

# Development and application of a set of breeder-friendly SNP markers for genetic analyses and molecular breeding of rice (*Oryza sativa* L.)

Haodong Chen · Hang He · Yanjiao Zou · Wei Chen · Renbo Yu ·  
Xia Liu · Yang Yang · Yong-Ming Gao · Jian-Long Xu ·  
Liu-Min Fan · Yi Li · Zhi-Kang Li · Xing Wang Deng

Received: 7 January 2011 / Accepted: 1 June 2011  
© Springer-Verlag 2011

**Abstract** Single nucleotide polymorphisms (SNPs) are the most abundant DNA markers in plant genomes. In this study, based on 54,465 SNPs between the genomes of two *Indica* varieties, Minghui 63 (MH63) and Zhenshan 97 (ZS97) and additional 20,705 SNPs between the MH63 and Nipponbare genomes, we identified and confirmed 1,633 well-distributed SNPs by PCR and Sanger sequencing. From these, a set of 372 SNPs were further selected to analyze the patterns of genetic diversity in 300 representative rice inbred lines from 22 rice growing countries worldwide. Using this set of SNPs, we were able to uncover the well-known *Indica*–*Japonica* subspecific differentiation and geographic differentiations within *Indica* and *Japonica*. Furthermore, our SNP results revealed some common and contrasting patterns of the haplotype diversity along different rice chromosomes in the *Indica* and

*Japonica* accessions, which suggest different evolutionary forces possibly acting in specific regions of the rice genome during domestication and evolution of rice. Our results demonstrated that this set of SNPs can be used as anchor SNPs for large scale genotyping in rice molecular breeding research involving *Indica*–*Japonica* and *Indica*–*Indica* crosses.

## Introduction

Rice (*Oryza sativa*) is one of the most important food crops around the world. With a broad geographic adaptation and rich phenotypic and molecular diversity, rice is an excellent model system for studying genetics and evolution of crop plants. Significant progress in rice functional genomics is being made since the completion of the international rice genome sequencing project (International Rice Genome Sequencing Project 2005), which offers tremendous opportunities for breeders to improve this important crop by molecular breeding. However, this expectation has been hindered by lack of low cost and robust marker technology that are breeder-friendly because a successful molecular breeding program requires detailed and comprehensive understanding of the genetic basis of target traits, fast and efficient characterization of the relationships among potential parental lines, and accurate design of crosses and trait selection schemes using molecular markers. This problem can now be readily solved by developing single nucleotide polymorphisms (SNPs) markers, the most abundant ones in the rice genomes. Although the availability of the Nipponbare (*temperate japonica*, *Japonica*) and 93-11 (*Indica*) sequences (Yu et al. 2002; International Rice Genome Sequencing Project 2005) has provided us with the most useful

---

H. Chen, H. He and Y. Zou contributed equally to this work.

---

Communicated by M. Wissuwa.

---

**Electronic supplementary material** The online version of this article (doi:10.1007/s00122-011-1633-5) contains supplementary material, which is available to authorized users.

---

H. Chen · H. He · Y. Zou · W. Chen · R. Yu · X. Liu ·  
Y. Yang · L.-M. Fan · Y. Li · X. W. Deng (✉)  
Peking-Yale Joint Center for Plant Molecular Genetics  
and Agro-biotechnology, State Key Laboratory of Protein  
and Plant Gene Research, College of Life Sciences,  
Peking University, 100871 Beijing, China  
e-mail: deng@pku.edu.cn

Y.-M. Gao · J.-L. Xu · Z.-K. Li (✉)  
Institute of Crop Sciences/National Key Facility for Crop Gene  
Resources and Genetic Improvement, Chinese Academy  
of Agricultural Sciences, 100081 Beijing, China  
e-mail: z.li@cgiar.org; zhkli@yahoo.com.cn

database to find large numbers of SNPs (Feltus et al. 2004; Shen et al. 2004), the discovery of polymorphisms within lines of the same cultivar type (*Indica* or *Japonica*) is more difficult (Monna et al. 2006). Recently, high-throughput sequence technologies provide the capability to produce more than 100 Gb sequences per run, and genome-wide SNPs selection has become easy to achieve (Chi 2008; Hillier et al. 2008; Ossowski et al. 2008; Schuster 2008; Yamamoto et al. 2010). Low, medium and high-resolution SNPs assays have been developed and used in dissecting phenotype–genotype associations in rice (Tung et al. 2010; McCouch et al. 2010).

SNPs have been increasingly used in linkage and association mapping studies of crop species. In maize, genetic properties of a maize nested association mapping population, generated by crossing 25 diverse inbred maize lines to the reference B73 line, have been analyzed and used for association mapping of a large number of complex traits (Buckler et al. 2009; McMullen et al. 2009). Barley geneticists also used a highly parallel SNP assay platform to identify and directly fine map traits in elite plant breeding materials (Waugh et al. 2009). The effects of SNP number and selection strategy on estimates of diversity and population structure for different types of barley germplasm have been evaluated (Moragues et al. 2010). Illumina GoldenGate assay has also been demonstrated suitable for SNP genotyping of homozygous tetraploid and hexaploid wheat lines (Akhunov et al. 2009). These high-throughput SNP genotyping platforms improved the efficiency of diversity and mapping analyses dramatically for different crops.

Until recently, several studies have reported the uses of SNP markers in characterizing the breeding history, determining genomic composition of genetically related varieties, and detecting associations between introgressed genomic regions and agronomic traits in rice. Resequencing microarrays have been used to interrogate 100 Mb of the unique fraction of the reference genome (Nipponbare) for 20 diverse varieties and landraces that capture impressive genotypic and phenotypic diversity of the domesticated rice, providing comprehensive SNP data for genotyping other varieties (McNally et al. 2009). A set of 1,536 SNP targets were selected from the data generated in the OryzaSNP project, and used to characterize the genome-wide patterns of polymorphisms in 395 diverse rice accessions (Zhao et al. 2010). Recently, a comparison between the resequencing data of an elite Japanese rice cultivar, Koshihikari and the reference Nipponbare sequence data resulted in the development of a high-throughput SNP typing array suitable for genotyping of *Japonica* cultivars (Yamamoto et al. 2010). A method for constructing high-density linkage maps composed of high-quality SNPs based on low-coverage sequences of

recombinant inbred lines has been reported (Xie et al. 2010). Recently, Bin Han and his colleagues have identified about 3.6 million SNPs by resequencing 517 rice landraces and a high-density rice genome haplotype map was constructed (Huang et al. 2010).

In this study, we developed and tested a small but useful set of 384 genome-wide SNPs identified primarily from the sequence data of Minghui 63 (MH63) and Zhenshan 97 (ZS97), the parental lines of an elite hybrid Shanyou 63, which has been the dominant hybrid cultivar for 15 years in China. Further, we applied a microbead-based SNP detection system to genotype a global set of 300 rice varieties which were used as the parental lines for the International Rice Molecular Breeding Network (IRMBN, Yu et al. 2003) and for developing “Green Super Rice” (Zhang 2007). Comparing with other types of molecular markers including isozymes (Glaszmann 1987; Li and Rutger 2000), RFLPs (Olufowote et al. 1997; Lu et al. 2002), and SSRs (Ni et al. 2002; Garris et al. 2005), SNPs showed obvious advantages. Our results demonstrated the power and usefulness of this set of SNPs in differentiating genomic differences between *Indica* rice cultivars in our rice molecular breeding program.

## Materials and methods

### Plant materials and DNA extraction

Three hundred rice varieties representing major geographic areas of rice growing countries in the world were chosen for SNP genotyping and characterization. These varieties are the parental lines of the IRMBN for developing “Green Super Rice Cultivars” for the Resource-Poor of Africa and Asia supported by Bill and Melinda Gates Foundation (Tables 1, S1). Most materials used in this study are publicly available in the Rice Genebank of the International Rice Research Institute, and the remaining ones may be obtained by sending the requests to Dr. ZK Li (one of the corresponding authors) and signing specific Material Transfer Agreements (MTAs). Rice leaves were collected from approximate 10 seedlings for each line. Plant DNA extraction kits (Qiagen) were used for the genomic DNA extraction.

### SNPs selection and confirmation

We collected the SNPs along the rice genome from two resources: OryzaSNP project (McNally et al. 2009), and a population sequence project (Xie et al. 2010). Because 78% of rice accessions in the collection of this study are *Indica* germplasm, most of our SNPs were selected between MH63 and ZS97. In cases where there was no

**Table 1** Summary information of the geographic origins of the 300 sampled rice inbred lines used for SNP analyses

|                 | No. | Asia |           |       |      | Africa | Europe | America |       |
|-----------------|-----|------|-----------|-------|------|--------|--------|---------|-------|
|                 |     | East | Southeast | South | West |        |        | North   | South |
| <i>Indica</i>   | 235 | 103  | 96        | 34    | 2    | 0      | 0      | 0       | 0     |
| <i>Japonica</i> | 65  | 41   | 12        | 3     | 0    | 2      | 2      | 2       | 3     |

SNP between MH63 and ZS97 in a large genome region, SNPs between MH63 and Nipponbare were used as supplements. The following are the criteria for SNP selection: (1) extract 201 bp sequence of each SNP site with its upstream and downstream 100 bp length genomic DNA; (2) compare the extracted sequences to the reference rice genome (Nipponbare), and select SNPs only from unique genomic regions. Together, 54,465 SNPs have been extracted between MH63 and ZS97, and an additional 20,705 SNPs between MH63 and Nipponbare.

The next step is to select and confirm the SNP sites distributed evenly along the whole genome. The SNPs were omitted if there's any other SNP or Indel around the target site. In this step, 1,633 SNPs have been confirmed with about one marker each centimorgan. For confirming these SNPs, we designed PCR primers up- and downstream of each SNP site. The PCR products of MH63, ZS97 and Nipponbare were further sequenced to validate these SNPs. All the confirmed SNPs were submitted to Illumina for designability rank score. The scores are ranging from 1.0 to 0, where a rank score of >0.6 indicates a high-success rate, 0.4 to <0.6 indicates a moderate success rate, and <0.4 indicates a low success rate for the GoldenGate assay. A total of 384 SNPs distributed evenly across the genome (one marker per 4 cM) were further selected for Oligo Pool Assay (OPA) synthesis, 381 of which show designability rank scores of 0.6 or higher, and the remaining 3 SNPs with scores between 0.5 and 0.6 were chosen due to the fact that there are no better ones in those target regions. Finally, 357 out of the 384 SNPs are between MH63 and ZS97, and the remaining 27 SNPs are between MH63 and Nipponbare.

### SNP genotyping

SNP genotyping was carried out using Illumina BeadXpress. For each sample, 250 ng of DNA were used. Several inbred rice accessions with known whole genome sequences (MH63, ZS97 and Nipponbare) and their crossed F1 samples were added into the experiment as controls. The system using GoldenGate assay and vericode technology can genotype 96 samples for 384 markers in a single plate. All SNP genotyping data (300 accessions by 384 SNPs) generated from BeadXpress system were scored using the Illumina GenomeStudio genotyping software with a no-call threshold of 0.25, which is a recommended

bound for obtaining a reliable genotype call. The SNP calling results were further adjusted based on the genotyping clusters of the control samples. The SNPs with call rates lower than 90% were removed. Finally, 372 makers for 300 accessions were retained for data analysis.

### Data analysis

The polymorphism information content (PIC) value (Botstein et al. 1980) was calculated as the measurement of gene diversity at each SNP marker, using the following formula:

$$PIC_i = 1 - \sum_{j=1}^n P_{ij}^2 - \sum_{j=1}^{n-1} \sum_{k=j+1}^n 2P_{ij}^2 P_{ik}^2$$

In this formula,  $P_{ij}$  and  $P_{ik}$  are the frequencies of  $j$ th and  $k$ th alleles for marker  $i$ , respectively. The heterozygosity value indicates the proportion of heterozygous loci. The gene diversity for marker  $i$  can be obtained using the above equation by dropping the last item, which indicates the probability that two alleles of randomly chosen test samples are different (Lu et al. 2009).

Allele frequency was calculated for characterizing the subspecific differentiation and geographic patterns of genetic diversity in the sampled rice accessions through three pairwise comparisons: *Japonica* versus *Indica*, East Asia (EA) *Indica* versus South and Southeast Asia (SA/SEA) *Indica*, and EA *Japonica* versus SA/SEA *Japonica*. To apply hierarchical clustering alignment, software Cluster (Eisen et al. 1998) was performed on the chosen 372 SNP markers, and the average linkage clustering method was adopted. Then, software Treeview was applied on the clustered genotypes to obtain the graphic view. Further cluster analysis using 258 markers with minor allele frequencies (MAF) >0.2 was carried out using MEGA 4 (Tamura et al. 2007). The Nei's genetic distance (Tamura and Nei 1993) and the neighbor-joining tree (NJ) method (Saitou and Nei 1987) were performed and interior branch test was applied.

Allele frequencies within different germplasm collections were used to identify SNP markers with unique or missing alleles in specific germplasm collections. SNP markers detecting significantly different allele frequencies have the potential to be used to distinguish rice lines from

different germplasm groups. Unique alleles indicate the alleles existing only in one germplasm collection but not in the others. Missing alleles indicate the alleles completely lacking in one germplasm collection while existing in all others (Lu et al. 2009).

## Results

Properties of the SNPs, classification and genetic diversity of the sampled rice accessions

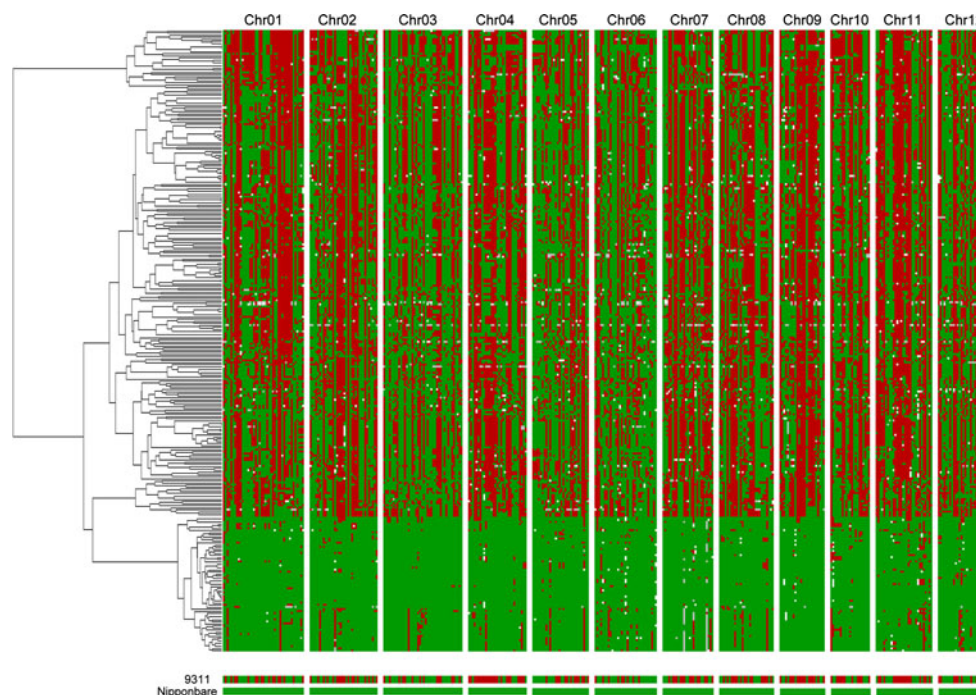
In this study, 1,633 SNPs distributed evenly across the rice genome or about one SNP each centimorgan, were originally selected and confirmed (Fig. S1). Of these, 675 (41.3%) SNPs locate in the intergenic regions. The majority SNPs (958) locate in the intragenic regions, including 373 in exon regions, 51 of which are nonsynonymous mutations, resulting in changes of the amino acids and thus possible associations with specific functions (Table S2).

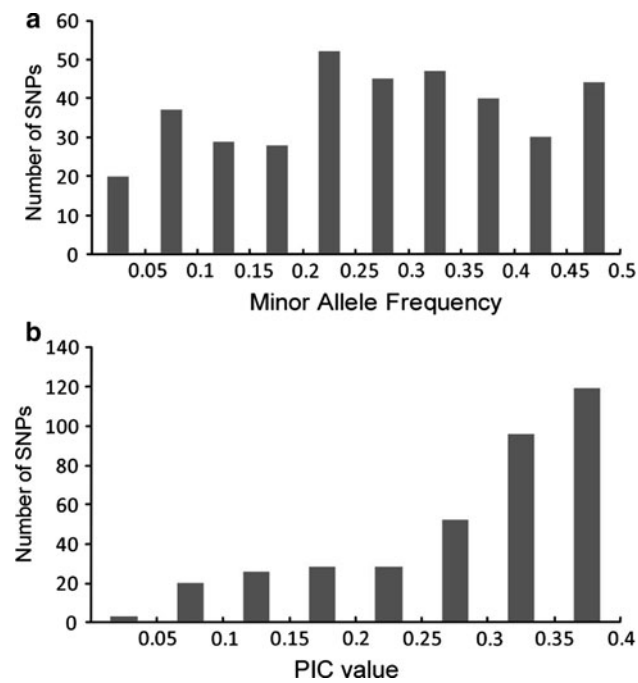
Then, a selected subset of 384 SNPs were used to genotype the 300 accessions using the BeadXpress system and the Illumina GenomeStudio genotyping software. A clear output is shown in Fig. S2, where three separated clusters, representing AA (pink), AB (purple), and BB (blue) genotypes were revealed by a single SNP. In the full genotypic dataset, 12 SNPs had call rates lower than 90% and thus removed from further analyses. The remaining

372 high-quality SNPs were confirmed with an estimated genotyping error rate lower than 1% based on the analysis of control sample Nipponbare. Base changes at the 372 SNPs include A/G (142) and C/T (111) accounting for 68% of the informative SNPs, and A/C (30), A/T (42), C/G (17) and G/T (30) accounting for the remaining 32%. Of the 372 SNPs, 234 (62.9%) SNPs locate in the intragenic regions, including 92 in exon regions, 10 of which are nonsynonymous mutations, resulting in changes of the amino acids and thus possible associations with specific functions (Fig. S3; Table S3). The remaining 138 (37.1%) SNPs locate in the intergenic regions (Table S3). A list of these SNPs and their diversity information, including base change, minor allele frequency (MAF), heterozygosity, gene diversity, and PIC estimates, are provided in Table S4.

Figure 1 shows the clustering of the 300 accessions based on the 372 SNPs. Of the 372 markers, only 20 (5.4%) showed  $MAF \leq 0.05$ , and 258 (69.4%) had  $MAF \geq 0.2$ , which were considered to have good differentiating power in distinguishing the tested rice accessions (Fig. 2a). There are 44 (11.8%) SNPs with two alternative alleles of approximately equal allele frequencies ( $MAF > 0.45$ ), which are excellent in differentiating the tested rice accessions. As expected, 744 alleles were detected at the 372 marker loci each with two alleles in the 300 accessions. The average PIC was 0.285, ranging from 0.010 to 0.375 (Table S4) with a peak distribution between 0.350 and 0.375 (Fig. 2b). The estimated average gene diversity

**Fig. 1** Clustering of 300 rice inbred lines based on 372 single nucleotide polymorphisms (SNPs). Homozygous alleles identical to Nipponbare were labeled as *green* color, different from Nipponbare as *red* color, and heterozygous alleles as *white*. Grey indicates no clear genotyping signal from the experiment





**Fig. 2** Frequency distribution of the minor alleles and their polymorphic information content (*PIC*) of 372 informative SNPs among the 300 rice inbred lines

of the SNPs was 0.358, ranging from 0.010 to 0.500. The tested accessions exhibited an average of heterozygosity of 0.7% at these SNPs.

To avoid the possible limitation of the 372 SNPs caused by fact that they were selected based on the sequence data of MH63, ZS97 and Nipponbare, a subset of 258 SNPs were further selected for genetic diversity analysis based on two factors: normal MAF ( $>0.2$ ), and high PIC values ( $>0.25$ ) (Tables S4, S5). Clearly, this subset of selected SNPs gave virtually the same picture regarding the

geographic patterns of genetic diversity in the tested 300 rice accessions as that obtained by the whole data set of the 372 SNPs (Table 2). Within the 300 sampled inbred lines, EA *Indica* accessions showed the highest average PIC value of 0.277 and gene diversity of 0.346. The average PIC and gene diversity estimates were also high for SA/SEA *Indica*, but very low for *Japonica* (Table 2).

#### Classification of the 300 sampled rice inbred accessions

Figures 3 and S4 show a neighbor-joining tree of the 300 tested accessions constructed using the subset of 258 markers with  $PIC > 0.20$ . The 300 tested accessions were grouped into two major groups, representing the two subspecies of *Oryza sativa*, including 235 *Indica* accessions, and 65 *Japonica* accessions. *Indica* accessions were well separated into five subgroups, covering geographic differentiations. All *Japonica* accessions were grouped into a single cluster with two subclusters, representing roughly the temperate and tropical *japonicas*. Generally, *Indica* accessions were separated better than *Japonica*, which was consistent with the expectation based on our SNP selection methods.

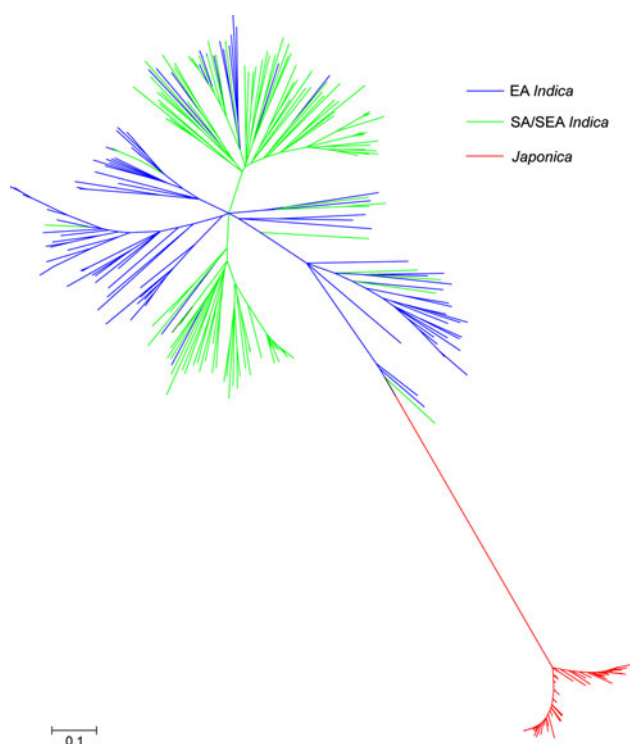
#### Genomic differentiation among rice varietal groups

To get more detailed information of the subspecific and geographic differentiation among the tested rice accessions, we performed three pairwise comparisons in allele frequencies of all SNPs: *Japonica* versus *Indica*, EA *Indica* versus SA/SEA *Indica*, and EA *Japonica* versus SA/SEA *Japonica* (Fig. 4). In the *Japonica* versus *Indica* comparison (Fig. 4a), 318 (85.5%) SNPs showed allele frequency difference larger than 10%, including 51 SNPs at which the allele frequency differences were greater than 80%. In the

**Table 2** Genetic diversity estimates within the sampled 300 rice inbred lines as revealed by selected subsets of SNP markers

|                        | <i>N</i> | MAF   | Gene diversity | Heterozygosity | PIC   |
|------------------------|----------|-------|----------------|----------------|-------|
| 372 markers            |          |       |                |                |       |
| <i>Indica</i>          | 235      | 0.233 | 0.319          | 0.008          | 0.257 |
| <i>Japonica</i>        | 65       | 0.053 | 0.080          | 0.004          | 0.068 |
| EA <i>Indica</i>       | 103      | 0.255 | 0.346          | 0.008          | 0.277 |
| SA/SEA <i>Indica</i>   | 130      | 0.206 | 0.280          | 0.008          | 0.226 |
| EA <i>Japonica</i>     | 41       | 0.042 | 0.068          | 0.003          | 0.060 |
| SA/SEA <i>Japonica</i> | 15       | 0.047 | 0.068          | 0.005          | 0.056 |
| 258 markers            |          |       |                |                |       |
| <i>Indica</i>          | 235      | 0.282 | 0.368          | 0.009          | 0.291 |
| <i>Japonica</i>        | 65       | 0.061 | 0.094          | 0.003          | 0.081 |
| EA <i>Indica</i>       | 103      | 0.290 | 0.376          | 0.009          | 0.297 |
| SA/SEA <i>Indica</i>   | 130      | 0.262 | 0.343          | 0.009          | 0.272 |
| EA <i>Japonica</i>     | 41       | 0.050 | 0.082          | 0.003          | 0.072 |
| SA/SEA <i>Japonica</i> | 15       | 0.058 | 0.082          | 0.003          | 0.068 |

EA East Asia, SA South Asia, SEA Southeast Asia, *N* the sample size, MAF minor allele frequency, PIC polymorphism information content



**Fig. 3** The neighbor-joining (NJ) tree for the 300 rice inbred lines based on Nei's genetic distance. Varieties labeled as *blue* indicate *Indica* accessions from East Asia, *green* for *Indica* accessions from South/Southeast Asia, and *red* for *Japonica* accessions. A different presentation of this tree with the names of the individual inbred lines was shown in Fig. S4

comparison between the EA *Indica* and SA/SEA *Indica* (Fig. 4b), 169 (45.4%) SNPs showed significant allelic differences >10% with the largest difference of 44.1%, including 17 SNPs with allele frequency difference larger than 30%. In the comparison between the EA *Japonica* and SA/SEA *Japonica* (Fig. 4c), 60 (16.1%) SNPs showed significant allelic differences >10%, and only 26 SNPs exhibited allele frequency differences greater than 30%. These results indicate that significant geographic differentiation has occurred within the *Indica* subspecies but to a lesser degree within the *Japonica* subspecies of *O. sativa* at many SNP loci. This result is consistent with previous results from isozymes (Glaszmann 1987; Li and Rutger 2000), RFLPs (Wang et al. 1992), SSRs (Ni et al. 2002; Yu et al. 2003) and SNPs (McNally et al. 2009).

To better evaluate the usefulness of the selected set of 372 SNPs in rice genetic analyses and molecular breeding, we calculated the number of SNPs between pairwise accessions (Table 3). When averaged from 15,275 pairwise comparisons between the 235 *Indica* accessions and the 65 *Japonica* varieties, the number of differentiating SNPs was 174.2 (46.8%), ranging from 131 (35.2%) to 232 (62.4%). When averaged from 27,495 pairwise comparisons within the 235 *Indica* accessions, the number of differentiating

SNPs was 120 (32.4%), ranging from 0 to 315 (84.7%). When averaged from 2,080 pairwise comparisons within the 65 *Japonica* accessions, the number of differentiating SNPs was 30.7 (8.3%), ranging from 0 to 64 (17.2%). This result indicates that this selected set of SNPs is more suitable for polymorphism analysis within *Indica* accessions and between *Indica* and *Japonica* accessions than within *Japonica* accessions. For distinguishing different germplasm collections, ten SNP markers representing the most significant allelic difference were selected from each of three pairwise germplasm comparisons (Table 4). The largest average allele frequency difference for the top ten differences was between *Indica* and *Japonica* (0.968), followed by *Japonica* in EA versus SA/SEA (0.611).

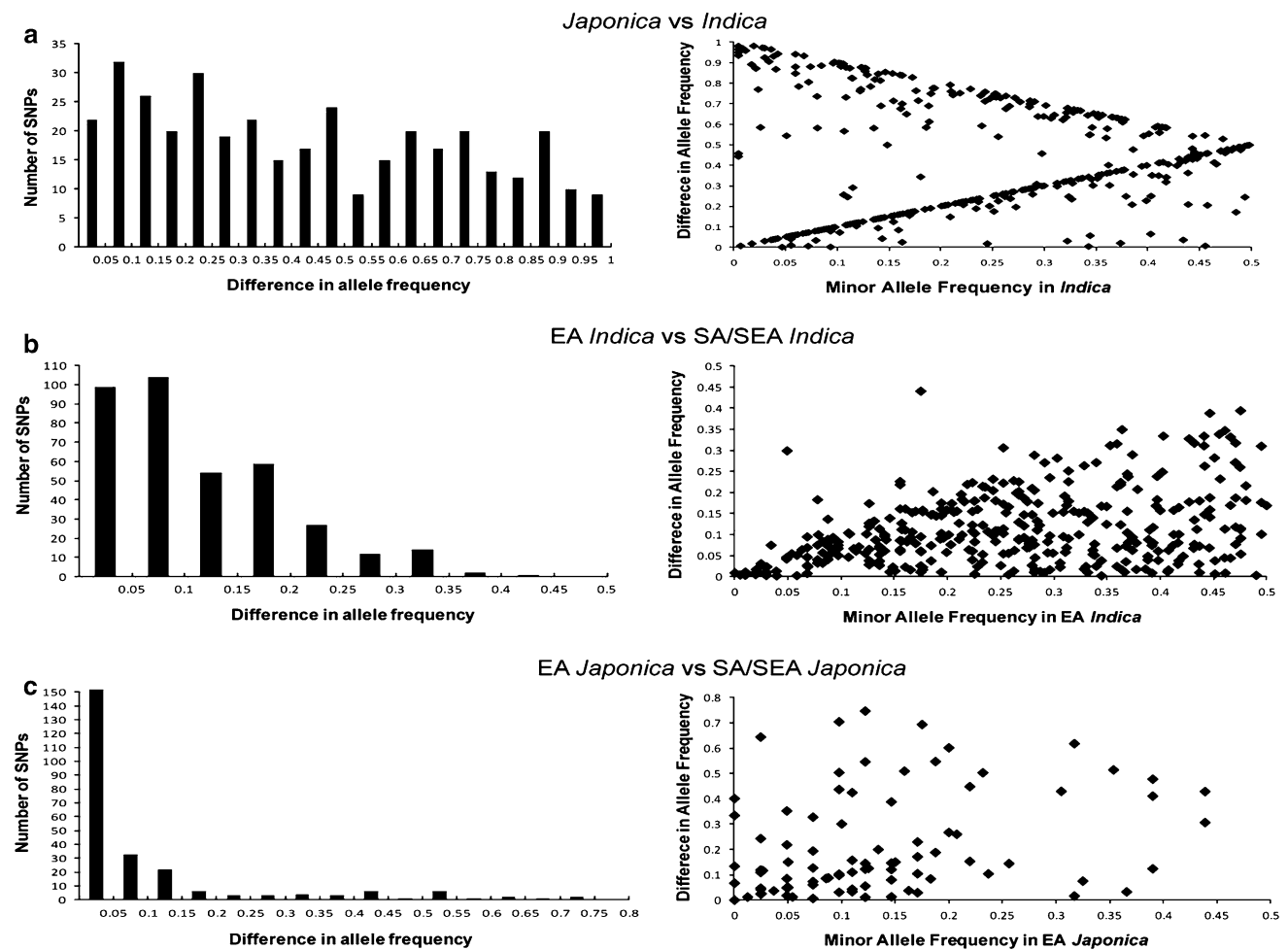
#### Unique alleles in different varietal groups

By comparing allele frequencies of a specific germplasm group with those in the entire sample (300 lines), nine unique alleles were identified that only existed in EA *Indica* accessions (Table 5). Interestingly, unique allele G at OS01-40348552-MZ showed 21.8% frequency in EA *Indica* accessions, which is relatively high. There are many markers showing missing alleles, especially in *Japonica*. Unique or missing alleles identified above, together with markers showing significant allele frequency difference among germplasm groups, can be combined to characterize rice varieties.

#### Genome haplotype diversity

Because chromosomal segments, instead of individual SNPs, are the units of inheritance, estimating the number of haplotypes per unit of genomic region in the tested accessions provides a better picture of their evolutionary history. Based on clear changes in the number of haplotypes, we were able to determine that a 10-SNP window of approximately 9.0 Mb in physical length is appropriate for estimating the haplotype diversity across the genome, which resulted in identification of a total of 19,961 haplotypes in the sampled rice accessions using the sliding-window approach.

Figure 5 shows the average haplotype diversity estimates of the 103 EA *Indica* accessions, 160 SA/SEA *Indica* accessions and 65 *Japonica* accessions along the 12 rice chromosomes. Again, the former two groups had the highest average number of haplotypes at 0.3/line, three times as much as that in the *Japonica* accessions (0.1/line). It is interesting to note that there are 15 genomic regions in the middle parts of chromosomes 3 and 4, on the top parts of chromosomes 2, 5, 7 and 11, and on the bottom parts of chromosomes 2–9, where haplotype diversity was at the highest in the *Indica* accessions, but lowest in the *Japonica*



**Fig. 4** Differentiation of allele frequencies between rice germplasm collections: **a** *Japonica* versus *Indica*, **b** *EA Indica* versus *SA/SEA Indica*, and **c** *EA Japonica* versus *SA/SEA Japonica*

**Table 3** The average number of SNPs of the selected set of 372 SNPs differentiating pairwise rice accessions

|  | $N^a$  | Mean  | Range   | Polymorphism level (%) <sup>b</sup> |
|--|--------|-------|---------|-------------------------------------|
| <i>Indica</i> versus <i>Japonica</i>             | 15,275 | 174.2 | 131–232 | 46.8                                |
| Within <i>Indica</i>                             | 27,495 | 120.0 | 0–315   | 32.4                                |
| <i>EA Indica</i> versus <i>SA/SEA Indica</i>     | 13,390 | 124.7 | 0–240   | 33.5                                |
| Within <i>EA Indica</i>                          | 5,253  | 131.2 | 0–315   | 35.3                                |
| Within <i>SA/SEA Indica</i>                      | 8,385  | 105.8 | 0–180   | 28.4                                |
| Within sister lines of IR81896                   | 55     | 29.6  | 1–46    | 8.0                                 |
| Within sister lines of IR81047                   | 10     | 36.0  | 16–62   | 9.7                                 |
| Within <i>Japonica</i>                           | 2,080  | 30.7  | 0–64    | 8.3                                 |
| <i>EA Japonica</i> versus <i>SA/SEA Japonica</i> | 615    | 33.6  | 0–64    | 9.0                                 |
| Within <i>EA Japonica</i>                        | 820    | 26.5  | 0–55    | 7.1                                 |
| Within <i>SA/SEA Japonica</i>                    | 105    | 27.1  | 0–47    | 7.3                                 |

<sup>a</sup>  $N$  is the total number of pairwise comparisons

<sup>b</sup> The approximate genome coverage estimated by the polymorphic SNPs between a pair of tested accessions

accessions. This result suggests the balancing selection has been operating on these regions in the former group, but the purifying selection has been acting in the latter. The opposite was true for 3 specific regions in the middle of

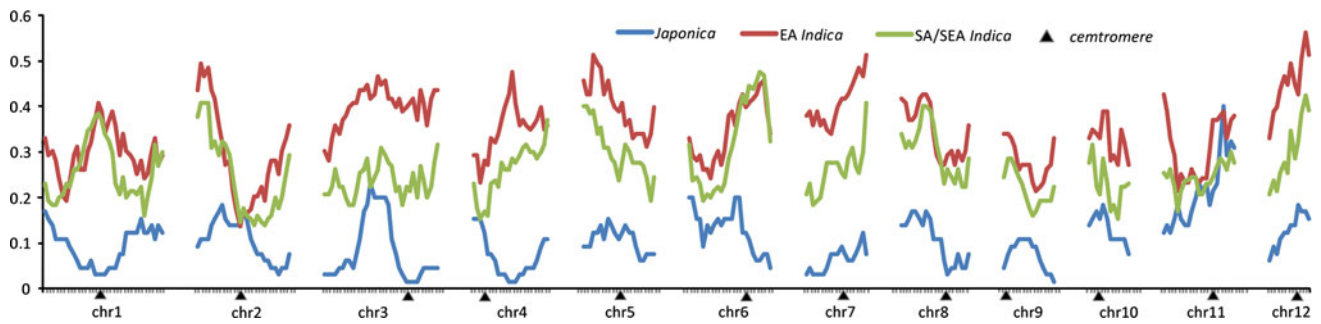
chromosomes 2 and 9 and the top of chromosome 4. In both the *Indica* and *Japonica*, the balancing selection appeared to be responsible for the observed high haplotype diversity genomic regions in the middle of chromosome 3,

**Table 4** Top ten single nucleotide polymorphisms (SNPs) with the greatest subspecific and geographic differentiation in allele frequency detected in the 300 sampled rice inbred lines

| SNP name  | Chr. | Physical position | Allele | Allele frequency |       | Allele frequency difference |
|---|------|-------------------|--------|------------------|-------|-----------------------------|
|   |      |                   |        | I                | II    |                             |
| <i>Japonica</i> (I) versus <i>Indica</i> (II)             |      |                   |        |                  |       |                             |
| OS01-27636997-MN  | 1    | 27636997          | T      | 0.981            | 0.000 | 0.981                       |
| OS02-22033189-MN  | 2    | 22033189          | T      | 0.996            | 0.015 | 0.980                       |
| OS02-27006317-MN  | 2    | 27006317          | T      | 0.973            | 0.000 | 0.973                       |
| OS01-31331538-MZ  | 1    | 31331538          | T      | 0.970            | 0.000 | 0.970                       |
| OS09-14379283-MN  | 9    | 14379283          | T      | 0.970            | 0.000 | 0.970                       |
| OS09-15128006-MN  | 9    | 15128006          | C      | 0.992            | 0.023 | 0.968                       |
| OS11-09392231-MN  | 11   | 9392231           | T      | 0.996            | 0.031 | 0.965                       |
| OS07-04493215-MN  | 7    | 4493215           | A      | 0.964            | 0.000 | 0.964                       |
| OS04-09625204-MN  | 4    | 9625204           | A      | 0.989            | 0.031 | 0.958                       |
| OS08-02202344-MN  | 8    | 2202344           | C      | 0.996            | 0.046 | 0.950                       |
| EA <i>Indica</i> (I) versus SA/SEA <i>Indica</i> (II)     |      |                   |        |                  |       |                             |
| OS01-22453166-MN  | 1    | 22453166          | A      | 0.825            | 0.385 | 0.441                       |
| OS07-23235458-MZ  | 7    | 23235458          | A      | 0.525            | 0.130 | 0.394                       |
| OS07-24653396-MZ  | 7    | 24653396          | C      | 0.553            | 0.165 | 0.388                       |
| OS08-24754479-MZ  | 8    | 24754479          | A      | 0.636            | 0.287 | 0.349                       |
| OS04-18754459-MZ  | 4    | 18754459          | T      | 0.539            | 0.192 | 0.347                       |
| OS01-08924578-MZ  | 1    | 8924578           | A      | 0.545            | 0.207 | 0.338                       |
| OS08-03003597-MZ  | 8    | 3003597           | T      | 0.597            | 0.263 | 0.334                       |
| OS11-01915704-MZ  | 11   | 1915704           | T      | 0.559            | 0.225 | 0.333                       |
| OS01-20824464-MZ  | 1    | 20824464          | G      | 0.534            | 0.203 | 0.331                       |
| OS10-19761586-MZ  | 10   | 19761586          | G      | 0.573            | 0.246 | 0.327                       |
| EA <i>Japonica</i> (I) versus SA/SEA <i>Japonica</i> (II) |      |                   |        |                  |       |                             |
| OS03-11176746-MZ  | 3    | 11176746          | C      | 0.878            | 0.133 | 0.745                       |
| OS02-04605390-MZ  | 2    | 4605390           | G      | 0.902            | 0.200 | 0.702                       |
| OS01-32202075-MZ  | 1    | 32202075          | C      | 0.825            | 0.133 | 0.692                       |
| OS04-33626126-MZ  | 4    | 33626126          | A      | 0.976            | 0.333 | 0.642                       |
| OS10-00391885-MZ  | 10   | 391885            | G      | 0.683            | 0.067 | 0.616                       |
| OS02-18307289-MZ  | 2    | 18307289          | T      | 0.800            | 0.200 | 0.600                       |
| OS07-20328300-MN  | 7    | 20328300          | C      | 0.813            | 0.267 | 0.546                       |
| OS07-28301346-MZ  | 7    | 28301346          | G      | 0.878            | 0.333 | 0.545                       |
| OS05-27647053-MN  | 5    | 27647053          | A      | 0.646            | 0.133 | 0.513                       |
| OS08-10157017-MZ  | 8    | 10157017          | C      | 0.841            | 0.333 | 0.508                       |

**Table 5** Nine unique SNP alleles in the 103 EA *Indica* accessions

| SNP name         | Chr. | Position | Unique allele | Allele frequency |
|------------------|------|----------|---------------|------------------|
| OS01-40348552-MZ | 1    | 40348552 | G             | 0.218            |
| OS02-23821861-MZ | 2    | 23821861 | T             | 0.092            |
| OS03-01094836-MZ | 3    | 1094836  | C             | 0.068            |
| OS03-02660732-MZ | 3    | 2660732  | A             | 0.087            |
| OS03-11798566-MZ | 3    | 11798566 | A             | 0.131            |
| OS03-31435766-MZ | 3    | 31435766 | A             | 0.160            |
| OS06-09397026-MZ | 6    | 9397026  | A             | 0.130            |
| OS09-21921271-MZ | 9    | 21921271 | G             | 0.126            |
| OS11-26048286-MZ | 11   | 26048286 | T             | 0.083            |



**Fig. 5** Haplotype diversity index values for the rice chromosomes in the EA *Indica* accessions, the SA/SEA *Indica* accessions and *Japonica* accessions. The diversity index was calculated using a 10-SNP window. The x axis shows the start position of the haplotype

window. *Triangles* show the positions of the centromeres. The y axis shows the haplotype diversity index, which is calculated as the number of haplotypes divided by the number of cultivars in each group

6 and 10, on the top part of chromosomes 1, 6, 8 and 10, and the bottom part of chromosomes 11 and 12, and so was for the purifying selection acting on the bottom parts of chromosomes 5, 8 and 10, and the top part of chromosome 12.

## Discussion

In this study, 1,633 SNPs distributed evenly across the rice genome were selected and confirmed, providing a resolution of  $\sim 1$  SNP per cM in the rice genome (Fig. S1). From these, a subset of 372 SNPs were used to genotype 300 inbred rice varieties from 22 major rice-growing countries around the world (Fig. S3; Table S5). These accessions were carefully selected to represent maximum geographic diversity of rice and have been used as parental lines in the China National Rice Molecular Breeding Network for 12 years (Yu et al. 2003). Thus, the results presented in this work provide rice scientists with very useful information and efficient tools in their genetic and breeding studies. Furthermore, we identified a subset of 258 high-quality SNP markers, which proven to be more efficient than the whole set SNPs in revealing the subspecific and geographic patterns of genetic diversity of the sampled rice accessions. Because of their evenly distribution along the whole rice genome and relatively low costs, this set of SNPs is thus recommended as the anchor set of rice SNPs for the current uses in large scaled genetic and breeding studies of rice, such as genetic diversity analysis and background selection during breeding processes.

Two additional properties of the SNPs developed in this study make them valuable. The first one was a set of SNPs with rare (allele frequency  $<0.05$ ) and unique alleles (Tables S4, 5). This set of SNPs was typically found in accessions of specific geographic origin and thus are very useful in linkage-based genetic mapping and differentiating unique germplasm accessions. The second one was the fact

that a significant portion of the SNPs used in this study are intragenic. Thus, many of these intragenic SNPs may be associated with loss/gain of function (trait). Future genetic and breeding studies using these SNPs may allow detection and validation of these allele–trait associations, and thus their directly application for trait improvement by genome selection and marker-assisted breeding.

Because all SNPs selected and confirmed in this study were based primarily on the genomic sequence differences between MH63 and ZS97, and partially on that between MH63 and Nipponbare, this set of SNPs are powerful in differentiating differences between *Indica* and *Japonica* and between *Indica* accessions, but have a limited power in differentiating *Japonica* accessions. Thus, this set SNPs expectedly underestimated the genetic diversity in the *Japonica* accessions. Fortunately, with the high density of SNPs in the rice genome (Feltus et al. 2004; Shen et al. 2004) and availability of the high-throughput sequencing technologies (Chi 2008; Hillier et al. 2008; Ossowski et al. 2008; Schuster 2008; Yamamoto et al. 2010), we are in the process of expanding this anchor set of SNPs to overcome this limitation and developing multiple SNP sets of different throughputs for different research purposes. Also, when combined with efforts of other researcher groups (Tung et al. 2010; McCouch et al. 2010; McNally et al. 2009; Huang et al. 2010; Xie et al. 2010; Yamamoto et al. 2010; Zhao et al. 2010), we expect rice SNP data resource increases quickly and SNP genotyping strategies with different throughputs have been and will be created to benefit all the rice researchers around the world.

The general subspecific and geographic patterns of genetic diversity in the sampled rice accessions revealed by the SNPs were consistent with the isozyme and SSRs results of largely the same set of materials (Li and Rutger 2000; Yu et al. 2003), but the SNP results provide more deliberate details of the population structure of the sampled rice accessions. In particular, some common and contrasting patterns of the haplotype diversity along different rice

chromosomes in the *Indica* and *Japonica* accessions suggest different evolutionary forces possibly acting in specific regions of the rice genome during domestication and evolution of rice, which raised interesting questions on what gene(s) and their associated function(s) or trait(s) in each of these genomic regions are responsible for the observed patterns.

**Acknowledgments** We are grateful to Drs. Qifa Zhang and Weibo Xie for sharing the SNP information between MH63 and ZS97 before their publication. We thank Hao Chen, Li Wang, Qiushi Huang, and Tiantian Zhu for their assistances in SNP verification. We appreciate Dr. Judy Lee's help for manuscript editing. This work was supported by grants from the Bill and Melinda Gates foundation (51587-5), the Ministry of Science and Technology of China (2009DFB30030 and 2010AA101806), the Ministry of Agriculture of China (2008ZX08012-005, 2009ZX08012-021B, 2011-G2B and 2006-G51) and the Generation Challenge Program (#12) of CGIAR.

## References

- Akhunov E, Nicolet C, Dvorak J (2009) Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. *Theor Appl Genet* 119:507–517
- Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32:314–331
- Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, Goodman MM, Harjes C, Guill K, Kroon DE, Larsson S, Lepak NK, Li H, Mitchell SE, Pressoir G, Peiffer JA, Rosas MO, Rocheford TR, Romay MC, Romero S, Salvo S, Sanchez Villeda H, da Silva HS, Sun Q, Tian F, Upadhyaya N, Ware D, Yates H, Yu J, Zhang Z, Kresovich S, McMullen MD (2009) The genetic architecture of maize flowering time. *Science* 325:714–718
- Chi KR (2008) The year of sequencing. *Nat Methods* 5:11–14
- Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95(25):14863–14868
- Fan JB, Gunderson KL, Bibikova M, Yeakley JM, Chen J, Wickham Garcia E, Lebruska LL, Laurent M, Shen R, Barker D (2006) Illumina universal bead arrays. *Methods Enzymol* 410:57–73
- Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) An SNP resource for rice genetics and breeding based on subspecies *indica* and *japonica* genome alignments. *Genome Res* 14:1812–1819
- Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. *Genetics* 169:1631–1638
- Glaszmann JC (1987) Isozymes and classification of Asian rice varieties. *Theor Appl Genet* 74:21–30
- Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson RK, Mardis ER (2008) Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Methods* 5:183–188
- 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, Lin Z, Buckler ES, Qian Q, Zhang QF, Li J, Han B (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 42:961–967
- International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436:793–800
- Li ZK, Rutger JN (2000) Geographic distribution and multilocus organization of isozyme variation of rice (*Oryza sativa* L.). *Theor Appl Genet* 101:379–387
- Lu BR, Zheng KL, Qian HR, Zhuang JY (2002) Genetic differentiation of wild relatives of rice as assessed by RFLP analysis. *Theor Appl Genet* 106:101–106
- Lu Y, Yan J, Guimaraes CT, Taba S, Hao Z, Gao S, Chen S, Li J, Zhang S, Vivek BS, Magorokosho C, Mugo S, Makumbi D, Parentoni SN, Shah T, Rong T, Crouch JH, Xu Y (2009) Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. *Theor Appl Genet* 120:93–115
- McCouch SR, Zhao K, Wright M, Tung C, Ebana K, Thomson M, Reynolds A, Wang D, DeClerck G, Ali ML, McClung A, Eizenga G, Bustamante C (2010) Development of genome-wide SNP assays for rice. *Breed Sci* 60:524–535
- McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, Brown P, Browne C, Eller M, Guill K, Harjes C, Kroon D, Lepak N, Mitchell SE, Peterson B, Pressoir G, Romero S, Oropeza Rosas M, Salvo S, Yates H, Hanson M, Jones E, Smith S, Glaubitz JC, Goodman M, Ware D, Holland JB, Buckler ES (2009) Genetic properties of the maize nested association mapping population. *Science* 325:737–740
- McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureau TE, Stokowski R, Ballinger DG, Frazer KA, Cox DR, Padhukasahasram B, Bustamante CD, Weigel D, Mackill DJ, Bruskiewich RM, Ratsch G, Buell CR, Leung H, Leach JE (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc Natl Acad Sci USA* 106:12273–12278
- Monna L, Ohta R, Masuda H, Koike A, Minobe Y (2006) Genome-wide searching of single-nucleotide polymorphisms among eight distantly and closely related rice cultivars (*Oryza sativa* L.) and a wild accession (*Oryza rufipogon* Griff.). *DNA Res* 13:43–51
- Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, Russell JR (2010) Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. *Theor Appl Genet* 120:1525–1534
- Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21:2194–2202
- Ni J, Colowit PM, Mackill DJ (2002) Evaluation of genetic diversity in rice subspecies using microsatellite markers. *Crop Sci* 42:601–607
- Olufowote JO, Xu Y, Chen X, Park WD, Beachell HM, Dilday RH, Goto M, McCouch SR (1997) Comparative evaluation of within-cultivar variation of rice (*Oryza sativa* L.) using microsatellite and RFLP markers. *Genome* 40:370–378
- Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* 18:2024–2033
- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425
- Schuster SC (2008) Next-generation sequencing transforms today's biology. *Nat Methods* 5:16–18
- Shen YJ, Jiang H, Jin JP, Zhang ZB, Xi B, He YY, Wang G, Wang C, Qian L, Li X, Yu QB, Liu HJ, Chen DH, Gao JH, Huang H, Shi TL, Yang ZN (2004) Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. *Plant Physiol* 135:1198–1205

- Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24:1596–1599
- Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10(3):512–526
- Tung CW, Zhao K, Wright MK, Ali ML, Jung J, Kimball J, Tyagi W, Thomson MJ, McNally K, Leung H, Kim H, Ahn SN, Reynolds A, Scheffler B, Eizenga G, McClung A, Bustamante C, McCouch SR (2010) Development of a research platform for dissecting phenotype–genotype associations in rice (*Oryza* spp.). *Rice* 3:205–217
- Wang ZY, Second G, Tanksley SD (1992) Polymorphism and phylogenetic relationships among species in the genus *Oryza* as determined by analysis of nuclear RFLPs. *Theor Appl Genet* 83:565–581
- Waugh R, Jannink JL, Muehlbauer GJ, Ramsay L (2009) The emergence of whole genome association scans in barley. *Curr Opin Plant Biol* 12:218–222
- Xie W, Feng Q, Yu H, Huang X, Zhao Q, Xing Y, Yu S, Han B, Zhang Q (2010) Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc Natl Acad Sci USA* 107:10578–10583
- Yamamoto T, Nagasaki H, Yonemaru J, Ebana K, Nakajima M, Shibaya T, Yano M (2010) Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. *BMC Genomics* 11:267
- Yu J, He S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y et al (2002) A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. indica). *Science* 296:79–92
- Yu SB, Xu WJ, Vijayakumar CHM, Ali J, Fu BY, Xu JL, Marghirang R, Domingo J, Jiang YZ, Aquino C, Virmani SS, Li ZK (2003) Molecular diversity and multilocus organization of the parental lines used in the International Rice Molecular Breeding Program. *Theor Appl Genet* 108(1):131–140
- Zhang Q (2007) Strategies for developing Green Super Rice. *Proc Natl Acad Sci USA* 104:16402–16409
- Zhao K, Wright M, Kimball J, Eizenga G, McClung A, Kovach M, Tyagi W, Ali ML, Tung CW, Reynolds A, Bustamante CD, McCouch SR (2010) Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. *PLoS One* 5:e10780