Breeding Science 60: 524–535 (2010)
doi:10.1270/jsbbs.60.524

Review

Development of genome-wide SNP assays for rice

Susan R. McCouch*[1)], Keyan Zhao[2,3)], Mark Wright[1,2)], Chih-Wei Tung[1)], Kaworu Ebana[4)], Michael Thomson[5)], Andy Reynolds[2)], Diane Wang[1)], Genevieve DeClerck[1)], Md. Liakat Ali[6)], Anna McClung[7)], Georgia Eizenga[7)] and Carlos Bustamante[2,3)]

[1)] *Department of Plant Breeding and Genetics, Cornell University*, Ithaca, NY, USA
[2)] *Department of Biological Statistics and Computational Biology, Cornell University*, Ithaca, NY, USA
[3)] *Department of Genetics, Stanford University*, Stanford, CA, USA
[4)] *National Institute of Agrobiological Sciences*, Tsukuba, Ibaraki 305-8602, Japan
[5)] *International Rice Research Institute*, Los Baños, Laguna, Philippines
[6)] *Rice Research and Extension Center, University of Arkansas*, Stuttgart, AR, USA
[7)] *USDA ARS, Dale Bumpers National Rice Research Center*, Stuttgart, AR, USA

Single nucleotide polymorphisms (SNPs) are the most abundant form of genetic variation in eukaryotic genomes. SNPs may be functionally responsible for specific traits or phenotypes, or they may be informative for tracing the evolutionary history of a species or the pedigree of a variety. As genetic markers, SNPs are rapidly replacing simple sequence repeats (SSRs) because they are more abundant, stable, amenable to automation, efficient, and increasingly cost-effective. The integration of high throughput SNP genotyping capability promises to accelerate genetic gain in a breeding program, but also imposes a series of economic, organizational and technical hurdles. To begin to address these challenges, SNP-based resources are being developed and made publicly available for broad application in rice research. These resources include large SNP datasets, tools for identifying informative SNPs for targeted applications, and a suite of custom-designed SNP assays for use in marker-assisted and genomic selection, association and QTL mapping, positional cloning, pedigree analysis, variety identification and seed purity testing. SNP resources also make it possible for breeders to more efficiently evaluate and utilize the wealth of natural variation that exists in both wild and cultivated germplasm with the aim of improving the productivity and sustainability of agriculture.

Key Words: single nucleotide polymorphism (SNP), rice (*Oryza sativa* L.), genotyping assay, next-generation sequencing, genetic variation, germplasm diversity, plant improvement.

SNPs in the context of plant improvement

The introduction of new sequencing technologies has dramatically changed the landscape for detecting and monitoring genome-wide polymorphism (Craig *et al.* 2008, Huang *et al.* 2009, Metzker 2005, Schuster 2008). Today, single nucleotide polymorphisms (SNPs) are rapidly replacing simple sequence repeats (SSRs) as the DNA marker of choice for applications in plant breeding and genetics because they are more abundant, stable, amenable to automation, efficient, and increasingly cost-effective (Duran *et al.* 2009, Edwards and Batley 2010, Rafalski 2002). SNPs are the most abundant form of genetic variation in eukaryotic genomes and they occur in both coding and non-coding regions of nuclear and plastid DNA (Kwok *et al.* 1996). As genetic markers, they represent sites in the genome where DNA sequence differs by a single base when two or more individuals are compared. They may be individually responsible for specific traits or phenotypes, or may represent neutral variation that is useful for evaluating diversity in the context of evolution.

SNPs are widely used in breeding programs for marker-assisted and genomic selection, association and QTL mapping, positional cloning, haplotype and pedigree analysis, seed purity testing and variety identification (Bernardo 2008, Eathington *et al.* 2007, Jannink *et al.* 2010, Lorenz *et al.* 2010, Moose and Mumm 2008). They allow breeders to assess the range of alleles available to them in diverse germplasm resources and to monitor the combinations of alleles that perform well in target environments (Collard and Mackill 2008, Heffner *et al.* 2009, Jannink *et al.* 2010, Kim *et al.* 2010, Moose and Mumm 2008, Xu *et al.* 2008). Using genome-wide SNP data, a breeder can examine linkage

Communicated by M. Yano
Received September 5, 2010. Accepted October 23, 2010.
*Corresponding author (e-mail: srm4@cornell.edu)

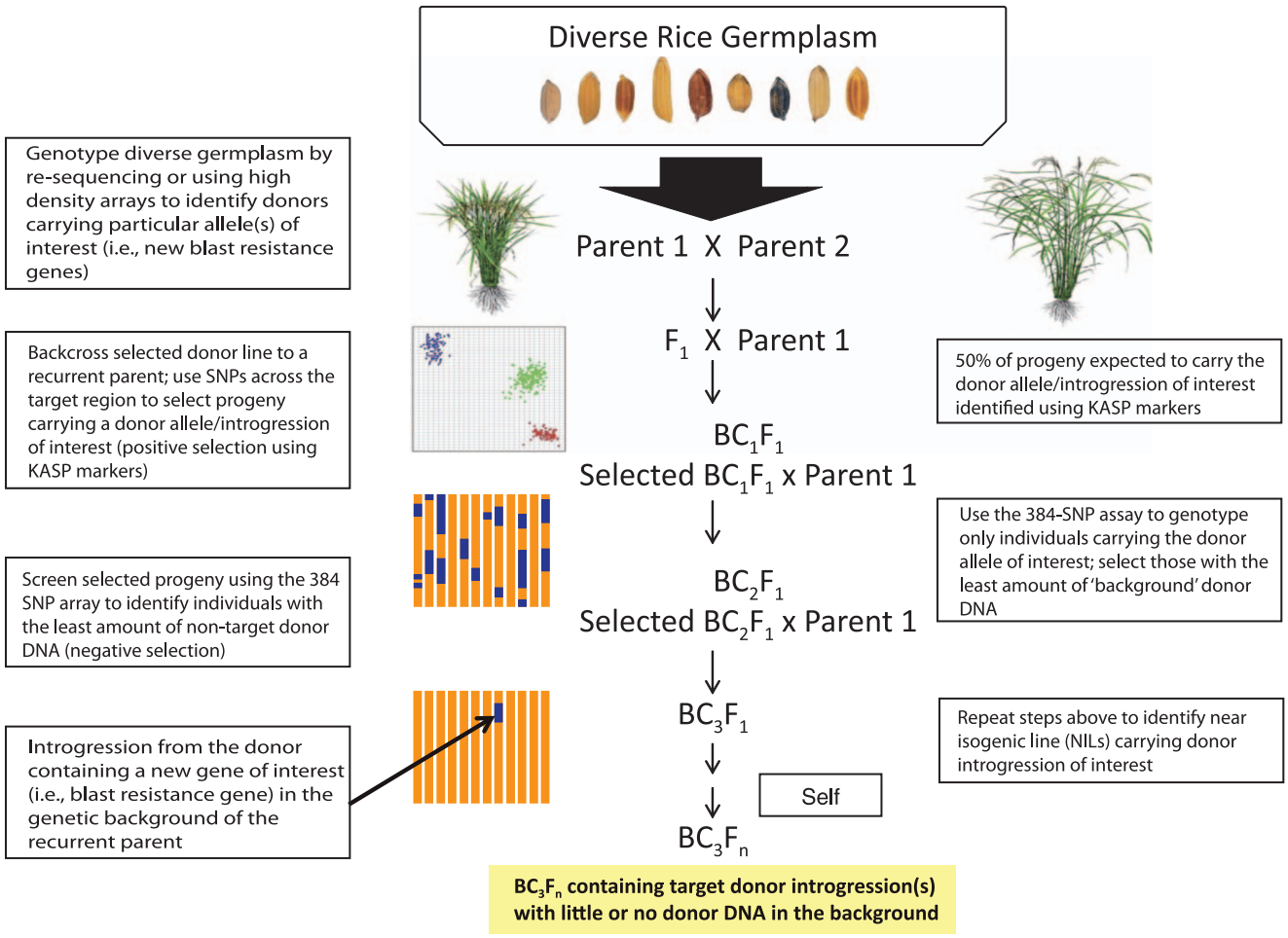

Fig. 1. Diagram showing a marker-assisted backcross-breeding scheme that utilizes information derived from SNP-discovery and SNP-detection platforms (illustrated in Fig. 2) to increase the efficiency of selection.

relationships among alleles along a chromosome or across the genome as a whole. Markers can be used to identify favorable recombinants that break linkage drag, develop near-isogenic lines, pyramid useful alleles, or capture positive transgressive variation through backcross breeding (Fig. 1) (Ashikari and Matsuoka 2006, Fukuoka *et al.* 2009, 2010, Tanksley and McCouch 1997, Yara *et al.* 2010). SNPs are also extensively used to determine population substructure (Ebana *et al.* 2010, Flint-Garcia *et al.* 2005, Hamblin *et al.* 2010, Hyten *et al.* 2007, Nordborg and Weigel 2008, Zhao *et al.* 2010), understand the history of domestication (Kovach *et al.* 2009, Li *et al.* 2010, Shomura *et al.* 2008, Sweeney *et al.* 2007, Takano-Kai *et al.* 2009), and to explore the ancestry of specific alleles, regions of chromosomes or populations (He *et al.* 2006, Yamamoto *et al.* 2010).

The integration of SNP-based selection along with other selection techniques in a breeding program requires significant changes in the way the program is organized. Handling large amounts of genotypic data in addition to all the phenotypic data, pollination schedules, seed management and decision making adds complexity to a breeder's job and requires considerable investment in technical and informatics infrastructure and support (Eathington *et al.* 2007, Lorenz *et al.* 2010, Moose and Mumm 2008, Pop and Salzberg 2008). That support must enable a breeder to obtain reliable marker information on an appropriate number of samples in real time, and the marker-supported genetic gain and improvements in breeding efficiency must compensate for the cost of the SNP genotyping (Heffner *et al.* 2010). Thus, while the integration of SNP genotyping capability promises to accelerate genetic gain and the breeding cycle, it also imposes a series of economic and technical hurdles that represent serious challenges for many public-sector breeding programs, particularly those in the developing world.

To start to address these challenges, SNP-based genotyping and information resources are being developed with public funding and made available for application in rice improvement (Table 1). These resources include diversity datasets emerging from large SNP-discovery and SNP-detection efforts in a broad array of germplasm, custom-designed SNP-genotyping assays for use in genomic and marker-assisted selection, and tools for identifying subsets of SNPs that allow researchers to construct their own SNP assays for targeted applications. The availability of commercial

Table 1. SNP-detection assays for rice

Description	# SNPs	Platform	Utility	Reference
O. sativa Diversity Array	1M	Affymetrix	Evaluate diversity within and between *O. sativa*, *O. rufipogon/O. nivara*, *O. glaberrima*, *O. barthii*	McCouch, per. Comm.
O. sativa Diversity Array	44,100	Affymetrix	Evaluate diversity within and between sub-populations of *O. sativa*	Tung *et al.* (2010)
Global Diversity Primer Set	4,357	Sanger sequencing	Evaluate diversity at 1,578 genes distributed throughout the genome	Ebana *et al.* (2010)
O. sativa Diversity OPA	1,536	Illumina GoldenGate	Captures variation within & between sub-populations of *O. sativa*	Zhao *et al.* (2010)
Japanese Core Set	768	Illumina GoldenGate	Captures variation within Japanese *temperate japonica* cultivars	Yamamoto *et al.* (2010) Nagasaki *et al.* (2010)
O. sativa_indica × *japonica*	384	Illumina BeadXpress	Evaluates variation between *indica* and *japonica*	Thomson *et al.* (2011)
O. sativa_indica/aus	384	Illumina BeadXpress	Evaluates variation within *indica* and *aus*	Thomson *et al.* (2011)
O. sativa_tropical japonica	384	Illumina BeadXpress	Evaluates variation within US *tropical japonica*	Thomson *et al.* (2011)
O. sativa_indica × *O. rufipogon*	384	Illumina BeadXpress	Evaluates variation between *indica* and *O. rufipogon*	Thomson *et al.* (2011)
O. sativa_japonica × *O. rufipogon*	384	Illumina BeadXpress	Evaluates variation between *japonica* and *O. rufipogon*	Thomson *et al.* (2011)

genotyping centers that offer competitive pricing, guaranteed quality and rapid turn-around time is helping to ensure broad access to SNP technology. These centers avoid the requirement for each institution to maintain costly physical infrastructure in-house by offering opportunities for researchers to send tissue or DNA samples, select from a menu of options, and obtain digital genotypes in return. Commercial service providers help bridge the gap between genomics discoveries and breeding applications because they make it their business to stay current with new technology and to actively support the activities of plant breeders from diverse institutions around the world. Using a common set of genotyping assays in a community also facilitates the exchange of information about critical germplasm resources, even when those resources cannot be formally exchanged due to intellectual property or biosafety constraints.

Re-sequencing for SNP discovery

The SNP discovery process is designed to identify SNPs that differ between two or more genomes. For inbreeding species like rice, lines to be re-sequenced are normally purified through one or two generations of inbreeding (via single seed descent) and then high quality DNA is extracted using a Qiagen column or a standard chloroform extraction procedure. The DNA sample is re-sequenced using second or third-generation sequencing technology and SNPs are called by comparing the re-sequenced genome to a reference genome (Metzker 2005, Schuster 2008). As they are discovered, SNPs are characterized in terms of their genome position and frequency in different populations or groups of samples.

Second-generation sequencing platforms produce millions of short-sequence reads, typically 25–400 bp in length, in multiple genomes simultaneously. Third-generation platforms use single-molecule sequencing without the requirement for DNA amplification, and they produce relatively long reads (>1 kb) in real time (Eid *et al.* 2009). In either case, the randomly generated sequences are aligned to a reference genome that has been sequenced to high quality and fully assembled, and SNPs are called in those reads that can be unequivocally aligned to a specific region in the reference genome.

In rice, Nipponbare is the only genome that has been sequenced to high accuracy (IRGSP 2005) and as a result, it is almost universally used as the reference against which other sequences are aligned. Thus, SNPs in rice are generally identified based on their nucleotide position in the Nipponbare reference genome. Alignment to a reference genome like Nipponbare allows researchers to take advantage of the genome annotation to predict whether a SNP falls within or near a gene of interest, and whether a genic SNP is expected to cause a functional change in the protein product (synonymous vs nonsynonymous change) that might alter the expression of the gene (Ondov *et al.* 2008). This can be very useful in determining whether a particular SNP is likely to be responsible for a phenotype of interest. Even when SNPs do not fall within genes, the frequency and distribution of polymorphisms can be used to construct haplotypes and determine ancestry or identify regions of the genome associated with traits of interest (Gupta *et al.* 2001, Rafalski 2002).

History of re-sequencing in rice

The first re-sequencing efforts in rice generated high-resolution shot-gun sequences of both the *japonica* variety, Nipponbare, and the *indica* variety, 9311 (Goff *et al.* 2002, Yu *et al.* 2002). The shot-gun datasets were assembled in

conjunction with the high quality Nipponbare genome that was being sequenced *de novo* clone-by-clone, at high accuracy (IRGSP 2005). By comparing the Nipponbare and 9311 sequences, 1.7 million SNPs were identified by Shen *et al.* (2004) and a set of 384,431 high quality SNPs were identified by Feltus *et al.* (2004) using the same dataset. The rate of SNP detection was reported to be >15x the rate of indel detection in the rice genome (Feltus *et al.* 2004).

A second major re-sequencing effort was undertaken by the OMAP project, where BAC end-sequences from a variety of wild forms of rice were aligned to the Nipponbare sequence and used to construct physical maps of 12 *Oryza* species (Ammiraju *et al.* 2006, 2010, Wing *et al.* 2007). The BAC libraries generated by the OMAP project provided 10–19x coverage of the rice genome, and alignment of the BAC-end sequences allowed SNPs and indels to be called, despite the variable quality of the sequence itself.

The third major re-sequencing effort in rice was performed by Perlegen BioSciences for the *Oryza*SNP project, where 159,879 SNPs were identified based on hybridization-based re-sequencing of 20 diverse *O. sativa* varieties (McNally *et al.* 2009). The quality of the SNPs in this study was very good, but the SNP discovery pool covered only about 100 Mb of the genome and had a low discovery rate due to technical limitations (~11% within the tiled 100 Mb region).

In addition to the genomic-scale re-sequencing approaches described above, there are several examples of Sanger sequencing-based strategies to discover SNPs on a genome-wide basis in rice. These targeted approaches require the design of specific primer pairs, which are generally located in exons and may span intronic as well as exonic regions. Examples of SNP discovery using this approach in rice include Caicedo *et al.* (2007) and Ebana *et al.* (2010).

The early re-sequencing efforts provided excellent starting material for the development of a high resolution SNP map for rice, but additional genome-wide re-sequencing is necessary to enlarge the SNP-discovery pool and more adequately represent the range of natural variation found in rice and its closest ancestors, as well as to identify SNPs within individual gene pools of interest to plant breeders (Tung *et al.* 2010, Yamamoto *et al.* 2010).

Expanding the SNP-discovery pool

To expand the SNP discovery pool for rice, hundreds of diverse *O. sativa*, *O. rufipogon/O. nivara*, *O. glaberrima* and *O. barthii* accessions are currently being re-sequenced by groups in several countries. Early re-sequencing efforts used single-end libraries, but there was a rapid transition to paired end libraries (~350 bp fragments) (Fullwood *et al.* 2009), and sequencing is currently being done using Illumina Genome Analyzer II (GAIIx) technology. Over time, read length has increased from 56 bp to 129 bp reads as Illumina's re-sequencing chemistry has improved, bringing greater efficiencies to the system. The GAIIx platform is well-suited for large-scale SNP discovery because it generates hundreds of millions of short, overlapping sequence-reads that can be used to enhance the confidence of each allele call (Imelfort *et al.* 2009). While the massive data capture associated with next generation sequencing has come at some cost in terms of data quality compared to Sanger sequencing, the loss of quality is generally compensated by the deep coverage. Most widely used re-sequencing efforts provide 5–50x genome coverage and provide the backbone for implementing strategies that provide much lower genome coverage (<0.5x) and which depend on statistical approaches to impute the missing data (Howie *et al.* 2009, Huang *et al*. 2010, Roberts *et al.* 2007, Servin and Stephens 2007, Sun and Kardia 2008).

Development of low, medium and high-resolution SNP assays

While many people in the rice community have the computational infrastructure and statistical expertise needed to generate and analyze the gigabytes of data generated by the re-sequencing of whole genomes, others will find it helpful to be able to access smaller subsets of SNP data that can be analyzed using an MS Excel spreadsheet, or to be able to request targeted SNP detection assays that are tailored for a particular purpose or population. Tailored SNP-detection platforms are designed to monitor variation at a smaller number of SNPs, but can currently do so at a fraction the cost of re-sequencing, particularly when the bioinformatics requirements are taken into account. The two most commonly used SNP-detection platforms include Affymetrix's custom-designed SNP genotyping arrays and Illumina's custom-designed SNP oligonucleotide pools assays (OPAs), though other platforms, such as one developed by KBiosciences offer similar options at lower prices. All require construction of a targeted assay designed to interrogate a subset of previously identified SNPs in an automated, high throughput manner.

Unlike microsatellite loci that are multi-allelic, SNP loci are generally bi-allelic and the individual base change that is detected as a SNP is expected to have occurred only once in evolutionary time. Thus, SNPs are generally informative (polymorphic) only for a particular set of genetic materials. For this reason, each low-resolution SNP assay must be optimized for the population under study. The pronounced subpopulation structure of *O. sativa* (Ebana *et al.* 2010, Garris 2005, Glaszmann 1987, Zhao *et al.* 2010) is apparent at all levels of analysis. Genome-wide SNP frequencies are characteristic of each subpopulation and/or breeding program, and make it possible to trace the ancestry of accessions (Hamblin *et al.* 2010, Yamamoto *et al.* 2010). For this reason, a comprehensive SNP discovery initiative requires thoughtful selection of materials for re-sequencing to ensure that patterns of variation are adequately sampled. If the objective is to identify SNPs that differentiate the major subpopulations, the goal can be easily met by shallow re-sequencing of a couple of representatives from each group,

while if the objective is to identify SNPs that differentiate closely related individuals, such as those found within a single breeding program, deep re-sequencing of elite varieties may be necessary (Yamamoto *et al.* 2010).

Currently, a combination of low-, medium- and high-resolution SNP assays are being developed to address a variety of interests within the rice research community (Table 1). These assays will be publicly available for genotyping of rice samples.

The low-resolution assays consist of 384-SNPs each and have been designed for the GoldenGate assay using Veracode technology on Illumina's BeadXpress Reader (http://www.illumina.com/systems/beadxpress.ilmn). They make use of allele-specific primer extension and require the design of specific primers bordering each SNP. The collection of primers is referred to as an oligonucleotide pool, or an OPA, which is hybridized to genomic DNA followed by allele-specific primer extension and labeling through a universal PCR reaction. The target SNPs are selected following a SNP discovery effort that provides information about the frequency of polymorphism in a particular population or collection of germplasm. These 384-SNP OPAs are attractive to the breeding and genetics communities because they are very reliable and require little technical adjustment once they are designed and optimized. They can rapidly assay hundreds or thousands of individuals within a short time window, and they are relatively inexpensive, given the time, labor and bioinformatics requirements of other marker technologies.

Several 384-SNP Illumina BeadXpress assays have been developed to assay diversity within and between specific subpopulations and species as summarized in Table 1. The SNPs that were included in these oligo pools (OPAs) were selected from larger SNP arrays (the 1536-SNP Illumina assay (Zhao *et al.* 2010) and the 44K-SNP Affymetrix array, discussed below). The conversion rate between assays is very high and the performance of the 384-SNP OPAs has been outstanding, as summarized by Thomson *et al.* (2011).

In another study, a set of 2,688 SNPs were used to genotype 151 Japanese rice cultivars released over the last 150 years (Yamamoto *et al.* 2010). Initially, whole-genome re-sequencing of the elite Japanese cultivar, Koshihikari, was undertaken and the short sequence reads were aligned to the Nipponbare genome. A total of 67,051 SNPs were discovered, and from this discovery pool, a set of 2,688 well-distributed SNPs (every 100–200 kb) were used to develop an Illumina GoldenGate SNP-detection assay. This assay was used to genotype the Japanese rice cultivars, and after eliminating poorly performing markers, 1,917 high quality SNPs were detected. These SNPs were used to develop a combination of 768-SNP OPAs for use on the Illumina GoldenGate BeadArray platform (Yamamoto *et al.* 2010).

More recently, a core set of 768-SNPs was selected to provide even distribution of polymorphisms along the chromosomes of 92 diverse Japanese *temperate japonica* rice accessions (Nagasaki *et al.* 2010) (Table 1). This core set was developed by selecting from a larger set of SNPs that were informative in the same set of materials. SNPs were discovered by re-sequencing three diverse Japanese cultivars (Koshihikari, Eiko and Rikuu132) and aligning the short reads to the Nipponbare reference genome. This effort complemented the first phase of the project in which only Koshihikari was re-sequenced (as outlined above), leaving regions of low SNP density due to the shared ancestry of Koshihikari and Nipponbare.

Using a similar approach, a set of 384 well-distributed SNPs was selected to provide genome-wide coverage of US *tropical japonica* cultivars for use in QTL mapping and pedigree analysis. The first attempt to design this assay was limited by the *Oryza*SNP discovery pool (McNally *et al.* 2009) where large regions of the rice genome appeared to be common by descent in US varieties. The regions that appeared devoid of polymorphism in the *Oryza*SNP dataset (Fig. 2) were found to contain abundant SNPs when US breeding lines were re-sequenced as the basis for SNP discovery.

A 1,536-SNP Illumina GoldenGate assay (Table 1) was designed to detect polymorphism within and between the five major subpopulations of *O. sativa* (Zhao *et al.* 2010). In this case the Perlegen-based re-sequencing of 20 diverse *O. sativa* varieties (McNally *et al.* 2009) provided an adequate SNP discovery pool and approximately 1% of the total Perlegen SNP pool was included in the 1,536-SNP OPA. This assay was used to genotype 395 diverse *O. sativa* accessions, determine their subpopulation structure, identify regions of introgression, and test the feasibility of genome-wide association studies and admixture mapping in this sample population (Zhao *et al.* 2010).

Ebana *et al*. (2010) used Sanger sequencing and 1,578 primer pairs developed from predicted gene sequences positioned roughly every 100–300 kb in cv Nipponbare (RAP-DB, 19 http://rapdb.dna.affrc.go.jp/), to generate ~567 bp of sequence/amplicon in 140 accessions. They observed a total of 4,357 SNPs, averaging 3.21 SNPs per sequenced site, or 4.87 SNPs/kb across the genome. This dataset was used for population structure analysis and to evaluate the diversity of different subpopulations of *O. sativa*. Information about the SNPs and their neighboring sequences can be found at the National Institute of Agrobiological Sciences *Oryza SNP Database 5* (http://oryza-snp.dna.affrc.go.jp/en/index_en.html).

To complement the assays described above, two higher-resolution Affymetrix custom arrays have recently been designed for rice (unpublished data) (Table 1). One consists of ~44,000 SNPs (hereafter the 44K array) and the other consists of ~1 million SNPs (1 M array). Both are being used to assay genome-wide patterns of diversity in world collections of wild and cultivated rice. These higher-resolution SNP detection platforms are more economical per data point than the lower resolution assays, and can rapidly and accurately generate large databases of information about SNP diversity on thousands of lines. The 44 K array provides approximately one SNP every 10 kb, which is expected to be sufficient for genome wide association mapping in rice. The 1 M array

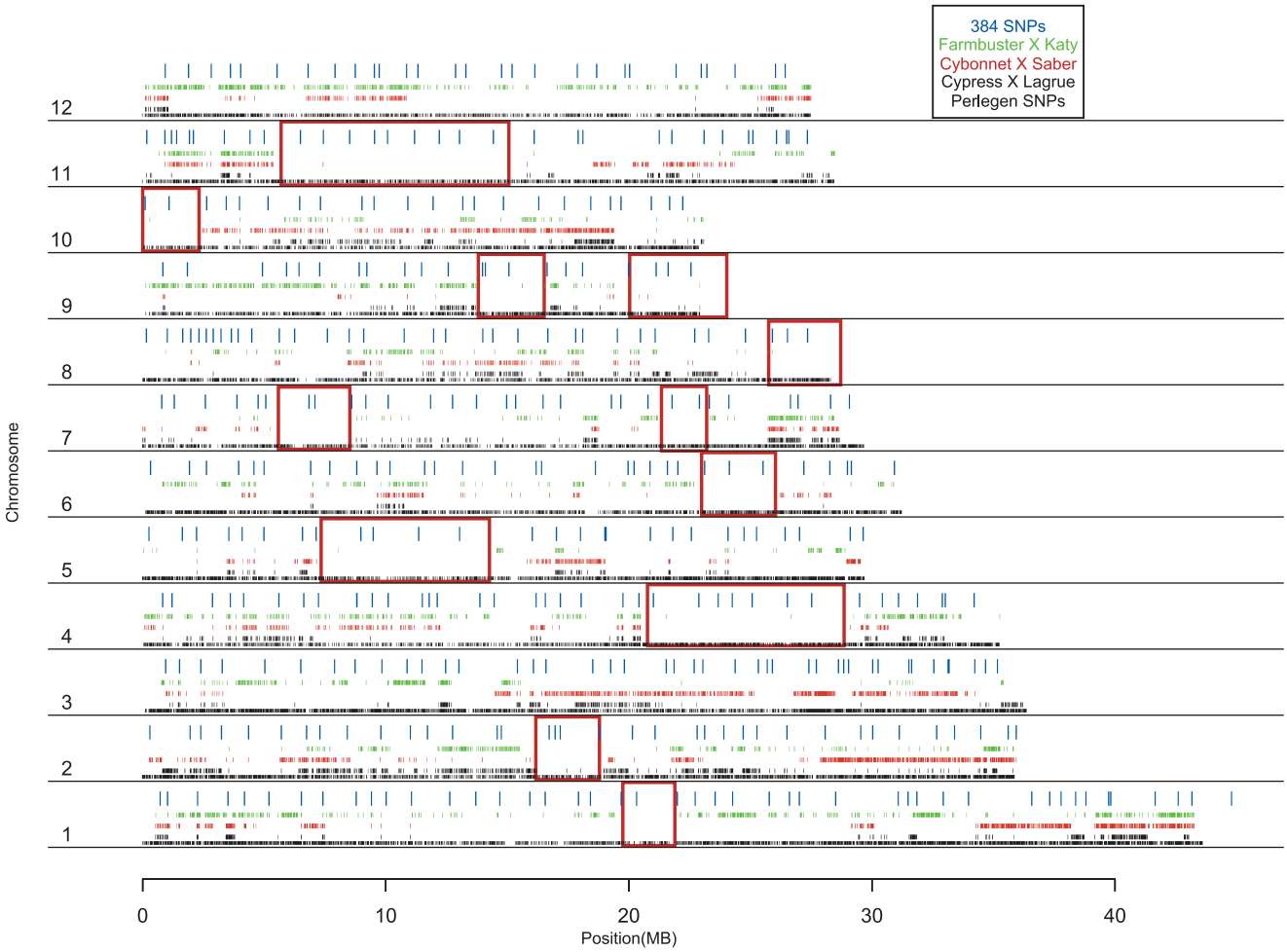

Fig. 2. Distribution of SNPs along the 12 chromosomes of rice. Tick marks along the chromosomes represent SNP locations; first row = Perlegen SNP discovery pool (McNally *et al.* 2009); second row = SNPs from the Perelgen set that were segregating in Cypress × LaGrue; third row = SNPs segregating in Cybonnet × Saber population; fourth row = SNPs segregating in Farmbuster × Katy; fifth row = 384-SNP assay developed for US *tropical japonica* varieties. Rectangles highlight genomic regions that were devoid of SNPs in all three US cross combinations based on the Perelegen SNP discovery pool. The final selection of 384 SNPs for this assay was heavily dependent on the discovery of additional SNPs based on re-sequencing of specific varieties of interest to US breeders.

will provide approximately one SNP every ~500 bp (of single copy DNA), with an expected 1–5 SNPs in every annotated, single copy gene in the rice genome, depending on the size and sequence complexity of the gene. These higher resolution tools allow rice breeders and geneticists to characterize natural variation across most of the functionally important fraction of the rice genome at a cost that invites large-scale analysis in thousands of diverse accessions from the major international rice germplasm repositories.

The International Rice Germplasm Collection (IRGC) at IRRI comprises the largest single species germplasm collection in the world, with more than 102,547 accessions of *O. sativa* (Asian cultivated rice), 1,651 accessions of *O. glaberrima* (African cultivated rice) and 4,508 accessions of wild ancestors (McNally *et al.* 2006). This collection is complemented by extensive rice germplasm collections in Japan (www.gene.affrc.go.jp/databases-core_collections_en.php), China (http://www.cgris.net/cgris_english.html), Taiwan (http://www.npgrc.tari.gov.tw/npgrc1/index_e.html), India (www.nbpgr.ernet.in/), Korea (www.genebank.go.kr/) the USA (www.ars-grin.gov/npgs/searchgrin.html) and many other countries. More in-depth characterization of these rice germplasm resources at both the genotypic and phenotypic level will help breeders more effectively discover and deploy valuable alleles and allele combinations to drive improvements in crop performance (Ebana *et al.* 2010, McNally *et al.* 2009, Tanksley and McCouch 1997, Yan *et al.* 2007, Zhao *et al.* 2010). At this time, a consortium of researchers is organizing an international effort to genotype several thousand diverse rice accessions selected to represent the range of variation available in the two domesticated species of rice, *O. sativa* and *O. glaberrima* and their most closely related wild ancestors, *O. rufipgon/O.nivara* and *O. barthii*.

SNP diversity data emerging from this project can be used to focus more intensively on genomic regions of interest or to develop more economical, low-resolution assays.

Proven SNPs from detection arrays provide ready templates for primer design and targeted marker evaluation useful for a diversity of downstream applications.

Genotyping using targeted SNP assays

SNPs may also be detected using targeted genotyping systems that permit users to make decisions on-the-fly about which SNPs to include in an assay. These systems may be of particular interest to breeders and researchers who are interested in analyzing a small number of specific loci in a large number of samples. There are several types of genotyping assays available on the market that will allow this to be done in a high throughput fashion. Here, we describe the KASP (**K**BioScience **A**llele-**S**pecific **P**olymorphism, KBioscience, UK) system as an example to illustrate how SNP and flanking sequence data generated by the high-resolution fixed arrays may be readily used for the design of allele-specific primers.

KASP is a competitive allele-specific PCR genotyping system that allows SNPs to be detected via FRET (**F**luorescence **R**esonance **E**nergy **T**ransfer). The reaction involves three primers: one common primer and two allele-specific primers that differ at their 3′ ends, where the target SNP is located. The design of a single KASP primer set (one common and two allele-specific primers) requires knowledge of not only the SNP, but of the flanking sequences surrounding the SNP. Because the regions flanking target SNPs on the 1 M, 44 K and Illumina 1536 arrays are known to be invariable across a diverse range of germplasm as well as unique within the genome, these fixed array data allow researchers to identify reliable target regions for design of KASP primers and negate the need to sample-sequence in order to find SNPs in material of interest.

Targeted genotyping systems may be used to evaluate large populations of specific loci to 1) identify recombination breakpoints in fine-mapping genes underlying QTLs in bi-parental crosses, 2) detect of functional SNPs within a subset of germplasm, 3) identify accessions carrying introgressed alleles from other subgroups, 4) perform marker-assisted breeding, and 5) retaining target regions in NIL development (Fig. 1).

SNP Quality Analysis and SNP Calling Algorithms

In rice, most samples are inbred and likely to be homozygous at nearly all loci. This poses a problem for the genotype calling software developed by Illumina (GenomeStudio) and Affymetrix (BRLMM-P) that are based on clustering methods. Both default programs often fail to call SNPs correctly because the heterozygote cluster is missing. As a result, the clustering algorithms typically assume one of the homozygous clusters to be heterozygous, or attempt to split one of the homozygous clusters into a heterozygous and homozygous group to meet the expectation of Hardy Weinberg equilibrium. To address this problem, we developed a statistical method, ALCHEMY, that reliably performs allele calling in inbred samples where one of the genotypic classes is missing (Wright *et al.* 2010). To validate ALCHEMY's accuracy and call rates, we compared ALCHEMY SNP calls to expected SNP genotypes across many re-sequenced rice samples and found a high concordance (average 99.1%, call rate 96.1%). Another advantage compared to existing software is that ALCHEMY works well even if only a few samples are processed. Additionally, the statistical treatment of the problem permits inbreeding to be explicitly considered and incorporated into the model. Simultaneously estimating and optimizing the inbreeding coefficient on a per-sample basis allows both out-bred and inbred samples to be analyzed simultaneously and improves both accuracy and call rates.

We are currently using ALCHEMY for genotype calling in all rice samples using both Illumina and Affymetrix SNP-detection platforms. The software has also shown strong performance on Human HapMap data and consistently high levels of performance on data from several other plant species. ALCHEMY is written in C and developed and used under the GNU/Linux environment. It is available to the public at no charge. Source code is expected to compile and run on any GNU/Linux platform, Mac OS X, and Unix environments with the GNU C compiler and associated tools installed (Wright *et al.* 2010, www.alchemy.sourceforge.net/).

Training and user support for utilizing SNPs in plant improvement

Publicly available rice SNP datasets can be accessed from www.ricediversity.org, from the Gramene database (www.gramene.org/db/diversity) or from the National Institute of Agrobiological Sciences (NIAS) Oryza SNP Database (oryza-snp.dna.affrc.go.jp/en/index_en.html) for further analysis and research. These datasets can be downloaded in several different file formats, including .txt, HapMap, PLINK or as a Flapjack project file. Data can also be queried and selectively downloaded using, for example, the SNP Query tool developed for the Gramene Genetic Diversity module (www.gramene.org/db/diversity/snp_query) or the SNP Data Search option from the NIAS website. These tools enable large datasets to be searched based on genome coordinates and/or cultivar subgroups, making it easy for users to download sections of SNP data in comma-separated-value format for manipulation in MS Excel, GeneFLow (www.geneflowinc.com) or some other analysis package of choice. For population genetic analysis, TASSEL (www.maizegenetics.net/tassel) can be launched through a web-browser within Gramene (www.gramene.org/diversity/tassel_launch.html).

Users who would like to browse data in Flapjack may install the program on their computer (bioinf.scri.ac.uk/flapjack/). Once Flapjack has been installed, datasets can be automatically downloaded from public databases, including the Gramene Genetic Diversity Data Download page (www.gramene.org/diversity/download_data.html) by

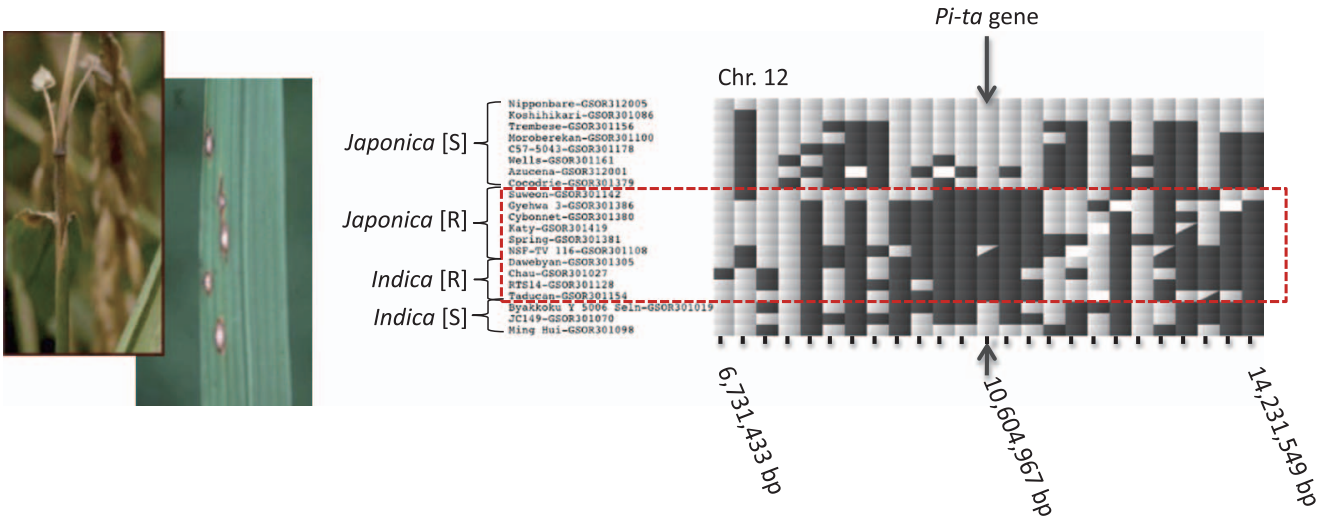

Fig. 3. SNP genotypes across a 7.5 MB region on chromosome 12 carrying the *Pi-ta* gene for resistance to rice blast disease. SNPs positions (in bp) are from the 1,536-SNP assay described by Zhao *et al.* (2010) and are shown as tick marks along the bottom of the graph. Cultivars highlighted with a dashed box carry the functional SNP conferring resistance at *Pi-ta*. The functional nucleotide polymorphism [G/T] in *Pi-ta* is located at 10,604,967 bp; a "G" at this position confers resistance [R], while a "T" confers susceptibility [S] (Bryan *et al.* 2000). The *indica* variety, Taducan (fourth from the bottom) is widely used as a donor of *Pi-ta* resistance and blast resistant *Japonica* cultivars shown here all carry an *indica* introgression around the *Pi-ta* gene (dark grey). In this figure, accession NSF_TV 116 is heterozygous at the functional SNP (half white, half grey); white fill indicates missing data.

clicking on any 'Flapjack' link (note, large datasets are downloadable by chromosome). Clicking on the 'Flapjack' hyperlink listed for the SNP dataset published by Zhao *et al.* (2010) will load roughly 1300 SNP allele calls for the 395 rice accessions assayed in this experiment. Flapjack will, by default, display the alleles with four colors corresponding to the four nucleotides. It is sometimes easier to see interesting trends by viewing the data using only two colors. Using these features, a Flapjack snapshot was created to illustrate how SNP data can be used to identify rice accessions carrying a rice blast disease resistance allele at the *Pi-ta* locus, as featured in Fig. 3.

To complement the technical and analytical aspects of the project, the Rice SNP Consortium is planning a series of training workshops and is developing a set of educational materials to demonstrate how SNPs can be managed and used to increase the efficiency of plant breeding. The outreach materials will be used to run workshops, with hands-on exercises to learn how to make use of the SNP chips and other information generated by this project. Workshop participants will have an opportunity to manipulate SNP data, select subsets of informative SNPs for marker-assisted selection, analyze seed purity and investigate patterns of SNP diversity in different gene pools.

Advantages and disadvantages of fixed SNP assays

While the cost of second and third generation re-sequencing approaches is rapidly coming down, the data management and data analysis requirements of the new technology command increasingly sophisticated computational infrastructure and bioinformatics expertise (Edwards and Batley 2010, Imelfort *et al.* 2009, Pop and Salzberg 2008). The newer, faster, cheaper sequencing platforms and strategies generate data that is initially riddled with sequencing errors (Craig *et al.* 2008) and while deeper genome coverage helps to minimize the discovery of false SNPs, it engorges the data pipelines with quantities of low-quality data that must be screened, sorted and ultimately, discarded. Diversity data generated today using high quality SNP-detection platforms (both high and low resolution) will retain value for years to come, despite changes in the technology used to identify the SNPs. Fixed arrays will be just as informative as resequencing approaches would be where the information content of the dataset is limited by the frequency of recombination rather than by the detection of rare alleles. Furthermore, array-based analysis can be readily followed up by more targeted, sequencing-based approaches to identify rare alleles in genomic regions or germplasm accessions of interest (Mitsui *et al.* 2010)

With rapid changes in technology and computational resources, the two-step paradigm of SNP-discovery followed by SNP-detection is rapidly merging into a single, sequence-based process of simultaneous SNP discovery and detection (Craig *et al.* 2008, Cronn *et al.* 2008, Huang *et al.* 2009, 2010). Once this becomes economically and logistically feasible for the rice community, direct sequencing will undoubtedly replace the use of fixed arrays. The most important advantage of direct re-sequencing is that it makes no assumptions about the SNP variation that may occur in a genome; it readily detects both frequent and rare SNPs and may detect novel SNP variation at a site that is not known to be variable.

While re-sequencing is a very powerful strategy for

characterizing genetic variation at the DNA level, it is not without its own limitations (Ku *et al.* 2010). Re-sequencing approaches generally depend on alignment of short sequence reads to a reference genome. Thus, SNPs falling in repetitive regions or in regions that are highly diverged from the reference genome do not align well to the reference genome and are often excluded from the pool of discovery SNPs. In the case of rice, novel genes, gene duplications, transposable elements, chromosomal rearrangements, or other structural variants larger than a few base pairs that distinguish an accession or group of accessions from Nipponbare, fall into this category. For example, the *Sub1* locus controlling submergence tolerance consists of a tandemly duplicated array of genes corresponding to *Sub1A*, *Sub1B* and *Sub1C* that are found in wild and cultivated populations (Xu *et al.* 2006). Sequence analysis indicates that the rise of the *Sub1A* gene that is functionally responsible for the varying degrees of submergence tolerance is due to a duplication that occurred only in the *aus/indica* lineage during rice domestication (Fukao *et al.* 2009). Re-sequencing and alignment to Nipponbare would have failed to detect the uniqueness of this gene because the functional copy is missing from the reference genome. Similarly, the *Snorkel1* and *Snorkel 2* genes that enable a plant to grow ahead of rising flood waters are not present in Nipponbare (Hattori *et al.* 2009). For this reason, SNPs corresponding to these genes would not be detected using any of the re-sequencing approaches in common use today. Nonetheless, SNPs in some of these variable regions can be recovered and identified, along with their flanking sequence, particularly if they are shared by several re-sequenced genomes. New re-sequencing technologies that expand the size of sequenced libraries and/or allow for *de novo* assembly of short sequence reads derived from variable length clones (such as bacterial artificial chromosomes, BACs) will help position SNPs along a genome, even those derived from highly divergent regions of the genome.

In rice, the *Oryza* Alignment Project (OMAP) is addressing the limitations of Nipponbare as a single reference genome by constructing deep-coverage large-insert bacterial artificial chromosome (BAC) libraries from 12 *Oryza* species, end-sequencing the clones, and constructing physical maps of all 12 genomes (Ammiraju *et al.* 2006, 2010, Wing *et al.* 2007). These physical maps will enable researchers to construct new reference genomes for rice and its wild relatives by combining the use of current re-sequencing technology with a BAC-pool approach. The existence of physical maps for the 12 divergent *Oryza* genomes is an essential tool for characterizing the kinds of structural variation that distinguish these genomes from Nipponbare (Ammiraju *et al.* 2006, 2010, Wing *et al.* 2007). This comparative genomics approach will greatly facilitate the identification of genes and non-coding DNA that are unique to individual species or subpopulations and may be useful targets for future plant breeding efforts.

The Rice SNP Consortium

To address the need for an expanded genomic toolkit and data resource for exploring and utilizing natural variation in rice, an international group of researchers from diverse public and private institutions in more than 10 countries has contributed expertise, germplasm, DNA, sequence information and financial support to the Rice SNP Consortium (www.ricesnp.org).

The goals of the Rice SNP Consortium are a) to develop a large SNP-discovery pool and data resource, b) to design, develop and distribute a suite of low-, medium- and high-resolution SNP genotyping platforms and analysis tools, c) to genotype a large collection of diverse rice germplasm accessions, and d) to develop a publically accessible database containing information on genomic diversity in wild and cultivated rice. The consortium is funded by several large competitive grants and many smaller contributions from public and private sector institutions (see list of contributions to the re-sequencing effort at www.ricesnp.org).

Members of the consortium contribute in different ways to this effort, providing purified germplasm samples, DNA, sequence information (in the form of fastq files from in-house efforts), computational expertise and/or financial support. The combined efforts of consortium members continue to expand the number, quality and diversity of rice samples being re-sequenced. With its open architecture, the Rice SNP Consortium welcomes any group or individual to contribute to the SNP-discovery pool. Expanding the repertoire of sequenced genomes will help ensure that the data has the broadest potential for applications in rice research worldwide. All genomes sequenced under this initiative will be made publicly available through the Amazon Cloud, and information about how to access the data will be posted on the consortium's website (www.ricesnp.org).

The long-term goal of the consortium is to unlock the wealth of natural variation that has remained largely untapped in both wild and cultivated genetic resources, and to constructively utilize it to improve the productivity and sustainability of rice agriculture. Understanding the evolutionary history of rice is expected to generate new insights about the synthesis of novel alleles that may expand the repertoire of useful traits in the future. In a broader context, the ability to link sequence and diversity information to physiological functions, plant development and agronomic traits in rice will greatly expand the foundation for comparative genomics using rice as a pivotal reference genome. As a result, this project will lay the groundwork for related applications in other major crop species.

Acknowledgements

We acknowledge financial support from contributors to the Rice SNP Consortium (see www.ricesnp.org), the NSF (PRGP award 0606461), and the USDA (AFRI award 2009-65400-05698). We are grateful to the USDA-NRI

*Oryza*SNP project (PIs: Jan Leach, Hei Leung, Robin Buell) for providing early access to the Perlegen Sciences re-sequencing data (McNally *et al.* 2009) that facilitated the design of the 1536-SNP assay (Zhao *et al.* 2010), to Masahiro Yano for sharing Koshihikari re-sequencing data ahead of publication and to Affymetrix for hosting workshops at the Plant and Animal Genome Meetings in San Diego and the Rice Genetics Symposium in Manila during 2008, 2009, 2010 and 2011 that promoted critical discussion among members of the international rice research community and helped coordinate the development of SNP detection platforms for rice.

Literature Cited

Ammiraju, J.S.S., X. Song, M. Luo, N. Sisneros, A. Angelova, D. Kudrna, H.R. Kim, Y. Yu, J.L. Goicoechea, M. Lorieux *et al*. (2010) The *Oryza* BAC resource: a genus-wide and genome scale tool for exploring rice genome evolution and leveraging useful genetic diversity from wild relatives. Breed. Sci. 60: 536–543.

Ammiraju, J.S., M. Luo, J.L. Goicoechea, W. Wang, D. Kudrna, C. Mueller, J. Talag, H. Kim, N.B. Sisneros, B. Blackmon *et al.* (2006) The *Oryza* Bacterial Artificial Chromosome library resource: Construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*. Genome Res. 16: 140–147.

Ashikari, M. and M. Matsuoka (2006) Identification, isolation and pyramiding of quantitative trait loci for rice breeding. Trends Plant Sci. 11: 344–350.

Bernardo, R. (2008) Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Sci. 48: 1649–1664.

Bryan, G.T., K.-S. Wu, L. Farrall, Y. Jia, H.P. Hershey, S.A. McAdams, K.N. Faulk, G.K. Donaldson, R. Tarchini and B. Valent (2000) A single amino acid difference distinguishes resistant and susceptible alleles of the rice blast resistance gene *Pi-ta*. Plant Cell 12: 2033–2045.

Caicedo, A.L., S.H. Williamson, R.D. Hernandez, A. Boyko, A. Fledel-Alon, T.L. York, N.R. Polato, K.M. Olsen, R. Nielsen, S.R. McCouch *et al.* (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet. 3: e163.

Collard, B.C.Y. and D.J. Mackill (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Phil. Trans. R. Soc. B 363: 557–572.

Craig, D.W., J.V. Pearson, S. Szelinger, A. Sekar, M. Redman, J.J. Corneveaux, T. Laub, G. Nunn, D.A. Stephan, N. Homer *et al.* (2008) Identification of genetic variants using bar-coded multiplexed sequencing. Nat. Methods 5: 887–893.

Cronn, R., A. Liston, M. Parks, D.S. Gernandt, R. Shen and T. Mockler (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing—by-synthesis technology. Nucleic Acids Res. 36: e122.

Duran, C., N. Appleby, T. Clark, D. Wood, M. Imelfort, J. Batley and D. Edwards (2009) AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. Nucleic Acids Res. 37: D951–953.

Eathington, S.R., T.M. Crosbie, M.D. Edwards, R.S. Reiter and J.K. Bull (2007) Molecular markers in a commercial breeding program. Crop Sci. 47: S154–S163.

Ebana, K., J.-I. Yonemaru, S. Fukuoka, H. Iwata, H. Kanamori, N. Namiki, H. Nagasaki and M. Yano (2010) Genetic structure revealed by a whole-genome single-nucleotide polymorphism survey of diverse accessions of cultivated Asian rice (*Oryza sativa* L.). Breed. Sci. 60: 390–397.

Edwards, D. and J. Batley (2010) Plant genome sequencing: applications for plant improvement. Plant Biotech. J. 8: 2–9.

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. Science 323: 133–138.

Feltus, A., J. Wan, S.R. Schulze, J.C. Estill, N. Jiang and A.H. Paterson (2004) An SNP resource for rice genetics and breeding based on subspecies *indica* and *japonica* genome alignments. Genome Res. 14: 1812–1819.

Flint-Garcia, S.A., A.-C. Thuillet, J. Yu, G. Pressoir, S.M. Romero, S.E. Mitchell, J. Doebley, S. Kresovich, M.M. Goodman and E.D. Buckler (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J. 44: 1054–1064.

Fukao, T., T. Harris and J. Bailey-Serres (2009) Evolutionary analysis of the *Sub1* gene cluster that confers submergence tolerance to domesticated rice. Annal. Bot. 103: 143–150.

Fukuoka, S., N. Saka, H. Koga, K. Ono, T. Shimizu, K. Ebana, N. Hayashi, A. Takahashi, H. Hirochika, K. Okuno *et al.* (2009) Loss of function of a proline-containing protein confers durable disease resistance in rice. Science 325: 998–1001.

Fukuoka, S., K. Ebana, T. Yamamoto and M. Yano (2010) Integration of genomics into breeding in rice. Rice 3: 131–137.

Fullwood, M.J., C.L. Wei and Y. Ruan (2009) Next-generation DNA sequencing for paired-end tags for transcriptome and genome analyses. Genome Res. 19: 521–532.

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

Glaszmann, J.C. (1987) Isozymes and classification of Asian rice varieties. Theor. Appl. Genet. 74: 21–30.

Goff, S.A., D. Ricke, T.-H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296: 92–100.

Gupta, P.K., J.K. Roy and M. Prasad (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Curr. Sci. 80: 524–535.

Hamblin, M.T., T.J. Close, P.R. Bhat, S. Chao, J.G. King, K.J. Abraham, T. Blake, W.S. Brooks, B. Cooper, C.A. Griffey *et al.* (2010) Population structure and linkage disequilibrium in US barley germplasm: Implications for association mapping. Crop Sci. 50: 556–566.

Hattori, Y., K. Nagai, S. Furukawa, X.-J. Song, R. Kawanok, H. Sakakibara, J. Wu, T. Matsumoto, A. Yoshimura, H. Kitano *et al.* (2009) The ethylene response factors *SNORKEL1* and *SNORKEL2* allow rice to adapt to deep water. Nature 460: 1026–1030.

He, G., X. Luo, F. Tian, K. Li, Z. Zhu, W. Su, X. Qian, Y. Fu, X. Wang, C. Sun *et al*. (2006) Haplotype variation in structure and expression of a gene cluster associated with a quantitative trait locus for improved yield in rice. Genome Res. 16: 618–626.

Heffner, E.L., M.E. Sorrells and J.-L. Jannink (2009) Genome selection for crop improvement. Crop Sci. 49: 1–12.

Heffner, E.L., A.J. Lorenz, J.-L. Jannink and M.E. Sorrells (2010) Plant breeding with genomic selection: gain per unit time and cost. Crop Sci. 50: 1681–1690.

Howie, B.N., P. Donnelly and J. Marchini (2009) A flexible and accurate genotype imputation method for the next generation of

genome-wide association studies. PLoS Genet 5: e1000529.

Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, C. Li, C. Zhu, T. Lu, Z. Zhang *et al*. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Gen. 42: 961-967.

Huang, X., Q. Feng, Q. Qian, Q. Zhao, L. Wang, A. Wang, J. Guan, D. Fan, Q. Weng, T. Huang *et al.* (2009) High-throughput genotyping by whole-genome resequencing. Genome Res. 19: 1068–1076.

Hyten, D.L., I.-Y. Choi, Q. Song, R.C. Shoemaker, R.L. Nelson, J.M. Costa, J.E. Specht and P.B. Cregan (2007) Highly variable patterns of linkage disequilibrium in multiple soybean populations. Genetics 175: 1937–1944.

Imelfort, M., C. Duran, J. Batley and D. Edwards (2009) Discovering genetic polymorphisms in next-generation sequencing data. Plant Biotech. J. 7: 312–317.

IRGSP (International Rice Genome Sequencing Project) (2005) The map-based sequence of the rice genome. Nature 436: 793–800.

Jannink, J., A.J. Lorenz and H. Iwata (2010) Genomic selection in plant breeding: from theory to practice. Brief. Funct, Genomics 9: 166–177.

Kim, J.S., S.-G. Ahn, C.-K. Kim and C.-K. Shim (2010) Screening of rice blast resistance genes from aromatic rice germplasms with SNP markers. Plant Pathol. J. 26: 70–79.

Kovach, M.J., M.N. Calingacion, M.A. Fitzgerald and S.R. McCouch (2009) The origins and evolution of fragrance in rice (*Oryza sativa* L.). Proc. Natl. Acad. Sci. USA 106: 14444–14449.

Ku, C.-S., E.Y. Loy, A. Salim, Y. Pawitan and K.S. Chia (2010) The discovery of human genetic variations and their use as disease markers: past, present and future. J. Hum. Genet. 55: 403–415.

Kwok, P.Y., Q. Deng, H. Zakeri, S.L. Taylor and D. Nickerson (1996) Increasing the information content of STS-based genome maps: identifying polymorphisms in mapped STSs. Genomics 31: 123–126.

Li, Y.H., W. Li, C. Zhang, L. Yang, R-Z. Chang, B.S. Gaut and L.-J. Qiu (2010) Genetic diversity in domesticated soybean (*Glycine max*) and its wild progrenitor (*Glycine soja*) for simple sequence repeat and single-nucleotide polymorphism loci. New Phytol. 242–253.

Lorenz, A., S. Cho, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, K.P. Smith, M.E. Sorrells and J.-L. Jannink (2010) Genomic selection in plant breeding: knowledge and prospects. Adv. Agron. (in press).

McNally, K.L., R. Bruskiewich, D. Mackill, C.R. Buell, J.E. Leach and H. Leung (2006) Sequencing multiple and diverse rice varieties. Connecting whole genome variation with phenotypes. Plant Physiol. 141: 26–33.

McNally, K.L., K.L. Childs, R. Bohert, R.M. Davidson, K. Zhao, B.J. Ulat, G.G. Zeller, R.M. Clark, D.R. Hoen, T.E. Bureau *et al.* (2009) Genome-wide SNP variation reveals relationships among landraces and modern varieties of rice. Proc. Natl. Acad. Sci. USA 106: 12273–12278.

Metzker, M.L. (2005) Emerging technologies in DNA sequencing. Genome Res. 15: 1767–1776.

Mitsui, J., Y. Fukuda, K. Azuma, H. Tozaki, H. Ishiura, Y. Takahashi and S. Tsuji (2010) Multiplexed resequencing analysis to identify rare variants in pooled DNA with barcode indexing using next-generation sequencer. J. Hum. Genet. 55: 448–455.

Moose, S.P. and R.H. Mumm (2008) Molecular plant breeding as the foundation for 21st century crop improvement. Plant Physiol. 147: 969–977.

Nagasaki, H., K. Ebana, T. Shibaya, J. Yonemaru and M. Yano (2010) Core single-nucleotide polymorphisms—a tool for genetic analysis of the Japanese rice population. Breed. Sci. 60: 648–655.

Nordborg, M. and D. Weigel (2008) Next-generation genetics in plants. Nature 456: 720–723.

Ondov, B.D., A. Varadarajan, K.D. Passalacqua and N.H. Bergman (2008) Efficient mapping of applied biosystems SOLiD sequence data to a reference genome for functional genomic applications. Bioinformatics 24: 2776–2777.

Pop, M. and S.L. Salzberg (2008) Bioinformatics challenges of new sequencing technology. Trends Genet. 24: 142–149.

Rafalski, A. (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr. Opin. Plant Biol. 5: 94–100.

Roberts, A., L. McMillan, W. Wang, J. Parker, I. Rusyn and D. Threadgill (2007) Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. Bioinformatics 23: i401–407.

Schuster, S.C. (2008) Next-generation sequencing transforms today's biology. Nat. Methods 5: 16–18.

Servin, B. and M. Stephens (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet. 3: e114.

Shen, Y.-J., H. Jiang, J.-P. Jin, Z.-B. Zhang, B. Xi, Y.-Y. He, G. Wang, C. Wang, L. Qian, X. Li *et al.* (2004) Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiol. 135: 1198–1205.

Shomura, A., T. Izawa, K. Ebana, T. Ebitani, H. Kanegae, S. Konishi and M. Yano (2008) Deletion in a gene associated with grain size increased yields during rice domestication. Nat. Genetics 40: 1023–1028.

Sun, Y.V. and S.L. Kardia (2008) Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks. Eur. J. Hum. Genet. 16: 487–495.

Sweeney, M., M.J. Thomson, Y.G. Cho, Y.J. Park, S.H. Williamson, C.D. Bustamante and S.R. McCouch (2007) Global dissemination of a single mutation conferring white pericarp in rice. PLoS Genetics 3: e133.

Takano-Kai, N., H. Jiang, T. Kubo, M. Sweeny, T. Matsumoto, H. Kanamori, B. Padhukasahasram, C. Bustamante, A. Yoshimura, K. Doi *et al.* (2009) Evolutionary history of *GS3*, a gene conferring grain length in rice. Genetics 182: 1323–1334.

Tanksley, S.D. and S.R. McCouch (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. Science 277: 1063–1066.

Thomson, M.J., K. Zhao, M. Wright, K.L. McNally, J. Rey, C-W. Tung, A. Reynolds, B. Scheffler, G. Eizenga, A. McClung *et al.* (2011) High-throughput SNP genotyping in rice using the GoldenGate assay on the BeadXpress platform. Mol. Breed. (submitted).

Tung, C.-W., K. Zhao, M. Wright, L. Ali, J. Jung, J. Kimball, W. Tyagi, M. Thomson, K. McNally, H. Leung *et al.* (2010) Development of a research platform for dissecting phenotye-genotype associations in rice (*Oryza* spp.). RICE (in press).

Wing, R., H. Kim, J. Foicoechea, Y. Yu, D. Kudrna, A. Zuccolo, J. Ammiraju, M. Luo, W. Nelson and J. Ma (2007) The Oryza map alignment project (OMAP): a new resource for comparative genome studies within *Oryza*. *In*: Upadhyaya, N.M. (ed.) Rice Functional Genomics, Springer, New York, pp. 395–409.

Wright, M., C.W. Tung, K. Zhao, A. Reynolds, S.R. McCouch and C.D. Bustamante (2010) ALCHEMY: A reliable method for automated SNP genotype calling for small batch sizes and highly homozygous populations. Bioinformatics doi: 10.1093/bioinformatics/btq533.

Xu, K., X. Xu, T. Fukao, P. Canlas, R. Maghirang-Rodriguez, S. Heuer, A.M. Ismail, J. Bailey-Serres, P.C. Ronald and D.J. Mackill (2006) *Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature 442: 705–708.

Xu, Y. and J.H. Crouch (2008) Marker-assisted selection in plant breeding: from publications to practice. Crop Sci. 48: 391–407.

Xu, Y., Y. Lu, J. Yan, R. Babu, Z. Hao, S. Gao, S. Zhang, J. Li, B. Vivek, C. Magorokosho *et al.* (2009) SNP chip-based genomewide scans for germplasm evaluation, marker-trait association analysis and development of a molecular breeding platform in maize. Proceedings of the 14th Australasian Plant Breeding Conference (APBC) & 11th Congress of the Society for the Advancement of Breeding Research in Asia and Oceania (SABRAO) 10–14 August 2009, Cairns Convention Centre, Cairns, Tropical North Queensland, Australia.

Yamamoto, T., H. Nagasaki, J. Yonemaru, K. Ebana, M. Nakajima, T. Shibaya and M. Yano (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.

Yan, W.G., J.N. Rutger, R.J. Bryant, R.G. Fjellstrom, M.H. Chen, T.H. Tai and A. McClung (2007) Development and evaluation of a core subset of the USDA rice germplasm collection. Crop Sci. 47: 869–876.

Yara, A., C.N. Phi, M. Matsumura, A. Yoshimura and H. Yasui (2010) Development of near-isogenic lines for *BPH25*(t) and *BPH26*(t), which confer resistance to the brown planthopper, *Nilaparvata lugens* (Stål.) in *indica* rice 'ADR52'. Breed. Sci. 60: 639–647.

Yu, J., S. Hu, J. Wang, G.K.-S. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296: 79–92.

Zhao, K., M. Wright, K.J. Kimball, G. Eizenga, A. McClung, M. Kovach, W. Tyagi, L. Ali, C.-W. Tung, A. Reynolds *et al.* (2010) Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. PLOS One 5: e10780.