A whole-genome SNP array (RICE6K) for genomic breeding in rice

Huihui Yu1,†, Weibo Xie2,‡, Jing Li1, Fasong Zhou1,* and Qifa Zhang2,‡

1Life Science and Technology Center, China National Seed Group Co., Ltd, Wuhan, China
2National Key Laboratory of Crop Genetic Improvement, National Center of Crop Molecular Breeding, Huazhong Agricultural University, Wuhan, China

Received 26 June 2013; revised 25 July 2013; accepted 30 July 2013.

*Correspondence (Tel 86-27-87570526 and 86-27-87282429; fax 86-27-87570511 and 86-27-87287092; emails zhoufasong@sinochem.com; qifazhang@hzuAU.edu.cn)
†These authors contributed equally to this work.
The authors H. Yu, J. Li and F. Zhou have commercial interest in RICE6K as employees of China National Seed Group Co., Ltd.

Keywords: RICE6K, SNP chip, genomic breeding, functional markers, rice (Oryza sativa L.).

Summary
The advances in genotyping technology provide an opportunity to use genomic tools in crop breeding. As compared to field selections performed in conventional breeding programmes, genomics-based genotype screen can potentially reduce number of breeding cycles and more precisely integrate target genes for particular traits into an ideal genetic background. We developed a whole-genome single nucleotide polymorphism (SNP) array, RICE6K, based on Infinium technology, using representative SNPs selected from more than four million SNPs identified from resequencing data of more than 500 rice landraces. RICE6K contains 5102 SNP and insertion–deletion (InDel) markers, about 4500 of which were of high quality in the tested rice lines producing highly repeatable results. Forty-five functional markers that are located inside 28 characterized genes of important traits can be detected using RICE6K. The SNP markers are evenly distributed on the 12 chromosomes of rice with the average density of 12 SNPs per 1 Mb and can provide information for polymorphisms between indica and japonica subspecies as well as varieties within indica and japonica groups. Application tests of RICE6K showed that the array is suitable for rice germplasm fingerprinting, genotyping bulked segregating pools, seed authenticity check and genetic background selection. These results suggest that RICE6K provides an efficient and reliable genotyping tool for rice genomic breeding.

Introduction
In the history of plant breeding, cross-pollination and transgenic technologies have twice revolutionized the way of crop improvement in the beginning and the end of last century, respectively. Plant breeders realized gene recombination within species through controlled pollination and introduced genes for alien traits, such as insect resistance and herbicide tolerance, into crops via transformation. The advances of genomic research in the last decade may have afforded tools and resources for a third technology breakthrough in plant breeding, which may be termed ‘genomic breeding’. By genomic breeding, plant breeders can explore the genomic information including DNA sequences and gene functions to design ideal genotypes and conduct selection to modify the whole genome for varietal improvement. In practice, genomic breeding would select the genes of interest for target traits using molecular markers and optimize genetic background based on genome-wide DNA polymorphisms. Compared to field selection based only on phenotype in conventional breeding, whole-genome selection can integrate the target genes into a better-defined genetic background with greatly improved efficiency.

Two main types of high-throughput genotyping platforms are now available in technology market that can be adopted for genomic breeding: DNA sequencing (Davey et al., 2011) and DNA array (Gupta et al., 2008). New sequencing technologies have been widely applied in genetic studies (Metzker, 2010; Varshney et al., 2009). In rice (Oryza sativa L.), Huang et al. (2009) reported a high-throughput resequencing method to genotype a recombinant inbred line (RIL) population (Huang et al., 2009). Xie et al. (2010) developed a genotype-imputing method to construct haplotypes using low-coverage genome resequencing data (Xie et al., 2010). Both methods were validated in genetic studies for gene discovery (Wang et al., 2011; Yu et al., 2011). Diverse germplasm collections of rice have been resequenced for genome-wide association studies of agronomic traits (Huang et al., 2010, 2011). Many valuable linkage disequilibria, quantitative trait loci (QTLs) and genes have been identified from the analyses, and millions of DNA polymorphisms were detected in the genomic data. Such unprecedented large amounts of information have laid a solid foundation for platform development for genomic breeding.

Various DNA array-based genotyping platforms have been developed and tested for genetic studies and germplasm characterization (Borevitz et al., 2003; Jaccoud et al., 2001; McNally et al., 2009; Miller et al., 2007a,b; Wang et al., 2010; Xie et al., 2009). For crops such as rice that has a complete genome sequence available, single nucleotide polymorphism (SNP) array has been taken as the preferred technique because of its high-density, assay accuracy, simple data analysis and easy data exchange between research programmes. In rice, three major SNP assay platforms built on different assay principles are available. Affymetrix gene-chip detects SNPs based on differential hybridization efficiency between DNA probes and template sequences; a Rice 44K SNP genotyping array was developed and applied in genome-wide association studies (Famoso et al., 2011; McCouch et al., 2010; Tung et al., 2010; Zhao et al., 2011). Illumina GoldenGate SNP Chip detects SNPs based on DNA extension and differential ligation; various SNP chips of this type were also developed for rice and used in different genetic analysis
and breeding projects (Boualaphanh et al., 2011; Chen et al., 2011; Nagasaki et al., 2010; Thomson et al., 2012; Yamamoto et al., 2010; Zhao et al., 2010). As the most recent SNP assay technology, Illumina Infinium SNP array is based on differential single nucleotide extension. Data of genome screening obtained in human, animals and plants demonstrated that this technique enjoyed the advantages of high specificity, reproducibility and call rate (Oliphant et al., 2002; Steemers and Gunderson, 2007). It has now been used in human disease diagnosis, and genetic and breeding studies in crops including maize (Cook et al., 2012; Ganal et al., 2011), wheat and barley (Miedaner and Korzun, 2012), and oilseed rape (Snowdon and Iniguez Luy, 2012). These molecular marker assay platforms have provided options for breeding applications.

Genotyping technologies for genetic studies and breeding share some commonalities but at the same time have significant distinctions. Both need the capacity to detect molecular markers evenly distributed in the genome. Genetic studies aim at revealing unknowns by discovering new genes or QTLs, whereas breeders need genotyping tools that can quickly and reliably find the known targets to identify the expected genotype. Moreover, high-throughput molecular marker assay platform for breeding application has to be cost-effective, time-saving, technically reliable, easy to use and widely adaptable for various breeding applications.

Here, we report on our effort in developing a SNP array for rice genomic breeding. We extracted millions of SNPs based on data from resequencing of rice germplasm collections and designed a rice whole-genome SNP array using Infinium technology. We showed that this platform is useful for a range of applications.

Results

Design of RICE6K

A rice whole-genome SNP array was designed essentially for efficient progeny screening in rice breeding with two considerations, genetic background selection and genotyping of target genes. Illumina BeadArray technology and Infinium SNP assay platform were chosen for the SNP array fabrication because of its demonstrated high specificity, reproducibility and accuracy in SNP call (Oliphant et al., 2002; Steemers and Gunderson, 2007). Two sorts of DNA variations were considered in designing the array: (i) SNPs with adequate coverage and representation of the genome diversity judged on the basis of resequencing diverse germplasm collections (Huang et al., 2010) and (ii) allelic variations of characterized functional genes controlling important breeding traits (Jiang et al., 2011).

Probes for genome diversity

The selection of the probes took several steps. In the first step, raw sequences of 4 236 029 SNP sites were identified from low-coverage (~1x) genome sequences of 520 rice accessions (Huang et al., 2010). In doing so, the released assembly version 6.1 of genomic pseudomolecules of japonica cv. Nipponbare (http://rice.plantbiology.msuk.edu) was used as the reference genome. Sequence reads of all accessions were aligned to the reference genome using software MAQ (Li et al., 2008). SNPs were identified using custom PERL scripts from output of MAQ. The following criteria were applied in the processes of sequence comparison and SNP identifications. The mapping quality of sequence reads must be ≥20. At each SNP site, there are at least ten sequence reads showing consensus to each of the two polymorphic nucleotides, of which at least five reads each had mapping quality >40 and the corresponding base quality >20. Because the average sequence coverage after sequence alignment was ~400 x, we limited the total number of sequence reads covering a SNP site between 50 and 1000 to avoid possible repeat sequences.

In a predominantly self-pollinating species like rice, a plant from a germplasm collection is expected to be highly homozygous. Thus, a SNP site would be removed from the candidate list if more than five germplasm accessions showed heterozygous genotype. A SNP site would also be removed if: (i) more than 800 or less than 80 sequence reads covering the SNP site were obtained or (ii) the added frequency of a minor allele in indica group (374 accessions), in japonica group (146 accessions) and in all germplasm accessions is less than 0.2 or (iii) there are other SNP sites within the flanking 50-bp sequences. This process reduced the number of candidate SNP sites to 1 559 745.

In the following step, the 50-bp flanking sequences on both sides of each selected SNP site were extracted and aligned against the genome sequence of Nipponbare using BLAST program (Kent, 2002), and the SNP site was removed from the candidate list if more than one matched hits were found with identity >85% of either side flanking sequence. The 50-bp flanking sequences on both sides of the SNP site were extracted from genome sequences of Zhenshan 97 and Minghui 63, two typical indica varieties, and compared to the reference Nipponbare sequence. The SNP was excluded if the 50-bp flanking sequences in both Zhenshan 97 and Minghui 63 were different from those in Nipponbare at both sides. This screen kept 1 055 959 candidate SNP sites in the list, of which 35.5% show polymorphism between japonica and indica (with an allele frequency >0.9 in one subspecies and <0.1 in the other), 42.1% between two random indica accessions and 16.9% between two japonica accessions.

To further reduce the number of candidate SNPs, the distribution of the selected 1 055 959 SNPs on chromosomes was displayed in windows of 100 kb. Linkage disequilibrium between closely linked SNP sites was evaluated by squared correlation coefficient ($r^2$) with the threshold value set at 0.64. Two SNP sites with the $r^2$ ≥ 0.64 were placed in the same group using a greedy algorithm (Carlson et al., 2004), which resulted in a total of 86 075 groups. With one or two SNPs selected from each group, a total of 187 284 SNPs were used as probe candidates. These selected SNPs (called tag-SNPs) and the corresponding flanking sequences were submitted to Illumina Inc. (http://www.illumina.com/) for probe screen. After removing the tag-SNPs with a design score <0.6, a total of 115 740 SNP sites met the Illumina Infinium probe designing criteria.

To select the final set of SNPs, we defined an In/Ja SNP such that it could differentiate between the main alleles (>90% frequency) of indica and japonica. Since indica–japonica differentiation represents most of the genome diversity in rice germplasms, we chose two In/Ja SNPs in each 100-kb region. Other types of SNPs were added when there were less than two In/Ja SNPs in an 100-kb region. In Infinium SNP assays, two bead types are used to detect an A/T or G/C SNP (Infinium I type SNPs), while only one bead type is needed for other types of SNPs, such as A/G, A/C, T/G, T/C (Infinium II type SNPs). In order to put as many SNPs as possible on the chip with a total of 6000 bead types, we defined an empirical scoring system: $S = MAF + T^2 + 3.5$, where MAF is the minor allele frequency (%) of the SNP site, and $T = 1$ for Infinium II SNPs (non-A/T or G/C SNPs) or $T = 0$ for Infinium I SNPs. A SNP site with $S ≥ 33$ would be selected.
Eventually a total of 5556 SNP sites were selected from the 115 740 tag-SNPs, which together with the corresponding flanking sequences were used for synthesizing the probes.

Probes for functional genes

More than 600 rice genes controlling important agronomic traits and biological processes had been identified and characterized at the time when the array was designed (Jiang et al., 2011). To incorporate these genes in the array, we identified SNP/InDel sequences inside 40 functional genes for important agronomic traits that were isolated by map-based cloning. A gene-specific probe (functional probe), either a SNP or an InDel sequence, that represents a characterized function or phenotype was designed for each selected gene.

To put functional markers (FMs) on RICE6K array, we selected genes isolated via map-based cloning that are functionally important to agronomy traits. First, different alleles of 40 functionally characterized genes were identified by searching publications, and the related allele sequences were downloaded from the public DNA database (http://www.ncbi.nlm.nih.gov). These sequences were aligned together and subsequently the polymorphic SNP/InDel markers were developed. These markers were then converted into Infinium probes. If the functionally characterized polymorphic site of a gene is a single SNP and has no other SNPs or mutations in the 50-bp flanking sequence of one side, the conserved 50-bp sequence next to the SNP was directly used as a candidate probe. If the functional polymorphic site was an InDel, two strategies were taken to develop functional probes (Figure S1). One strategy was to convert the InDel marker to a SNP marker: the probe was developed from the conserved side of the insertion/deletion, and the nucleotide to be detected was the first base of the insertion sequence or the base following the deletion if a polymorphism between functional and non-functional alleles existed. The other strategy was to directly use the specific insertion sequence as a probe, and thus, the genomic sequence of the insertion allele could be detected with a strong signal, but the counterpart allele would show very low signal. The first-type FM is codominant and the second type is dominant.

In total, 80 functional probes for 40 genes controlling traits like grain yield, grain quality, heading date, hybrid fertility, biotic and abiotic stress resistance were included in the array.

Quality assay of the RICE6K array

A total of 5636 markers including 5556 SNPs for genetic diversity and 80 for specific gene functions were synthesized and put on RICE6K chip (Table S1). Genotyping accuracy and SNP call rate of RICE6K were tested following the recommended protocols. In order to establish an accurate genotyping procedure, RICE6K was used to genotype rice varieties and a F2 population derived from a cross between Balilla and Nanjing 11. Consistent with the corresponding phenotypes (Table 2). Additionally, the functional marker, Os07g15770.2, could distinguish three alleles of the pleiotropic gene Ghd7. The RICE6K array detected three genotypes in our tested lines, ‘C’, ‘A’ and ‘NC’ (no call due to low detecting signal), corresponding to the functional allele (e.g. Minghui 63) and two non-functional alleles, Ghd7-0a, a premature termination in the predicted coding region (e.g. Mudanjiang 8), and Ghd7-0, with the Ghd7 locus completely deleted (e.g. Zhenshan 97) (Xue et al., 2008). In total, these assays identified 4473 high-quality markers including 45 functional markers that are evenly distributed on the 12 chromosomes with an average of 12 markers per Mb (Figure S2).

We also tested reproducibility of the RICE6K array in genotyping by assaying four independent DNA samples of indica variety ‘93-11’. Among the 5102 markers of the RICE6K array, only 0–5 SNPs (<0.1%) showed different genotypes between any two tested samples and no difference was detected in the results from the 4473 high-quality markers, indicating high reproducibility of the array in genotyping.

The genotyping data of the 106 unique accessions selected from the 112 tested inbred lines (six duplicates identified in this work were removed in the following analysis; Table S3) were used to predict the number of polymorphic markers between any two varieties. The 106 rice varieties were clearly clustered into two groups based on the genotypes of 4473 high-quality

Table 1 Markers on the RICE6K array

<table>
<thead>
<tr>
<th>Types</th>
<th>Synthesized markers</th>
<th>Detected markers</th>
<th>High-quality markers</th>
<th>%*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic diversity</td>
<td>5556</td>
<td>5034</td>
<td>4428</td>
<td>87.96</td>
</tr>
<tr>
<td>Gene function</td>
<td>80</td>
<td>68</td>
<td>45</td>
<td>66.18</td>
</tr>
<tr>
<td>Total</td>
<td>5636</td>
<td>5102</td>
<td>4473</td>
<td>87.67</td>
</tr>
</tbody>
</table>

*Percentage of high-quality markers = High-quality markers/Detected markers × 100%.
SNP/InDel markers, and all 18 japonica varieties were clustered in one group and the rest 88 indica ones in the other. Additionally, subgroups in japonica and indica were also defined with a fine resolution (Figure S3). The average high-quality polymorphic markers between any two tested varieties are 1559, and the ones between two tested indica varieties, two tested japonica varieties or between tested indica and japonica are 1053, 824 and 2853, respectively (Figure 1). This result suggests that RICE6K array can be widely used to genotype different populations derived from different crosses, not only for interspecies between indica and japonica, but also for intra-subspecies.

Applications of the RICE6K

We validated the usefulness of the RICE6K array in a range of applications including genomic breeding and genetic analysis.

Table 2 Genotypes of Balilla and Nanjing11 at gene functional markers detected by RICE6K array

<table>
<thead>
<tr>
<th>Gene</th>
<th>MSU locus</th>
<th>Probe</th>
<th>Genotype</th>
<th>Table 2 Genotypes of Balilla and Nanjing11 at gene functional markers detected by RICE6K array</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sd-1</td>
<td>LOC_Os01g66100</td>
<td>ID01g005D1.1</td>
<td>Semi-dwarf plant</td>
<td>Balilla</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ID01g005D1.2</td>
<td>Semi-dwarf plant</td>
<td>High plant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Os01g66100.1</td>
<td>Semi-dwarf plant</td>
<td>High plant</td>
</tr>
<tr>
<td>TAC1</td>
<td>LOC_Os09g35980</td>
<td>Os09g35980.1</td>
<td>Compact plant</td>
<td>Nanjing11</td>
</tr>
<tr>
<td>Gn1a</td>
<td>LOC_Os01g10110</td>
<td>Os01g10110.1</td>
<td>Small spike</td>
<td>Spread-out plant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Os01g10110.2</td>
<td>Small spike</td>
<td>Big spike</td>
</tr>
<tr>
<td>SaF</td>
<td>LOC_Os01g39670</td>
<td>Os01g39670.1</td>
<td>Japonica type</td>
<td>indica type</td>
</tr>
<tr>
<td>S5</td>
<td>LOC_Os06g11010</td>
<td>Os06g11010.1</td>
<td>Japonica type</td>
<td>indica type</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Os06g11010.2</td>
<td>Japonica type</td>
<td>indica type</td>
</tr>
<tr>
<td>GW2</td>
<td>LOC_Os02g14720</td>
<td>Os02g14720.2</td>
<td>Wide grain</td>
<td>Narrow grain</td>
</tr>
<tr>
<td>qSWS5/GW5</td>
<td>(GeneBank: AB433345)</td>
<td>Os05g000GW5.1</td>
<td>Wide grain</td>
<td>Narrow grain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ID05g000GW5.3</td>
<td>Wide grain</td>
<td>Narrow grain</td>
</tr>
</tbody>
</table>

Figure 1 Polymorphism marker number distributions between two varieties. For each histogram, x-axis shows the number of polymorphism marker between two varieties and y-axis shows the number of pairs.

Genetic background selection in breeding process

The improvement of a specific trait by backcrossing is an important breeding strategy, aiming to transfer a desired trait from a donor germplasm into the genome of an elite variety without disturbing the genetic background. It is critical to be able to track the DNA fragments from the different genomes. Kongyu 131, a japonica cultivar widely grown in north-east China in the last decade, has become highly susceptible to rice fungal blast (Magnaporthe grisea) in recent years. A genomics-based introgression of Pi1 (Hua et al., 2012) and Pi2 (Zhou et al., 2006) from the donors into Kongyu 131 has been implemented using SSR markers. In BC4F1, 29 plants with the introduced Pi1 or Pi2 genes were examined using the RICE6K array, which provided an unambiguous graphic genotype for each individual (Figure 2). The results showed that the genetic backgrounds of several tested plants were similar to the...
recurrent parent Kongyu 131 (e.g. L16, L22, L24, L25 and L28), but the genomic regions containing Pi1 and Pi2 had large dragged fragments from the donor parents, which may have potential adverse genetic effects (Figure 2). Indeed, a flowering gene Hd1 (9.3 Mb on Chr06) (Yano et al., 2000) was linked to Pi2 (10.4 Mb on Chr06). Transferring late flowering allele into Kongyu 131 is undesirable for the improvement of this variety to be planted in targeted area. This result suggested selection for recombination must be performed in early generations of backcrossing as suggested by Chen et al. (2000). Nonetheless, it demonstrated that the RICE6K array can provide a powerful tool for genotype selection.

Genotyping biparental segregating populations

We genotyped individual lines/plants and their parents in several biparental cross-populations including (i) an F2 population from a cross between an indica variety Nanjing 11 and a japonica variety Balilla, (ii) a RIL population derived from a cross between two indica varieties, (iii) a RIL population from a cross between an indica and a japonica variety and (iv) CSSL (chromosomal segmental substitution lines) from three different crosses (Table 3). Among the 4473 high-quality SNP/InDel markers on the RICE6K array, the number of markers detected to be homozygous and polymorphic for the two parents of each tested population was ranged from 1336 to 3775: more than 3000 in inter-subspecific populations and more than 1000 in indica populations (Table 3). For each population, the genotypes of the called SNPs were assigned as ‘AA’ (female parental genotype), ‘BB’ (male parental genotype) or ‘AB’ (heterozygous genotype). As a result, the called high-quality polymorphic SNPs could provide high-density graphical genotypes for each individual (Figure 3), which can be used for genetic investigations. For example, genotyping the 197 lines in the ZX-RIL population using the RICE6K array resulted in a high-density genetic linkage map consisting of 1495 recombination bins covering 1591.2 cM with average length of 1.1 cM per bin (Tan et al., 2013). The total length of the genetic map was similar to the ones reported in previous studies using sequence-based genotyping method (Huang et al., 2009; Yu et al., 2011). These tests indicated that RICE6K SNP array is robust and efficient in population genotyping. 

Varietal identity and purity tests

Identity and purity of seeds are always of main concern in crop production and seed industry. In China, a set of 24 SSR markers...
has been officially used for testing identity and purity of rice varieties (Chinese Agricultural Industry Standard, NY/T 1433-2007), in which two samples are regarded as different varieties if two or more markers show polymorphisms, as related varieties if one marker is different, and as the same variety if no polymorphism is detected in all the 24 markers. Although the markers were carefully selected with even distribution on the 12 chromosomes and showing high polymorphisms in tested varieties (Zhuang et al., 2006), these 24 SSR markers cover only a very small partition of the rice genome. We used the RICE6K array to fingerprint rice varieties that produced results that challenge the notion of ‘varieties’. For example, YU is not only an elite inbred rice variety in China, but also the parent of an elite hybrid, which has been widely planted for more than 5 years. A sample of bulked seeds of YU was fingerprinted using RICE6K, which revealed several genomic regions that were heterogeneous and

Table 3 Population tested on RICE6K array

<table>
<thead>
<tr>
<th>Population</th>
<th>Cross</th>
<th>Type</th>
<th>Polymorphism markers</th>
<th>%*</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZM-RIL</td>
<td>ZS 97 (R) × MH 63 (R)</td>
<td>Indica × Indica</td>
<td>1336</td>
<td>26.18</td>
</tr>
<tr>
<td>ZX-RIL</td>
<td>Zhenshan 97 × Xizang 2</td>
<td>Indica × Japonica</td>
<td>3362</td>
<td>65.90</td>
</tr>
<tr>
<td>BN-F2</td>
<td>Balilla × Nanjing 11</td>
<td>Indica × Japonica</td>
<td>3775</td>
<td>73.99</td>
</tr>
<tr>
<td>ZN-CSSL</td>
<td>Zhenshan 97 × Nipponbare</td>
<td>Indica × Japonica</td>
<td>3709</td>
<td>72.70</td>
</tr>
<tr>
<td>ZM-CSSL</td>
<td>Zhenshan 97 × Minghui 63</td>
<td>Indica × Indica</td>
<td>1342</td>
<td>26.30</td>
</tr>
<tr>
<td>ZI-CSSL</td>
<td>Zhenshan 97 × Oryza rufipogon (IRGC-105491)</td>
<td>Indica × Oryza rufipogon</td>
<td>1789</td>
<td>35.06</td>
</tr>
</tbody>
</table>

RIL, recombinant inbred line; CSSL, chromosomal segmental substitution line.

*Percentage of polymorphism markers detected by the RICE6K array.
thus still segregating (Figure 4). The two main heterogeneous regions were 9–11.5 Mb on Chr02 and 0–7 Mb on Chr09. Using one of the plants designated YU001 as a reference, four major types of heterogeneous plants were identified from the population. However, none of the heterogeneous regions could be identified using the 24 SSR markers. These results indicated that genetic heterogeneity would be retained in a variety for a long period, which might be the cause of variety deterioration. This makes selection and subsequent field tests necessary a few years after variety release, which also suggests that genotyping the selected lines using a method like the RICE6K before varietal release may help maintaining the quality and purity of the varieties.

**Bulked segregant analysis**

Bulked segregant analysis (BSA) is an efficient method for rapidly identifying molecular markers linked to any specific genes or genomic regions (Michelmore et al., 1991). We tested the application of RICE6K array in BSA by mapping the fertility-restorer gene-controlling cytoplasmic male sterility (CMS). The three-line hybrid F1 plants, derived from a cross between CMS line JN 2A and restorer line JH 3, were self-pollinated to generate an F2 population of about 2000 plants. The F2 population was planted in a field nursery in Sanya, China, and 94 plants in the population were examined for spikelet fertility. The ratio of the fertile to sterile plants was 74 : 20 ($\chi^2 (3 : 1) = 0.695, \ P = 0.405$), indicating that the fertility was controlled by a single locus. DNA bulks from ten fertile plants and 20 sterile plants were separately prepared and assayed using the RICE6K array. The result showed that the main difference between the two bulks was in region 18.1–19.9 Mb on Chr10, where the bulk from fertile plants was heterozygous and the bulk of sterile plants was homozygous (Figure 5). There are two characterized rice fertility-restorer genes in this region, one is Rf1/Rf1a/Rf5 (18.8 Mb) and the other is Rf1b (18.9 Mb) (Akagi et al., 2004; Hu et al., 2012; Wang et al., 2006), and some other fertility-restorer genes have been also mapped to this region, such as Rf4 and Rf6 (Ahmad-ikhah and Karlov, 2006; Liu et al., 2004). Thus, the RICE6K can be used to quickly locate the gene to the genomic region. Moreover, the polymorphic SNP markers in the region identified by RICE6K array can be used for fine mapping of this gene.

In conclusion, the results show that the RICE6K chip provides a robust tool for a range of applications in genotyping.

**Discussion**

One of the biggest challenges that plant breeders have to face is to stack multiple target genes in a favourable genetic background. This requires the breeders to be able to track many functional genes in a segregating population in a short time and reliable manner at a reasonable cost. RICE6K accommodated 80 functional markers covering 40 rice genes. In our tests using various genetic materials, at least 45 functional markers representing 28 genes, 1–5 markers per gene, performed well. The parallel interrogation of dozens functional genes can greatly facilitate gene stacking.

Some of the genes have several functional mutations, for example, Hd1 for heading date (Takahashi et al., 2009). We thus designed multiple probes in the RICE6K to detect these alleles, which detected several haplotypes for determining functional type of the gene. For functional mutations located inside repeated motifs of a gene, for example, Pi2/Pi9/Piz-t (Qu et al., 2006; Zhou et al., 2006), a locus for blast resistance with multiple alleles, it is difficult to design a probe to differentiate the alleles. In such case, we also recommend to use multiple SNP probes differentiating functional allele haplotypes for the screening.

Agronomically important traits like hybrid vigour, grain yield and quality are affected by many genes/QTLs each with a small effect on the trait. Superior varieties or hybrids are the results of gradual accumulation of many favourite alleles through repeated

---

**Figure 4** Comparative genotyping of different plants in ‘YU’ population. The plant ‘YU001’ is the reference genotype. The figure shows the maps of different plants compared with the reference plant ‘YU001’ genotyped by the RICE6K array. The single nucleotide polymorphism (SNP) genotype is assigned to ‘AA’ when the detected genotype of the plant is the same as the referent plant ‘YU001’ and is not shown. The SNP genotype is assigned to ‘BB’ when the genotypes of the plant and the reference are different and both of them are homologous. The SNP genotype is assigned to ‘AB’ when the genotype of the reference is homologous and one of the plants is heterozygous. (a) Genotyping map of the plant ‘YU003’, (b) genotyping map of the plant ‘YU010’, (c) genotyping map of the plant ‘YU033’ and (d) genotyping map of mixture sample harvested from random 20 plants of the population. The triangle arrows indicate the positions of 24 SSR markers recommended by the Chinese Agricultural Industry Standard (NY/T 1433-2007).
crossing and phenotype selections in the long breeding history. These varieties probably represent the best combinations of genes controlling yield, quality and adaptation to climate and cultivation conditions. However, they will inevitably become susceptible to diseases or other biotic stresses because of emergence of new strains of the pathogens. In this case, a quick backcrossing scheme is usually performed to incorporate a specific gene to improve the resistance with minimal disturbance of the genetic background. In previous reports, breeders used linked markers (RFLP, SSR or AFLP) to perform selection for recombination between the target gene and the genetic background, and a set of unlinked markers for genetic background screening (Chen et al., 2000; Jorasch, 2005; Liu et al., 2003), which produced variable results. Chip-based high-density genome fingerprinting can greatly improve the efficiency of backcrossing by selecting for precise recombinations and the background of the recurrent parent. We showed that RICE6K array could provide detailed information of the genomic composition of the selected progeny at resolution of <100 kb. This may lead to highly accurate prediction of the performance of the selected individuals. Thus, the RICE6K array would be particularly useful for backcross breeding to introgress genes into elite backgrounds. In addition, knowledge of the genes in the array would also help predict the performance of the selected individuals in other types of breeding programmes.

Genomic information including genome sequences and functional genes is still accumulating on accelerated pace (Jiang et al., 2011), which provides practically unlimited resources for improving genotyping tools like the RICE6K array. Ideally, the genotyping information should provide adequate for detecting polymorphisms in both inter-subspecific (indica/japonica) and intra-subspecific (indica/indica and japonica/japonica) crosses of rice. This may be readily addressed with the present results from the already available sequencing data as well as the still ongoing efforts. It is also desired that the genotyping tools can incorporate information on functional genes as they become available. These will entail economically practical higher-density SNP arrays, which should now be placed in the pipeline for future development.

Experimental procedures
The assay protocol of the RICE6K

First, rice genomic DNA samples were extracted and their quality was examined. Two types of tissues, either seed or leaf, were used for DNA extraction. For seed, about 20 dry seeds were dehulled, mixed and ground after freezing with liquid nitrogen, and the genomic DNA was extracted using CoWin SurePlant DNA Kit (Beijing CoWin Bioscience Co., Ltd., Beijing, China). For leaf, fresh young seedling leaflet of 3–5 cm in length were harvested, ground after freezing with liquid nitrogen, and the genomic DNA was extracted using Wizard Magnetic 96 DNA Plant System Kit (Promega Corporation, Madison, WI). DNA quality was checked by 1%–1.5% agarose gel electrophoresis. The DNA samples with high quality (>10-kb fragments) and appropriate concentration (10–50 ng/µl) were used for SNP assays.

DNA amplification, fragmentation and chip hybridization, washing, and staining were performed according to Infinium assay standard protocol (Infinium HD Assay Ultra Protocol Guide,http://www.illumina.com). HiScan scanner (Illumina Inc., San Diego, CA) was used for chip scanning, and GenomeStudio software was used for raw data analysis. R platform was employed for further analysis, for example, genotype identification, comparison and map drawing (R Development Core Team, 2011).

Acknowledgements
This work was supported by the Introduction of International Advanced Agricultural Science and Technology Program of China (948 Program, Grant No. 2012-G2), the National High Technology Research and Development Program of China (863 Program, Rice Functional Genomics Research Project, Grant No. 2012AA10A304), the National Natural Science Foundation of China (Grant No. 31100962), the Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20111046120013). The lines/plants used in this study for RICE6K SNP chip tests were provided by Yuqing He, Sibin Yu, Xingming Lian and Yongzhong Xing in Huazhong Agricultural University. We highly appreciate the generous support.

References


Supporting information

Additional Supporting information may be found in the online version of this article.

Figure S1 An example of gene functional probes design.

Figure S2 Distribution of the 4473 high-quality markers of RICE6K on rice genome.

Figure S3 Clustering analysis of rice inbred lines based on genotyping data using RICE6K array.

Table S1 Information of 5636 SNP/InDel markers on the RICE6K array.

Table S2 Information of 45 high-quality functional markers on the RICE6K array.

Table S3 The genotypes of 45 functional markers in 112 tested rice inbred lines assayed by RICE6K.