Development of genomics-based genotyping platforms and their applications in rice breeding
Haodong Chen1,3, Hang He1,3, Fasong Zhou2, Huihui Yu2 and Xing Wang Deng1,3,4

Breeding by design has been an aspiration of researchers in the plant sciences for a decade. With the rapid development of genomics-based genotyping platforms and available of hundreds of functional genes/alleles in related to important traits, however, it may now be possible to turn this enduring ambition into a practical reality. Rice has a relatively simple genome comparing to other crops, and its genome composition and genetic behavior have been extensively investigated. Recently, rice has been taken as a model crop to perform breeding by design. The essential process of breeding by design is to integrate functional genes/alleles in an ideal genetic background, which requires high throughput genotyping platforms to screen for expected genotypes. With large amount of genome resequencing data and high-throughput genotyping technologies available, quite a number of genomics-based genotyping platforms have been developed. These platforms are widely used in genetic mapping, integration of target traits via marker-assisted backcrossing (MABC), pyramiding, recurrent selection (MARS) or genomic selection (GS). Here, we summarize and discuss recent exciting development of rice genomics-based genotyping platforms and their applications in molecular breeding.

Addresses
1 Peking-Yale Joint Center of Plant Molecular Genetics and Agrobiotechnology, State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing 100871, China
2 Life Science and Technology Center, China National Seed Group Co., Ltd., Wuhan 430075, China
3 Shenzhen Institute of Crop Molecular Design, Shenzhen 518107, China
4 Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520-8104, USA

Corresponding authors: Chen, Haodong (chenhaodong@pku.edu.cn) and Deng, Xing Wang (xingwang.deng@yale.edu)

Current Opinion in Plant Biology 2013, 16:247–254
This review comes from a themed issue on Genome studies and molecular genetics
Edited by Qifa Zhang and Rod Wing
For a complete overview see the Issue and the Editorial
Available online 21st May 2013
1369-5266/$ – see front matter, © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.pbi.2013.04.002

Introduction
The concept of ‘Breeding by Design’ was proposed by Peleman and van der Voort in 2003, aiming to control allelic variations of all important trait-related genes [1]. However, without enough genome information and high throughput genotyping tools, this concept was hardly practiced in empirical breeding. Since the rice genome was sequenced [2–4], significant advancements have been made in the functional genomics of rice (Oryza sativa L.) that have offered breeders numerous tools and resources to practice breeding by design. The fast accumulation of rice genome resequencing data not only assisted in identification of functional Quantitative Trait Loci (QTL) or genes, but also provided numerous polymorphic genome sequences for molecular marker-development [5*–6,7*,8*]. At the same time, a variety of molecular marker assay platforms with different throughputs have also recently been developed [5*,7*,9*,10]. All these achievements have thus turned the ‘design’ concept into a practical breeding activity. Here we focus on recent development of genomics-based genotyping platforms and their applications in rice molecular breeding by design.

Development of genotyping platforms
Molecular markers are widely used in genetic studies and new variety development. Various types of genotyping technologies have been developed to meet the requirements of genetic research or breeding programs in rice. Restriction Fragment Length Polymorphism (RFLP) and Simple Sequence Repeats (SSR) are the representatives of first and second generation markers [11,12]. They played important roles in construction of rice genetic maps and identification of trait-related loci/genes. SSR markers are still frequently used by breeders to assist their genotype screen for target traits. In this review, we will focus on the most recent development of high throughput genotyping platforms, specifically on DNA array platforms and next-generation sequencing (NGS) technologies.

Array-based genotyping
Recently, a number of array-based genotyping technologies have been developed and some of the representatives including Restriction Site-Associated DNA (RAD), Single Feature Polymorphism (SFP) and Single Nucleotide Polymorphism (SNP) have been widely tested. With short DNA tags as probes RAD markers identify genetic polymorphisms linked to particular restriction sites throughout the genome. RAD markers have been successfully used in genotyping both individuals and segregating bulks [13]. SFP markers detect polymorphic signals resulted from differential hybridization of various alleles to DNA probes arrayed on a microchip. Affymetrix
microarrays, composed of 25 nucleotide probes, have been shown to be efficient and convenient in detecting very large number of SFPs in rice [14–17].

SNPs resulted from single-base pair variations are the most abundant DNA markers that are evenly distributed on a whole genome [18]. Almost any gene or locus can be tagged by SNP markers. Extremely high density of a SNP array can assay large number of SNP markers in a high throughput manner. SNP array-based genotyping platform has been considered as one of most favorite options for gene/QTL mapping and marker-assisted crop breeding in the past decade [19]. Furthermore, with the rapid accumulation of rice genome resequencing data, SNP-based markers will continue to be more widely used than any other type of markers [5**,6,7**,20]. Although SNPs can also be detected via PCR or Sanger sequencing, array-based detection techniques are preferable given that they can satisfy different genotyping requirements. Currently, SNPs can be detected in a variety of throughputs, depending on the objectives. Many SNP assay systems have been developed by different companies, such as Illumina’s SNP chip platforms (http://www.illumina.com), Affymetrix’s SNP array platform (http://www.affymetrix.com), GenomeLab’s SNPstream genotyping system (Beckman Coulter, https://www.beckmancoulter.com), and the TaqMan OpenArray genotyping system (Applied Biosystems, http://www.appliedbiosystems.com). Given that Illumina and Affymetrix SNP arrays are more commonly used in the rice community, they will be discussed here in more detail.

A combination of Veracode and GoldenGate technology on Illumina’s BeadXpress Reader can be used to genotype 48-SNP, 96-SNP, 192-SNP or 384-SNP per sample. SNPs and their flanking sequences are used to design locus-specific and allele-specific primers for GoldenGate assay. This platform, which can be used to assay thousands of individual samples in a short time period, is both reliable and relatively inexpensive. Recently, several 384-SNP assays on this platform have been developed and used for both variation evaluation and genetic diversity analysis in rice [21,22,23*,24*]. This low-density SNP array platform is useful for the genotyping of early generation breeding materials due to its high-throughput capacity in sample processing.

GoldenGate SNP Beadarrays can provide medium resolution genotyping results. For example, a 1536-SNP GoldenGate array was designed to detect polymorphism within and between the five major subpopulations of O. sativa [25]. In another study, a set of 2688 SNPs were used to genotype 151 Japanese rice cultivars that had been released over the last 150 years [26].

The high density Affymetrix SNP chip and the Illumina Infinium SNP chip can be used to perform whole-genome selection. Two high-resolution Affymetrix custom arrays have been designed for rice, one consisting of ~44,000 SNPs (44 K array) and another consisting of ~1 million SNPs (1 M array) [9**,27]. Both have been used to assay genome-wide patterns of genetic variation in worldwide collections of wild and cultivated rice accessions. Recently, we developed another high-resolution SNP array for rice based on Illumina’s Infinium platform with 51,478 markers, dubbed RiceSNP50 (unpublished results). SNP probes on this array were selected and designed based on an analysis of more than 10,000,000 SNPs extracted from the sequence data of 801 rice accessions. These SNPs were both preferentially located in genes and evenly distributed across the genome. Thus, it is feasible for the rice community to use this high-resolution SNP array for wide variety of research objectives.

Next-generation sequencing (NGS) technologies
NGS platforms, including Illumina HiSeq2500, ABI 5500xl SOLiD, Roche 454, Ion Torrent and PacBio RS, have been rapidly developed as of late. These platforms make whole-genome sequencing accessible to regular laboratories, especially those seeking to re- sequence species for which there are complete reference genome sequenced in existence, such as rice. With advances in NGS technologies, the traditional two-step paradigm of SNP discovery and subsequent assay has been simplified into a single process, in which bioinformatics tools simultaneously analyze the sequence data for both SNP discovery and genotyping [5**,6,7**,8,28,29*].

Identification of genetic variations controlling rice agronomic traits
For designing ideal rice with superior genotypes, the prerequisite is to understand the genetic basis of agronomically important traits and the allelic variation at those loci. With the development of genomics platforms, more and more loci involved in agronomic traits have been mapped, and their allelic variations have been assessed.

The traditional method for gene isolation or marker–trait association analysis in rice is QTL mapping (also known as linkage mapping), which has been used extensively to identify natural mutations of agriculturally important traits. NGS has been used to construct a genetic map for 150 rice recombinant inbred lines (RILs), which was 35 times more precise in recombination breakpoint determination [28]. Further, 49 QTL with phenotypic effect ranging from 3.2 to 46.0% for 14 agronomic traits were detected in these RILs, which indicated that NGS could provide a powerful solution to map QTL with high resolution [30]. On the basis of low-coverage sequences of RILs, a high-density linkage map was constructed using high-quality SNPs in RILs without genotype data of the parental lines [29*]. The new SNP map detected more QTL with precise map locations, showing advantages in detecting power and resolution compared to the RFLP/SSR map [31].
Association mapping (also known as linkage disequilibrium mapping) is increasingly being adopted as the preferred rice mapping method. With genomics-based platforms in use, it is now feasible to perform genome-wide association studies (GWAS) in natural populations of rice. Bin Han and his colleagues performed a GWAS using 517 China landraces that had been selected from ~50,000 rice accessions. Their analysis of 14 agronomic traits localized 6 previously cloned genes to regions less than 26 kb [5**]. Furthermore, they extended the use of this methodology to 950 worldwide collected rice varieties and performed an additional GWAS that identified 32 new loci associated with flowering time and ten grain-related traits [7**]. Another study generating genome sequences from 446 geographically diverse accessions of the wild rice species *Oryza rufipogon* and 1083 cultivated *Indica* and *Sativa* varieties constructed a comprehensive map of rice genome variation and identified 55 selective sweeps that have occurred during domestication [6]. Also, another group re-sequenced 50 accessions of cultivated and wild rice and identified thousands of genes showing significantly lower diversity in cultivated rice than wild rice. These regions thus may have been selected during domestication [8*]. Recently, an Affymetrix array containing 44,000 SNPs has been used to genotype a diversity panel consisting of 413 *O. sativa* accessions, which were selected to represent the genetic diversity of domesticated rice. Aluminum tolerance and 34 other complex traits were systematically analyzed via GWAS [27,32]. On the basis of these successful cases, GWAS was considered to be even more successful in plants than in humans [33].

T-DNA and EMS-induced mutants are also important sources for gene cloning. For T-DNA mutants, the insertion position can be directly identified via Tail-PCR. In contrast, the traditional method of identifying the mutation position for EMS-induced mutants has generally been to generate a cross population with another variety before mapping. Recently, a new method of gene cloning based on whole-genome resequencing of pooled DNA has been developed, called MutMap. In MutMap, a mutant is crossed into its original wild type lines, and the DNA of F2 progenies with phenotypic differences is pooled and sequenced. This method is therefore distinct from the conventional gene isolation strategies that involve crosses between genetically distant lines, and thus avoid F2 Gaussian distribution [34**]. This breakthrough will dramatically increase the efficiency of gene identification.

With more tools and resources available to genetic investigations, an increasing number of rice genes related to important traits have been identified and characterized [35]. By September 2012, 886 genes had been cloned. The chromosomal and functional category distribution for these cloned genes is shown here in Figure 1. These genes, together with QTL and their allelic variations, provide the basis for molecular breeding by design. With accumulated genes information, genic and functional markers can be developed to avoid the recombination between markers and genes for the target traits during marker assisted selection (MAS). A total of 8575 SNPs within these 886 genes regions (from upstream 5 kb to downstream 5 kb) have been included in our second version high density SNP array RiceSNP90 (unpublished result).

Genomics-assisted molecular breeding by design

Since the concept of ‘Breeding by Design’ was proposed, different opinions on molecular breeding by design were brought up. Nevertheless, three major steps were usually considered to be important [36–39]. The first step in breeding by design is to identify genes/QTL or chromosome segments that affect target traits, and to understand the gene effects and their interactions. The second step is to formulate an ideal genotype to meet the breeding objectives. The third step is to develop the most efficient breeding strategies and technologies for identifying the expected genotype. On the basis of our experience and recent publications, we summarized the key steps in molecular breeding by design in Figure 2, emphasizing the important roles of genomics-based genotyping platforms in genes/QTL identification and progeny screen. Through these processes, it is feasible to develop an ideal variety with a desired genome haplotype. Computer simulation has been used and will play more important roles in plant breeding for optimization of breeding process to develop superior variety [40].

With the rapid development of different throughput genotyping platforms and the recent advancements in gene and QTL identification, genomics-assisted molecular breeding by design will become a common practice in the near future. Theoretically, we can design genotypes at the genome-wide level, and then put all the target genes together to breed a superior variety with aid of genomics-based genotyping platforms. Although we are still a long way off from understanding the functions of all the genes and interactions responsible to agronomic traits, it is time to attempt molecular breeding by design through practicing genomics-based genotyping and field phenotyping.

Germplasm characterization and parental selection

The selection for right parental lines is a key step in all breeding programs. For specific trait improvement breeders tend to use varieties with similar genetic background so that homozygous progenies can be obtained in early generations. In contrast, breeders prefer to use varieties with dramatically different genetic background to develop a novel variety with multiple new traits, so that progenies have more possibility to integrate different
allelic variations. With the aid of genomics-based genotyping platforms, breeders can easily identify the genetic background to their expectation \([23,27,41]\). Furthermore, high density genetic markers can also be used in combinability prediction to obtain superior hybrid vigor in rice \([42]\).

**Genomics-assisted trait integration**

Breeding process is a gradual improvement of overall performance in a variety. Most of the time, breeders just want to integrate a few better traits into a popular cultivar to remove its bottleneck. In these cases, backcrossing breeding is a commonly used approach. If breeders want to integrate multiple traits, approaches such as recurrent selection and pyramiding need to be used. By screening for trait-associated molecular markers, plants with desirable traits can be identified in early stages of breeding programs, which can significantly improve the breeding efficiency. This is especially useful when selecting for those traits that are difficult to examine under normal growing conditions.
The commonly practised marker-assisted backcrossing (MABC) is now evolving into a genomics-assisted activity. Besides the marker-assisted selection for target traits, high density SNP arrays are used to remove the adverse genetic drags through minimizing the length of chromosome fragment of donor parent. This process greatly accelerates the recovery of the recipient parent genome during backcrossing [43]. For example, foreground and background selection were integrated to select lines with target quality traits (xa13, Xa21, awaxy and fertility restorer genes), maximum recovery of Basmati rice genome, and minimum non-targeted donor chromosomes [44]. With the recently developed high resolution SNP array RiceSNP50, we can do MABC with a much better resolution (unpublished result).

In a biparental population, both parents can contribute favorable alleles, and the ideal individual should contain mosaic chromosome fragments. Marker-assisted recurrent selection (MARS) uses markers at each generation to target all important traits. Although with known QTL is helpful, MARS can start without QTL information, and marker–trait association can be established during the MARS. Simulation studies revealed that MARS was more efficient than phenotypic selection [45]. This idea can be extended to more than two parents.

Pyramiding is another breeding strategy for taking genes or QTL from different parents and stacking them in one progeny. Marker-assisted gene pyramiding has been successfully applied to disease resistance improvement against multiple pathogens, for example, double resistance to blast and bacterial blight diseases [46]. In another report, Bph14 and Bph15 were recently stacked in Minghui 63 to enhance its resistance to brown plant hoper [47]. For complex trait improvement, such as yield, combination of marker-assisted and phenotype selection (MAPS) has been tested in order to pyramid multiple QTL. Eight QTL controlling 1000-grain weight and spikelet number per panicle were pyramided to breed a new elite variety.
simulation analysis in this study demonstrated that the MAPS strategy allowed for pyramiding 24 QTL in a single hybridization derived progeny [48**].

Genomic selection (GS)
Although MAS has been applied successfully in the breeding of many crops, it has limitations such as that the mapped beneficial allelic genes cannot improve the target trait efficiently or even bring deleterious effects to other traits, due to the different genome backgrounds. Thus a new strategy named as genomic selection (GS) was proposed as the development of genome-wide genotyping technologies [49]. GS refers to marker-based selection without identifying the association between markers and the traits, but uses all data from the markers and traits simultaneously to estimate the accurate effects of all markers.

GS usually consists of a ‘training population’ and a ‘candidate population’. The genomic estimated breeding values (GEBVs) were calculated from the genotyped and phenotyped data of the ‘training population’ and then test on ‘candidate population’. To calculate GEBVs, the central challenge is to estimate all marker predictor effects, p, with available observations, u, when p >> u. To overcome this problem, a variety of methods, for example, best linear unbiased prediction [50], ridge regression [51], Bayesian regression [49], kernel regression [52] and machine learning methods [53], have been proposed to develop prediction models for GS. To estimate accuracy of different statistical models, biostatistics researchers used simulated populations. Meuwissen et al. [49] and Habier et al. [54] applied ridge regression and BayesB on about 50 QTL in an ideal population size of 1000, and the accuracies were 0.66 and 0.64 for ridge regression and 0.79 and 0.69 for BayesB. Then Zhong et al. [55] took 1040 actual markers’ data from 42 barleys and simulated them into a population size of 500, the accuracies were 0.62 and 0.61 for ridge regression and BayesB. Recently, development of genome-wide genotyping assays promoted GS from simulating data to applying actual markers’ data in bull, maize, fruit fly and barley. VanRaden et al. [56] genotyped with 38,416 SNPs on the training population contained over 3500 Holstein bulls. The accuracies of ridge regression and BayesB were similar, from 0.44 to 0.79 for traits ranging in heritability from 0.04 to 0.50. Riedelheimer et al. [57**] crossed a 285*2 hybrid maize population and predicted 7 biomass-related and bioenergy-related traits using 56,110 SNPs and 130 metabolites, the accuracies of BayesB ranged from 0.72 to 0.81 for SNPs and from 0.60 to 0.80 for metabolites. These GS samples applied SNP microarrays on individuals and confirm that statistical models will improve breeding trait predictions efficiently. On the other hand, the researches for methodology of GS are just beginning. Jia et al. [58] proposed that statistical models combining multiple traits would include more information thus offer better accuracy in GS practices.

With the experience of GS in other crops, we believe that GS will be used in rice breeding soon.

Obstacles on the way to breeding by design
Breeding by design is an ideal approach to the development of new varieties, but currently its full potential can hardly be realized because of our limited knowledge in the molecular and biochemical networks that control trait development and the effect of gene–environment interactions. With the complete genome sequence annotated, we know almost all the rice gene structures, but the modes of action for most of them are unclear. We know little about gene interactions and their role in trait development. Rich allelic variations existing in rice germplasm need to be further explored although a large number of accessions have been re-sequenced. Besides the need for high throughput genotyping platforms, breeders also long for accurate phenotyping tools in addition to the measurement of environmental effect on the plant growth and reproduction. More practically, the cost of high throughput genotyping via SNP array or NGS is still beyond the affordable range to most breeders. A joint effort from the whole rice research and breeding community is encouraged to overcome all these obstacles before breeding by design can be widely practiced.

Conclusions
Two genomics-based genotyping platforms, SNP array and NGS, are likely to play key roles in the future practice of molecular breeding by design. Both low-resolution and high-resolution SNP arrays will remain valuable in the years to come. Low-resolution SNP arrays allow breeders to perform foreground selections on target genes. Complementarily, medium or high-resolution SNP arrays will assist breeders in genetic background analysis of breeding parents and background selection of improved elite lines. The advantage of NGS in genotyping is the capability in detection of unknown DNA variations, including rare alleles that are not usually included in SNP arrays. As compared with NGS, SNP arrays can better meet breeders’ need for quick identification of expected genotypes by screening large number of samples in shorter periods of time and producing data that is easier to analyze, though the data may not be as informative as those from NGS. In contrast, NGS can help breeders explore new genetic variations, but requires more effort on data analysis. In the future, both SNP array and NGS platforms will most likely be used in combination to pave the way for breeders to perform molecular breeding by design. As the obstacles on the way are overcome one by one, ‘breeding by design’ in rice and other crops, will be practised more and benefit our human beings.

Acknowledgements
We appreciate the assistance of Abigail Coplin and Alex Roth in manuscript editing. This work was supported by grants from the National Program on Key Basic Research Project of China (973 Program: 2011CB100010), the
References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:
• of special interest
○ of outstanding interest


These two studies showed that the low-density and medium-density SNP assays were quite flexible for different research goals and breeding objectives. They are especially suitable to survey the genetic diversity of large scale germplasm collection and genotype breeding material in early generations.


See Ref. [23].


The authors reported a method for constructing ultrahigh-density linkage maps composed of SNPs based on low-coverage sequencing of recombinant inbred lines. This strategy dramatically increased the efficiency of QTL mapping, which is generally applicable in other mapping studies.


The authors have developed a mapping strategy called MutMap. In MutMap, a mutant is crossed directly to the original wild-type line and then selfed, allowing unequivocal segregation in second filial generation (F2) progeny of subtle phenotypic differences. This study provides an approach that is particularly applicable to crop species because it minimizes the number of genetic crosses and mutant F2 progenies that are required. Also, it avoids the disadvantage of conventional crossing between genetically distinct lines, in which segregation of particular phenotypes in F2 follows a Gaussian as opposed to a discrete distribution. The strategy MutMap will dramatically facilitate the efficiency of gene isolation and thus promote molecular breeding by design.


This study pyramided eight grain yield related QTL and obtained new lines showing increased panicle and spikelet size as compared with the parent variety. The authors further proposed a novel pyramid breeding scheme, allowing pyramiding of as many as 24 QTL at a single hybridization without massive cross work. This study is a good example showing how genomics-based platforms may contribute to rice molecular breeding by design.


The authors applied genomic selection on a hybrid population in maize. To predict combining abilities of the inbred lines for 7 traits of the hybrids, SNP and metabolite data were fitted into models respectively, and both of them provide high accuracy. This strategy has potential to screen large population of parents and predict the superior hybrids.


The authors simulated a barley population and discussed the statistical models for multiple traits genome selection. This model significantly improved prediction accuracies for low heritability traits after taking advantage of correlated high heritability traits information. And multiple trait genomic selection required less phenotypes data. This multivariate model provided a good attempt for linking different traits dynamically.