods including STRUCUTRE, PCA and NJ tree. The presence of population structure in populations of *indica* rice varieties has been reported using different panels and marker types (Wang et al., 2014; Xie et al., 2012; Zhang et al., 2013). The existence of population structure within *indica* rice was not out of expectation. In fact, for the majority of crop species and particularly self-pollinated crop species, panels of varieties and advanced breeding lines have population structure and unequal familial relationship (Cockram et al., 2010; Edae et al., 2014; Flint-Garcia et al., 2005; Pasam et al., 2012), partially caused by the selection for local ecological environments, spatial isolation and assortative mating (i.e. good $\times$ good) adopted by breeders.

Big differences were found for the genetic relatedness (estimated as relative kinship) between many pairs of lines (Fig. 4). This was expected, since rice breeders usually mate a small number of elite lines in many crosses. Moreover, unequal familial relationship is another important confounding factor in AM using breeding materials.

4.2. Linkage disequilibrium was comparable to other panels

Successful GWAS also depends on the extent of LD and the number and distribution of markers. It is recommended that the average marker density should be higher than the LD decay in order to achieve a successful GWAS (Wang et al., 2005; Zhu et al., 2008). Significant LD was detected across the genome of the current rice panel. The patterns of LD varied among chromosomes and subpopulations. For the whole panel the maximum r^2 was 0.25 of the whole genome and ranged from 0.21 to 0.34 among the individual chromosome. Subpopulation 1 showed generally lower LD ($r^2 = 0.36$) than subpopulation 2 ($r^2 = 0.65$) for the whole genome. The genome-wide LD decay in the whole population was around 200 kb, while it was 190 kb in the subpopulation 1 and 350 kb in the subpopulation 2. This was comparable to the estimates reported previously using different *indica* rice panels (Kumar et al., 2015; Mather et al., 2007; McNally et al., 2009; Phung et al., 2014; Xu et al., 2012). In a panel of 220 accessions which consisted mostly of the *indica* accessions, the LD decayed to half of its maximum value at around 300 kb, with the LD ranging from 69 kb on chromosome 3–125 kb on chromosome 8 and 12 (Kumar et al., 2015). Significant differences in LD decay among chromosomes were also found in the present study, ranging from 130 kb for chromosomes 2 and 9–250 kb for chromosomes 8 and 12. The differences in LD were mainly caused by the difference in recombination rate, selection history or structural difference among populations/panels. An uneven marker distribution and the difference in marker MAF could also contribute to the large variance associated with estimated LD decay.

4.3. Familial relationship was more important than population structure

Several models that were developed to reduce false associations in AM have been successfully applied in previous studies (Price et al., 2006; Pritchard et al., 2000b; Yu et al., 2006). The most commonly used ones in plant association studies are the Q, K, QK

and PK models. The Q model accounts for (large-scale) population structure only while the K model takes into account of cryptic familial relatedness only. The QK and PK models are the most stringent because they control both population structure and cryptic familial relatedness. Most of applied plant geneticists have used the QK model, even though the QK model may reduce the power of detecting associations in some cases due to a loss of degrees of freedom (Yu et al., 2006). Indeed, there is not a universal best model for all populations, traits and growing environments.

In the present study, the Q model was significantly worse than the other three models, which behaved similarly. This observation was consistent to Yu et al. (2006) who reported that the GLM approach with the Q matrix (Q model) was probably more sensitive than the MLM approach (QK) in some cases (Stich et al., 2008). The effects of Q matrix and K matrix for controlling false associations were different for different rice panels. Shao et al. (2011) reported that for their whole rice panel including red and black rice, the QK model performed similar to the Q model, with only slightly fewer significant markers identified than the Q model. In the white rice panel, the QK model performed well with 60% reduction in significant markers at different p values. Li et al. (2011) found that taking the kinship into account did not improve the mapping QTL for GY using the USDA rice mini-core collection. The K model performed the worst. The PCA model, which controlled the effect of population structure through PCA, showed the best approximation to the expected cumulative distribution of p values, followed by the PK, Q and QK models. This might be due to the fact that the accessions of the mini-core collection were not genealogically related, since the core-collection was developed by maximizing the genetic diversity. In such cases, population structure was the main confounding factor and as a result the models taking into account of population structure performed better than K model which accounts for unequal relatedness only.

For GWAS in rice, we expected that for association panels consisting of breeding lines of a single subspecies/ecotype, unequal relatedness among genotypes was the main confounding factor and as a result the K model could be sufficient and for panels which consisted of accessions from different subspecies, QK or PK models might be better since both population structure and cryptic relatedness were likely to be important.

4.4. Novel QTL for grain yield and other traits were identified

A total of 43 QTL for eight out of 11 traits were identified with no QTL being identified for PB, SB and SR. Those QTL were involved in 18 trait-environment combinations. The numbers of QTL detected for various traits varied from two (pH, SN and TGW) to 26 (PN). Four of the identified QTL in the present study were located in the genomic regions where QTL for yield or related traits have been reported previously. The DTF QTL *qDTF-3-1* detected on chromosome 3 corresponds to the location of *Hd9*, a known QTL for heading date (Lin et al., 2002). Another DTF QTL *qDTF-4* identified on chromosome 4 located within the region of previously reported *qGN4-1* for GN (Deshmukh et al., 2010). The PN QTL *qPN-1-1* identified on chromosome 1 was located in the region of known *Gw1-1* for TGW (Yu et al., 2008). The QTL region between 26.2 and 26.7 Mb of chromosome 8, which affected GN and TN, contains the known gene *OsSPL16* (Wang et al., 2012), which is synonymous with *GW8* and controls grain size and shape. *OsSPL16* encodes a protein that positively regulates cell proliferation. Over expression of this gene promotes cell division and grain filling, with positive consequences for grain width and GY (Wang et al., 2012). Further studies are needed to find whether *OsSPL16* is the underlying candidate gene for GN and TN.

A total of 39 QTL were found in chromosomal regions where none have been reported previously. Considering the large num-

ber of trait-environment combinations in the present study this number was small. Nevertheless, our results provided further evidence that novel QTL/MTA could be identified using breeding population. This was the second rice GWAS study using a breeding population as the association panel. Comparing to the first study by reported Begum et al. (2015) using breeding population as association panel, the lines in the present panel were more diverse with two distinct subpopulations. Even though the identified QTL accounted for a relatively small proportion of the phenotypic variation ($R^2 = 4.7\%$ to 9.6%), they should be effective in breeding programs since they were identified from a heterogeneous population with relatively small numbers of lines being from various breeding programs and a stringent statistical model was used.

The current panel of 327 lines is small and has relatively low genetic diversity (Liang et al., 2016), since the majority of its lines were from IRRI's breeding program. QTL with large effects might have been fixed or their MAF were very low. For instances, there were only 22 out of 327 lines that had the favourable allele for *qPN-6-3* which accounted for the largest proportion of the phenotypic variation (9.6%) among the detected QTL. For the largest effect QTL for GY on chromosome 9, which explained 8.6% of the phenotypic variation in DS2, only 22 lines carried the favourable allele for increasing GY. In some QTL regions, the LD between the markers and the trait-controlling genes were low due to low marker density. The QTL *qPN-1-1* covered an interval of 0.83 Mb on chromosome 1 with 329 markers within this region and the marker density was about 2.5 kb/SNP. The QTL *qPN-6-3* only contained three significant markers and the marker density in the 0.83 Mb interval was around 8.4 kb/SNP. Therefore, it is essential for increasing genetic diversity of the base population (Liang et al., 2016) and for gene discovery to add more diverse lines. One of the options for increasing genetic diversity of the current panel (base breeding population) is to have more lines from large *indica* breeding programs in the subtropical regions such as China (Liang et al., 2016). However, increasing sample size by pooling samples from different breeding programs may not necessarily increase the power to detect MTA, as demonstrated in wheat (Mohammadi et al., 2015). MTA identifiable in one of the population may become undetectable in the combined population. Mohammadi et al. (2015) did identify new QTL using the combined datasets of three programs. Our own study in rice in a slightly different context also found that many new MTA were identified by using the whole population but not the subpopulations, although some MTA were only identifiable in the subpopulations (Liang et al., 2016).

4.5. QTL-by-environment interaction was significant

Testing environments showed significant effects on the association analysis. Among the 43 QTL identified, 25 QTL were detected in DS2, 11 in WS2, 10 in WS1, seven in WS3, two in DS1 and SC, respectively, and one in JX. No QTL was detected in DS3. Only three QTL were detected from the average values from all the environments. The medium nitrogen rate yielded the highest number of significant QTL, which was 25 in DS2 and 11 in WS2, followed by no nitrogen. The majority of the QTL for a trait of interest were identified only in one environment with none of them being identified from all the testing environments. QTL-by-environment interaction (QEI) has been reported in previously reported rice GWAS. Zhao et al. (2011) found that only one of the 10 genomic regions associated with candidate genes for flowering time was detected in more than one location and that several GWAS peaks associated with candidate genes were significant in only a single year at the same location. In a different GWAS conducted in two environments using the USDA rice mini-core subset, only six out of the 60 markers detected from two environments were significant in both

environments (Nawaz et al., 2015). The environment-specific QTL could serve as the genetic variability that could be used to produce optimal cultivars once the beneficial stable major genes have been fixed in breeding programs (Pauli et al., 2014).

4.6. QTL affecting multiple traits were present

Eight of the identified QTL were associated with more than one trait with five were associated with strongly correlated traits measured in the same or different environments. Most of the QTL for DTF were located on a similar position to these for some of the yield components (PN, GN and TN), suggesting that some maturity genes may have significant effects on yield through pleiotropic effects on yield components. The QTL for PN on chromosome 6 and QTL for GN and SN on chromosome 9 were located at the same or similar position to that for GY, thus selecting for these QTL should be effective in improving GY. They were the QTL region of 22.64 Mb on chromosome 6 (*qGY-6* or *qPN-6-3*) affecting GY in DS2 and PN in average environment, the genomic region between 31.08 Mb and 31.61 Mb on chromosome 2 (*qPN-2-1* or *qTN-2*) affecting PN in WS2 and WS3 and TN in WS3, the QTL region between 26.16 Mb and 26.71 Mb of chromosome 8 (*qGN-8* or *qTN-8*) affecting GN in WS1 and TN in WS3, QTL region of 12.22 Mb on chromosome 9 (*qPN-9* or *qTN-9-2*) affecting PN and TN in WS2 and the QTL region between 22.23 Mb and 22.50 Mb on chromosome 9 (*qGY-9* or *qGN-9*) affecting GY in DS2 and GN in WS3. Three QTL affected different traits with a low correlation. They were QTL region between 0.52 Mb and 1.27 Mb on chromosome 3 (*qPN-3-1* or *qDTF-3-1*) affecting PN and DTF in DS2, the genomic region between 6.70 Mb and 6.91 Mb on chromosome 9 (*qPH-9* or *qTN-9-1*) affecting pH in DS1 and TN in WS2, and the genomic region between 2.73 Mb and 2.89 Mb on chromosome 10 (*qDTF-10* or *qPN-10-1*) affecting DTF in JX and PN in DS2. Multi-trait QTL were often identified in rice genetic studies. Most of the traits sharing same QTL were closely related. For examples, grain length, width, and length-breadth ratio (Begum et al., 2015); surface area of unhusked seeds, length of unhusked seeds and TGW (Yonemaru et al., 2014); filled grains, productive tillers, yield/plant, percent spikelet fertility under stress and percent spikelet fertility SSI (Kumar et al., 2015). Multi-trait QTL were also found for those showing no obvious relationships, i.e. amylase content and flowering time with blast resistance and flag leaf width (Zhao et al., 2011); heading date with surface area of unhusked seeds (Yonemaru et al., 2014). Pleiotropic effects of a single gene and linkage of QTL for different traits are the biological mechanisms of multi-trait QTL. Considering that the present panel consisted of lines from different well established breeding program, linkage might be a more likely mechanism. Rice breeders commonly make crosses between similar phenotype (assortative mating) and as a result many linkage blocks were formed and maintained. These linkage blocks also represent untapped genetic variability that could be utilized through breakage of these blocks through recombination, and lead to new genetic gain in populations with unfavourable linkage (Simmonds, 1979). Without cloning or mapping with extremely high resolution it is impossible to tell which of the mechanism is responsible.

Both pleiotropic effects and linkage of the identified QTL can be either beneficial or troublesome in plant breeding (Guo and Ye, 2014). QTL alleles with favourable effects on different traits are clearly beneficial. Similarly, favourable alleles linked in desirable phase form valuable linkage blocks that would be explored more easily in practical breeding. On the other hand, for QTL linked in undesirable phase more efforts are needed to break the linkage, which is the main impediment of using QTL identified from unadapted germplasm in breeding. An allele with favourable effects on some target traits but unfavourable effects on other traits could be utilized if its negative effects could be compensated. Attention must be paid to the effects on multiple essential traits of the target

QTL since transferring QTL even by well designed MAS is challenging and time consuming (Ye and Smith, 2010).

5. Conclusions

A panel of 327 *indica* rice varieties or advanced breeding lines was genotyped by the GBS method and phenotyped for GY and 10 related traits in eight environments. The whole panel was divided into two subpopulations. LD decay varied across chromosomes and the subpopulations. LD decay in the genome was around 200 kb in the whole population. GWAS based on a mixed linear model which incorporates PCA and kinship matrix identified a total of 43 QTL for eight out of 11 measured traits but PB, SB and SR. A significant interaction existed between QTL and environments with the majority of the QTL being only identified in a specific environment. Four QTL were located in the regions containing known genes/QTL for yield related traits. Eight multi-trait QTL regions were detected. The effects of identified QTL were relatively small with the largest percentage of variance explained by a single QTL being 9.6%. The present study demonstrated that MTA/QTL could be identified in a heterogeneous population of breeding lines through GWAS even if population structure and unequal familial relationship were present.

Conflict of interest

The authors declare they have no conflict of interest.

Author contribution

GY and DM Conceived and designed the experiments, SL, GR, and XZ Performed the experiments, SL, GY and LW Analyzed the data and wrote the manuscript and DM and MZ revised the manuscript.

Acknowledgments

SL was supported by a scholarship provided by the Monsanto's Beachell-Borlaug International Scholars Program. The research of GY was partially supported by the National Natural Science Foundation of China (grant 31461143014).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.fcr.2016.03.008>.

References

- Agrama, H.A., Eizenga, G.C., Yan, W., 2007. Association mapping of yield and its components in rice cultivars. *Mol. Breed.* 19, 341–356, <http://dx.doi.org/10.1007/s11032-006-9066-6>.
- Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., Jiang, R., Mulyati, N.W., Zhang, X., Amer, M.A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., Ecker, J.R., Faure, N., Kniskern, J.M., Jones, J.D.G., Michael, T., Nemri, A., Roux, F., Salt, D.E., Tang, C., Todesco, M., Traw, M.B., Weigel, D., Marjoram, P., Borevitz, J.O., Bergelson, J., Nordborg, M., 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465, 627–631, <http://dx.doi.org/10.1038/nature08800>.
- Begum, H., Spindel, J.E., Lalusin, A., Borromeo, T., Gregorio, G., Hernandez, J., Virk, P., Collard, B., McCouch, S.R., 2015. Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (*Oryza sativa*). *PLoS One* 10 (3), e0119873, <http://dx.doi.org/10.1371/journal.pone.0119873>.
- Bernardo, R., 2008. Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Sci.* 48, 1649–1664, <http://dx.doi.org/10.2135/cropsci2008.03.0131>.
- Bordes, J., Goudemand, E., Duchalais, L., Chevarin, L., Oury, F.X., Heumez, E., Lapiere, A., Perretant, M.R., Rolland, B., Beghin, D., Laurent, V., Le Gouis, J., Storlie, E., Robert, O., Charmet, G., 2014. Genome-wide association mapping of

- three important traits using bread wheat elite breeding populations. *Mol. Breed.* 33, 755–768, <http://dx.doi.org/10.1007/s11032-013-0004-0>.
- Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., Buckler, E.S., 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635, <http://dx.doi.org/10.1093/bioinformatics/btm308>.
- Cockram, J., White, J., Zuluaga, D.L., Smith, D., Comadran, J., Macaulay, M., Luo, Z., Kearsy, M.J., Werner, P., Harrop, D., Tapsell, C., Liu, H., Hedley, P.E., Stein, N., Schulte, D., Steuernagel, B., Marshall, D.F., Thomas, W.T.B., Ramsay, L., Mackay, I., Balding, D.J., Waugh, R., O'sullivan, D.M., 2010. Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. *Proc. Natl. Acad. Sci. U. S. A.* 107, 21611–21616, <http://dx.doi.org/10.1073/pnas.1010179107>.
- Cui, D., Xu, C.-Y., Tang, C.-F., Yang, C.-G., AX-X, Xu, F.-R., Zhang, J.-G., Han, L.-Z., 2013. Genetic structure and association mapping of cold tolerance in improved japonica rice germplasm at the booting stage. *Euphytica* 193, 369–382, <http://dx.doi.org/10.1007/s10681-013-0935-x>.
- Deshmukh, R., Singh, A., Jain, N., Anand, S., Gacche, R., Singh, A., Gaikwad, K., Sharma, T., Mohapatra, T., Singh, N., 2010. Identification of candidate genes for grain number in rice (*Oryza sativa* L.). *Funct. Integr. Genomics* 10, 339–347, <http://dx.doi.org/10.1007/s10142-010-0167-2>.
- Edae, E.A., Byrne, P.F., Haley, S.D., Lopes, M.S., Reynolds, M.P., 2014. Genome-wide association mapping of yield and yield components of spring wheat under contrasting moisture regimes. *Theor. Appl. Genet.* 127, 791–807, <http://dx.doi.org/10.1007/s00122-013-2257-8>.
- Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6 (5), e19379, <http://dx.doi.org/10.1371/journal.pone.0019379>.
- Evanno, G., Regnaut, S., Goudet, J., 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620, <http://dx.doi.org/10.1111/j.1365-294X.2005.02553.x>.
- Falush, D., Stephens, M., Pritchard, J.K., 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587, <http://dx.doi.org/10.1111/j.1471-8286.2007.01758.x>.
- Flint-Garcia, S.A., Thuitet, A.C., Yu, J., Pressoir, G., Romero, S.M., Mitchell, S.E., Doebley, J., Kresovich, S., Goodman, M.M., Buckler, E.S., 2005. Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* 44, 1054–1064, <http://dx.doi.org/10.1111/j.1365-3113.2005.02591.x>.
- GRiSP, 2010. Global Rice Science Partnership (GRiSP) Proposal.
- Guo, L., Ye, G., 2014. Use of major quantitative trait loci to improve grain yield of rice. *Rice Sci.* 21, 65–82, [http://dx.doi.org/10.1016/S1672-6308\(13\)60174-2](http://dx.doi.org/10.1016/S1672-6308(13)60174-2).
- Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z., Li, M., Fan, D., Guo, Y., Wang, A., Wang, L., Deng, L., Li, W., Lu, Y., Weng, Q., Liu, K., Huang, T., Zhou, T., Jing, Y., Li, W., Lin, Z., Buckler, E.S., Qian, Q., Zhang, Q.-F., Li, J., Han, B., 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–967, <http://dx.doi.org/10.1038/ng.695>.
- Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., Li, W., Guo, Y., Deng, L., Zhu, C., Fan, D., Lu, Y., Weng, Q., Liu, K., Zhou, T., Jing, Y., Si, L., Dong, G., Huang, T., Lu, T., Feng, Q., Qian, Q., Li, J., Han, B., 2012. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* 44, 32–39, <http://dx.doi.org/10.1038/ng.1018>.
- Juliano, B.O., 1993. Rice in Human Nutrition. Food and Agriculture Organization of the United Nations Rome (Italy) and International Rice Research Institute, Los Baños, Laguna (Philippines). 10.1017/CBO9781107415324.004.
- Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., Eskin, E., 2008. Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723, <http://dx.doi.org/10.1534/genetics.107.080101>.
- Kumar, V., Singh, A., Mithra, S.V.A., Krishnamurthy, S.L., Parida, S.K., Jain, S., Tiwari, K.K., Kumar, P., Rao, A.R., Sharma, S.K., Khurana, J.P., Singh, N.K., Mohapatra, T., 2015. Genome-wide association mapping of salinity tolerance in rice (*Oryza sativa*). *DNA Res.* 22, 133–145, <http://dx.doi.org/10.1093/dnares/dsu046>.
- Li, X., Yan, W., Agrama, H., Jia, L., Shen, X., Jackson, A., Moldenhauer, K., Yeater, K., McClung, A., Wu, D., 2011. Mapping QTL for improving grain yield using the USDA rice mini-core collection. *Planta* 234, 347–361, <http://dx.doi.org/10.1007/s00425-011-1405-0>.
- Li, X., Yan, W., Agrama, H., Jia, L., Jackson, A., Moldenhauer, K., Yeater, K., McClung, A., Wu, D., 2012. Unraveling the complex trait of harvest index with association mapping in rice (*Oryza sativa* L.). *PLoS One* 7 (1), e29350, <http://dx.doi.org/10.1371/journal.pone.0029350>.
- Liang, S., Ren, G., Liu, J., Zhao, X., Zhou, M., Mcneil, D., Ye, G., 2015. Genotype-by-environment interaction is important for grain yield in irrigated lowland rice. *Field Crops Res.* 180, 90–99, <http://dx.doi.org/10.1016/j.fcr.2015.05.014>.
- Liang, S., Sun, C., Ren, G., Zhao, X., Zhou, M., Mcneil, D., Ye, G., 2016. Usefulness of the cloned and fine-mapped genes/QTL for grain yield and related traits in *indica* rice breeding for irrigated ecosystems. *Field Crops Res.* 187, 58–73, <http://dx.doi.org/10.1016/j.fcr.2015.11.014>.
- Lin, H., Ashikari, M., Yamanouchi, U., Sasaki, T., Yano, M., 2002. Identification and characterization of a quantitative trait locus Hd9, controlling heading date in rice. *Breed. Sci.* 52, 35–41, <http://dx.doi.org/10.1270/jsbbs.52.35>.
- Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., Zhang, Z., 2012. GAPIT: Genome association and prediction integrated

- tool. *Bioinformatics* 28, 2397–2399, <http://dx.doi.org/10.1093/bioinformatics/bts444>.
- Mather, K.A., Caicedo, A.L., Polato, N.R., Olsen, K.M., McCouch, S., Purugganan, M.D., 2007. The extent of linkage disequilibrium in rice (*Oryza sativa* L.). *Genetics* 177, 2223–2232, <http://dx.doi.org/10.1534/genetics.107.079616>.
- McNally, K.L., Childs, K.L., Bohnert, R., Davidson, R.M., Zhao, K., Ulat, V.J., Zeller, G., Clark, R.M., Hoen, D.R., Bureau, T.E., Stokowski, R., Ballinger, D.G., Frazer, K.A., Cox, D.R., Padhukasahasram, B., Bustamante, C.D., Weigel, D., Mackill, D.J., Bruskiewich, R.M., Röttsch, G., Buell, C.R., Leung, H., Leach, J.E., 2009. Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc. Natl. Acad. Sci. U. S. A.* 106, 12273–12278, <http://dx.doi.org/10.1073/pnas.0900992106>.
- Mohammadi, M., Blake, T.K., Budde, A.D., Chao, S., Hayes, P.M., Horsley, R.D., Obert, D.E., Ullrich, S.E., Smith, K.P., 2015. A genome-wide association study of malting quality across eight U.S. barley breeding programs. *Theor. Appl. Genet.* 128, 705–721, <http://dx.doi.org/10.1007/s00122-015-2465-5>.
- Nawaz, Z., Kakar, K.U., Li, X., Li, S., Zhang, B., Shou, H., Shu, Q., 2015. Genome-wide association mapping of quantitative trait loci (QTL) for contents of eight elements in brown rice (*Oryza sativa* L.). *J. Agric. Food Chem.* 63, 8008–8016, <http://dx.doi.org/10.1021/acs.jafc.5b01191>.
- Nei, M., Li, W.H., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* 76, 5269–5273, <http://dx.doi.org/10.1073/pnas.76.10.5269>.
- Nguyen, N., Ferrero, A., 2006. Meeting the challenges of global rice production. *Paddy Water Environ.* 4, 1–9, <http://dx.doi.org/10.1007/s10333-005-0031-5>.
- Ordóñez Jr., S.A., Silva, J., Oard, J.H., 2010. Association mapping of grain quality and flowering time in elite japonica rice germplasm. *J. Cereal Sci.* 51, 337–343, <http://dx.doi.org/10.1016/j.jcs.2010.02.001>.
- Pasam, R.K., Sharma, R., Malosetti, M., van Eeuwijk, F.A., Haseneyer, G., Kilian, B., Graner, A., 2012. Genome-wide association studies for agronomical traits in a world wide spring barley collection. *BMC Plant Biol.* 12, 16, <http://dx.doi.org/10.1186/1471-2229-12-16>.
- Pauli, D., Muehlbauer, G., Smith, K., Cooper, B., 2014. Association mapping of agronomic QTL in US spring barley breeding germplasm. *Plant Genome* 7, 3, <http://dx.doi.org/10.3835/plantgenome2013.11.0037>.
- Phung, N.T.P., Mai, C.D., Mourmet, P., Frouin, J., Droc, G., Ta, N.K., Jouannic, S., Lê, L.T., Phung, V.N., Gantet, P., Courtois, B., 2014. Characterization of a panel of Vietnamese rice varieties using DArT and SNP markers for association mapping purposes. *BMC Plant Biol.* 14, 1–16, <http://dx.doi.org/10.1186/s12870-014-0371-7>.
- Pozniak, C.J., Clarke, J.M., Clarke, F.R., 2012. Potential for detection of marker-trait associations in durum wheat using unbalanced, historical phenotypic datasets. *Mol. Breed.* 30, 1537–1550, <http://dx.doi.org/10.1007/s11032-012-9737-4>.
- Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909, <http://dx.doi.org/10.1038/ng1847>.
- Pritchard, J.K., Stephens, M., Donnelly, P., 2000a. Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.
- Pritchard, J.K., Stephens, M., Rosenberg, N.A., Donnelly, P., 2000b. Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170–181, <http://dx.doi.org/10.1086/302959>.
- R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (URL <http://www.R-project.org/>).
- Romay, M.C., Millard, M.J., Glaubitz, J.C., Peiffer, J.A., Swarts, K.L., Casstevens, T.M., Elshire, R.J., Acharya, C.B., Mitchell, S.E., Flint-Garcia, S.A., McMullen, M.D., Holland, J.B., Buckler, E.S., Gardner, C.A., 2013. Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.* 14, R55, <http://dx.doi.org/10.1186/gb-2013-14-6-r55>.
- Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425 ([citeulike-article-id:93683](http://dx.doi.org/10.1093/molbev/4.4.406)).
- Shao, Y., Jin, L., Zhang, G., Lu, Y., Shen, Y., Bao, J., 2011. Association mapping of grain color, phenolic content, flavonoid content and antioxidant capacity in dehulled rice. *Theor. Appl. Genet.* 122, 1005–1016, <http://dx.doi.org/10.1007/s00122-010-1505-4>.
- Simmonds, N.W. (Ed.), 1979. *Longman, London*.
- Stich, B., Melchinger, A.E., 2009. Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and *Arabidopsis*. *BMC Genomics* 10, 94, <http://dx.doi.org/10.1186/1471-2164-10-94>.
- Stich, B., Möhring, J., Piepho, H.-P., Heckenberger, M., Buckler, E.S., Melchinger, A.E., 2008. Comparison of mixed-model approaches for association mapping. *Genetics* 178, 1745–1754, <http://dx.doi.org/10.1534/genetics.107.079707>.
- Storey, J.D., Tibshirani, R., 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* 100, 9440–9445.
- Storey, J.D., 2002. A direct approach to false discovery rates. *J. R. Stat. Soc. B* 64, 479–498.
- Wang, W.Y.S., Barratt, B.J., Clayton, D.G., Todd, J.A., 2005. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* 6, 109–118, <http://dx.doi.org/10.1038/nrg1522>.
- Wang, S., Wu, K., Yuan, Q., Liu, X., Liu, Z., Lin, X., Zeng, R., Zhu, H., Dong, G., Qian, Q., Zhang, G., Fu, X., 2012. Control of grain size, shape and quality by *OsSPL16* in rice. *Nat. Genet.* 44, 950–954, <http://dx.doi.org/10.1038/ng.2327>.
- Wang, K., Qiu, F., Dela Paz, M.A., Zhuang, J., Xie, F., 2014. Genetic diversity and structure of improved indica rice germplasm. *Plant Genet. Resour.* 12, 1–7, <http://dx.doi.org/10.1017/S1479262113000579>.
- Xie, F., Guo, L., Ren, G., Hu, P., Wang, F., Xu, J., Li, X., 2012. Genetic diversity and structure of indica rice varieties from two heterotic pools of southern China and IRRI. *Plant Genet. Resour.* 10 (1), <http://dx.doi.org/10.1017/s147926211200024X>.
- Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F., Li, X., Dong, Y., Gutenkunst, R.N., Fang, L., Huang, L., Li, J., He, W., Zhang, G., Zheng, X., Zhang, F., Li, Y., Yu, C., Kristiansen, K., Zhang, X., Wang, J., Wright, M., McCouch, S., Nielsen, R., Wang, J., Wang, W., 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* 30, 105–111, <http://dx.doi.org/10.1038/nbt.2050>.
- Ye, G., Smith, K.F., 2010. Marker-assisted gene pyramiding for cultivar development. In: Janick, J. (Ed.), *Plant Breeding Reviews*, vol. 33. John Wiley & Sons Inc., Hoboken, NJ, USA.
- Ye, G., Collard, B.C.Y., Zhao, X.Q., Nissila, E., 2013. Enhancing rice breeding efficiency: the role of breeding informatics. *Sabao J. Breed. Genet.* 45, 143–158.
- Yonemaru, J.-I., Mizobuchi, R., Kato, H., Yamamoto, T., Yamamoto, E., Matsubara, K., Hirabayashi, H., Takeuchi, Y., Tsunematsu, H., Ishii, T., Ohta, H., Maeda, H., Ebana, K., Yano, M., 2014. Genomic regions involved in yield potential detected by genome-wide association analysis in Japanese high-yielding rice cultivars. *BMC Genomics* 15, 346, <http://dx.doi.org/10.1186/1471-2164-15-346>.
- Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S., Buckler, E.S., 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208, <http://dx.doi.org/10.1038/ng1702>.
- Yu, S., Yang, C., Fan, Y., Zhuang, J., Li, X., 2008. Genetic dissection of a thousand-grain weight quantitative trait locus on rice chromosome 1. *Chinese Sci. Bull.* 53, 2326–2332, <http://dx.doi.org/10.1007/s11434-008-0281-x>.
- Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., Buckler, E.S., 2010. Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360, <http://dx.doi.org/10.1038/ng.546>.
- Zhang, D.L., Zhang, H.L., Qi, Y.W., Wang, M.X., Sun, J.L., Ding, L., Li, Z.C., 2013. Genetic structure and eco-geographical differentiation of cultivated Hsien rice (*Oryza sativa* L. subsp. *indica*) in China revealed by microsatellites. *Chinese Sci. Bull.* 58, 344–352, <http://dx.doi.org/10.1007/s11434-012-5396-4>.
- Zhao, K., Tung, C.-W., Eizenga, G.C., Wright, M.H., Ali, M.L., Price, A.H., Norton, G.J., Islam, M.R., Reynolds, A., Mezey, J., McClung, A.M., Bustamante, C.D., McCouch, S.R., 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* 2, 467, <http://dx.doi.org/10.1038/ncomms1467>.
- Zhu, C., Gore, M., Buckler, E.S., Yu, J., 2008. Status and prospects of association mapping in plants. *Plant Genome* 1, 5–20, <http://dx.doi.org/10.3835/plantgenome2008.02.0089>.