Development and application of 96- and 384-plex single nucleotide polymorphism (SNP) marker sets for diversity analysis, mapping and marker-assisted selection in rice

M.J. Thomson,[1] K. Zhao,[2] M. Wright,[2] K.L. McNally,[1] H. Leung[1] and S.R. McCouch[2]*
[1] International Rice Research Institute (IRRI), DAPO Box 7777, Metro Manila, Philippines; [2] Cornell University, Ithaca, NY 14853, USA.

Abstract
Marker-assisted selection can enable more precise and rapid breeding strategies, but limitations in genotyping techniques have prevented markers from being integrated into mainstream breeding programs. Multiplexed single nucleotide polymorphism (SNP) markers have the potential to increase the speed, efficiency and cost-effectiveness of genotyping, provided that an optimal SNP density is used for each application. To test the usefulness of multiplexed SNP genotyping in rice, we designed four GoldenGate VeraCode oligo pool assay (OPA) sets for the Illumina BeadXpress Reader. Validated markers from existing 1536 Illumina SNPs and 44K Affymetrix SNP chips developed at Cornell University were used to select subsets of SNPs for maximum information and even distribution for each application. A 96-plex OPA was developed for assigning a sample into one of the five *Oryza sativa* population sub-groups. One 384-plex OPA was designed to have evenly spaced polymorphic markers for QTL mapping and background selection for *indica* × *indica* crosses, two additional 384-plex OPAs were selected for use in *indica* × *japonica* crosses, while all of them can be used for genetic diversity analysis and DNA fingerprinting. More research is required to identify relevant SNPs for African rice germplasm, including *O. glaberrima* and NERICA varieties, and wild *Oryza* species to provide cost-effective and robust SNP genotyping assays for breeding programs focused on improving rice for Africa. The availability of optimized SNP sets will help increase the efficiency of genetic mapping and marker-assisted selection to meet the challenges of rice improvement in the future.

Introduction
The demand for rice is growing quickly in Africa and rice production cannot keep up with consumption. To meet future demand, stress-tolerant rice varieties that are adapted to local conditions and produce higher and more stable yields need to be developed. Although plant breeders have made significant progress using conventional breeding methods, the period since about 1990 has seen the potential of molecular markers gain prominence in breeding programs. Marker-assisted selection (MAS) techniques promise to enable greater gains in rice breeding by allowing for precise manipulation of the desired allele combinations. Toward this end, markers are now being widely used to provide information on genetic relatedness, linkage to important traits, and detection of donor introgressions in segregating populations.

While early success was made with mapping genes for qualitative traits, many traits in rice breeding are quantitative, showing a normal distribution, due to control by multiple genes and gene–environment interactions. Mapping quantitative trait loci (QTLs), the chromosomal regions that contain a gene or genes contributing to a trait phenotype, has provided the means to select specific introgressions conferring desirable traits. Once major QTLs are identified, marker-assisted backcrossing (MABC) can be used to quickly transfer these beneficial QTLs into breeding lines using foreground markers at the target locus and background markers across the rest of the genome (Collard and Mackill, 2008). Thus, the new paradigm for non-transgenic molecular breeding follows a pathway from identifying donors for the trait of interest, to QTL discovery and fine-mapping of the most promising loci, and finally using MABC to rapidly transfer the QTLs into popular varieties. The power and efficacy of this approach was proven through the discovery, fine-mapping and cloning the submergence-tolerance QTL *Sub1* from the submergence-tolerant landrace FR13A, followed by successful MABC transfer of *Sub1* into the popular variety Swarna and five other mega-varieties (Neeraja *et al.*, 2007; Septiningsih *et al.*, 2009). However, MAS using current genotyping techniques has limitations in cost and speed, thus slowing uptake by breeding programs.

The successful implementation of MAS strategies is dependent on having an efficient and robust genotyping system in place (Collard *et al.*, 2008). For many years, simple sequence repeat (SSR) markers have been the marker system of choice due to their high polymorphism rates, ease of polymerase chain reaction (PCR), and the ability to run them on the gel-electrophoresis equipment found in most labs. While SSRs have proven useful for diversity analysis, QTL mapping and marker-assisted breeding, they present difficulties in scoring precise allele sizes, running in high-throughput systems, and reducing costs through mulitiplexing. Routine integration of markers into modern breeding programs will require high-throughput genotyping platforms that can handle large numbers of samples at a low cost. Thus, a new generation of markers based on single nucleotide polymorphisms (SNPs) is now rapidly overtaking SSRs because of the new SNP genotyping platforms that offer multiplexed sets of markers for different applications. SNPs have the potential to greatly

[*] Corresponding author (email: m.thomson@cgiar.org).

increase the speed and reduce the cost of molecular-marker genotyping, which will make it feasible to 'mainstream' MAS into conventional breeding programs.

Materials and methods

Re-sequencing and SNP discovery

The completion of the high-quality DNA sequence of the rice genome by the International Rice Genome Sequencing Project was a landmark achievement (Matsumoto *et al.*, 2005). Although it took over 8 years and was a massive undertaking, the fact remains that it was a single rice genome, of the temperate-*japonica* variety Nipponbare, that was sequenced out of the thousands of different rice genomes existing in the diverse germplasm from around the world. Nonetheless, there is tremendous value in having a 'gold standard' reference sequence available, since it can be used to align lower-coverage whole-genome sequence reads, providing a rapid means for SNP discovery. While initial SNP identification made use of comparing the Nipponbare reference with the 93-11 *indica* sequence, a subsequent study used 20 diverse rice accessions to identify 160 000 SNPs across the genome (McNally *et al.*, 2009; www.oryzasnp.org). Likewise, next-generation sequencing techniques, such as the Illumina Genome Analyzer IIx, have been used for rapid re-sequencing of dozens of additional rice varieties, which is generating massive amounts of sequence data that will provide a valuable SNP-discovery pool to the rice community. At the same time, it will be important to have *de novo* high-quality *indica* and Aus genome sequences as well, since there may be unique portions of the genome in those sub-groups that are not present in Nipponbare, that would be hidden from any sequence assembly depending on the Nipponbare sequence, as was seen in a study on the *Pup1* QTL region which had several genes unique to Kasalath (Heuer *et al.*, 2009). However, with sequencing costs decreasing rapidly, it will become increasingly feasible to sequence large numbers of rice accessions, allowing for more rapid and comprehensive SNP discovery.

Selecting sub-sets of SNPs for marker applications

The value of having a large pool of available SNPs is that a sub-set containing the most informative and useful SNPs can be selected for different applications and sets of germplasm. For certain applications, it may be important to select a sub-set of evenly spaced SNP markers that are polymorphic between targeted germplasm groups, while others may use trait-specific functional SNPs that are diagnostic of desirable alleles. A key step in this process is validating SNPs for their performance with specific marker assays, since not every SNP will be able to be converted into a reliable marker across different genotyping systems. Fixed SNP chips, such as those from Affymetrix and Illumina, provide a cost-effective means to genotype a selected sub-set of SNPs across large numbers of accessions. In rice, efforts have been made to validate SNP markers using a 1536-plex Illumina SNP chip and a high-resolution Affymetrix SNP chip with 44 000 SNPs (Susan McCouch, Cornell University). These large genome-wide SNP sets will provide valuable data on marker assay conversion success, SNP frequency and polymorphism rates across different germplasm groups, and ultimately associations between SNP markers and traits of interest. These validated SNP sets enable the selection of more targeted sub-sets of SNPs for running on lower-cost systems aimed at larger numbers of samples on specific germplasm pools.

SNP genotyping using the BeadXpress Reader

The current study selected sub-sets of previously validated SNPs for running as 96- and 384-SNP multiplexed sets. The SNP sets were designed for the Illumina GoldenGate assay, which uses locus and allele-specific oligos with cy3/cy5 labeling to detect SNP alleles at each locus. These custom Oligo Pool Assay (OPA) sets were then run on the Illumina BeadXpress Reader as 96- and 384-plex VeraCode assays. Veracode uses cylinder microbeads with an internal barcode to differentiate bead types which correspond to different SNP loci (384 bead types are used for a 384-plex SNP set), and each microbead is coated with oligos that contain a unique address that hybridizes with the labeled products. During scanning on the BeadXpress Reader, the beads are aligned in a groove plate, and the bead codes and cy3/cy5 signal intensities are measured across replicated sets of beads to assign the SNP alleles. This procedure allows for rapid, high-quality SNP calling of 96 samples by 384 SNPs without requiring fixed arrays. While the GenomeStudio software from Illumina clusters alleles based on the ratio of the cy3/cy5 signal intensities to call the three genotype classes, an improved algorithm, called ALCHEMY, has been developed at Cornell University to provide more accurate allele calls, especially with largely inbred lines (Wright *et al.*, 2010).

The assays require 5 µL of a 50 ng/µL DNA sample, and was successfully tested at the International Rice Research Institute (IRRI) using DNA from rice leaves extracted with Qiagen DNeasy Plant Mini kits or a high-quality chloroform extraction and ethanol precipitation. A 2-day protocol is required to run one plate of 96 DNA samples with either the 96- or 384-plex SNP sets, and two plates can be run through the protocol at the same time by a single researcher. Scanning a 384-plex plate on the BeadXpress Reader takes approximately 4 hours, enabling a maximum throughput of up to 288 samples × 384 SNPs per day (assuming multiple researchers work in tandem to process the plates through the initial protocol).

Results

Four custom SNP sets were developed for use in rice: one 96-plex SNP set was designed to differentiate the five population sub-groups in *Oryza sativa*, one 384-plex OPA was designed to have evenly spaced polymorphic markers for QTL mapping and background selection for *indica* × *indica* crosses, while two additional 384-plex OPAs were selected for use in *indica* × *japonica* crosses. All three of the 384-plex SNP sets can also be used for genetic diversity analysis and DNA fingerprinting.

Multiple plates of 96 DNA samples of rice varieties and lines were successfully run on all four SNP sets, validating the OPA designs. A set of 12 F_1 and parental DNAs was also run as control and gave an average 99.7% consistency rate and 94% call rate. These custom BeadXpress SNP sets are being actively used at IRRI for diversity analysis, DNA fingerprinting, QTL mapping and MAS. For the rice varieties and lines tested, the GenomeStudio software had difficulty in clustering the heterozygous genotype class. Although the clusters can be manually edited, this is a time-consuming and subjective process, so in practice we have used the ALCHEMY software at a 0.95 threshold for calling the alleles. The resulting SNP calls were then re-formatted for subsequent data analysis for SNP visualization using Flapjack graphical genotyping software (http://bioinf.scri.ac.uk/flapjack), diversity analysis using PowerMarker (http://statgen.ncsu.edu/powermarker/), and population structure using Structure (http://pritch.bsd.uchicago.edu/structure.html).

Conclusions

SNP discovery efforts have begun to make large pools of SNPs available to the rice community. While high-density SNP chips will be useful for genome-wide association studies, their relative high cost per sample prevents their wider use in breeding programs. In contrast, lower-density SNP assays can provide a more reasonable cost per sample, which enables rapid, low-cost genotyping for a number of different genetics and breeding applications. Having a high-throughput SNP platform with automated genotyping and allele calling can allow more efficient strategies that were not feasible using previous gel-based genotyping systems. The low cost for SNP genotyping (in most cases under US$ 0.10/SNP data point for a 384-SNP assay) opens up this technology for use in a wide range of research programs and applications, as described below.

Diversity analysis

The most straightforward use of these SNP sets is to characterize the relatedness of a set of rice germplasm through a genetic diversity analysis. In this case, a key advantage of SNP markers is that they are bi-allelic: this not only provides more reliable and consistent allele scoring, but it also allows for data to be easily merged from different studies. In contrast, previous SSR studies often had to deal with multiple alleles of different sizes — which gave different results on each of the different genotyping systems. For instance, allele sizes obtained from manual acrylamide gels could not be directly merged with SSR data from fluorescently labeled capillary electrophoresis systems, while SNP data from different genotyping platforms and different groups will be more amenable to combining into a common diversity database. The only caveat is that each SNP set has an inherent bias due to the process of selecting the most informative SNPs for each population group, which can affect the interpretation of the diversity results. For example, a 384-plex SNP set designed to differentiate *indica* and *japonica* accessions will tend to amplify the differences between these two sub-groups, while causing accessions within each group to cluster more tightly. Conversely, the *indica* × *indica* SNP set will allow more detailed definition of sub-groups within *indica* and between *indica* and Aus, while causing *japonica* and aromatic accessions to cluster more tightly. However, in practice this has not been a problem, as 384 SNPs provide ample information compared with previous studies that had used only 30–50 SSR markers.

DNA fingerprinting

Likewise, DNA fingerprinting is a natural application for SNP markers, since robust allele calling and low-cost genotyping can allow unique SNP fingerprints to be assigned to large numbers of breeding lines and released varieties. For example, a 384-plex SNP set can be used to rapidly genotype important lines to develop a SNP fingerprint database that can be subsequently used for varietal identification, quality control, seed tracking, and impact assessment of variety uptake in farmers' fields. By using a common SNP set across groups, these data can also be shared to develop a more comprehensive global SNP fingerprint database.

QTL mapping

Although SSRs have an advantage of high polymorphism rates across any germplasm, SNPs need to be more carefully selected on the basis of their frequency rates within and between different germplasm groups. The 384-plex SNP sets were specifically designed to be evenly spaced across the genome and to provide reasonably high polymorphism rates for their targeted germplasm pools. Thus, the correct SNP set needs to be selected for each population — for example, 384 SNPs in an *indica* × *japonica* SNP set will be largely monomorphic for *indica* × *indica* crosses. Once a suitable polymorphism rate is confirmed (preferably >200 markers out of a 384-

plex SNP set), a QTL study can be completed very quickly due to having all of the markers multiplexed in the same plate: instead of running 96 samples on 100 gels for an SSR map, an equivalent SNP map can be obtained through a single run on the BeadXpress Reader.

Marker-assisted selection

SNPs can be integrated into many different MAS strategies to increase the efficiency of selection. For example, SNP markers can be used for rapid foreground and background selection in an MABC scheme. High-density SNP chips can also provide detailed genotype data for genome-wide selection strategies. However, an even more promising technique will be to identify functional SNPs that are diagnostic of desirable alleles for each trait of interest. Intense research efforts around the world are focused on dissecting the molecular mechanisms and genetic pathways underlying key traits in rice — leveraging the value of the complete genome sequence for gene discovery and functional analysis. One potential outcome of this massive research investment in rice functional genomics is that the end result will allow the identification of functional nucleotide polymorphisms, which are changes in the specific genes that cause the desired phenotype, for development of functional markers (Andersen and Lübberstedt, 2003). By relying directly on the causal polymorphism, these 'perfect' markers will be diagnostic of the favorable allele, since they will always co-segregate with the trait phenotype. As more genes controlling key traits in rice are characterized, there will be more opportunities to mine superior alleles from germplasm collections and to develop functional SNP markers for more precise and efficient MAS.

Future perspectives and SNP resources for African germplasm

While many of the initial SNP genotyping assays were designed for Asian rice germplasm, there are currently efforts for SNP discovery across African germplasm, which will enable the development of SNP markers that are optimized for use in programs breeding improved rice for Africa. For example, reference sequencing of the *O. glaberrima* accession CG14 is in progress (R. Wing, University of Arizona, personal communication). Additional sequencing efforts will identify relevant SNPs across a range of African rice germplasm, including *O. glaberima* and NERICA varieties, and wild *Oryza* species, which can then be used to provide cost-effective and robust SNP genotyping assays for African germplasm. As these genotyping tools are made more accessible to rice breeders, the full potential of marker-assisted breeding in mainstream breeding programs will finally be realized.

Acknowledgements

We are grateful to Ma. Ymber Reveche and Jessica Rey for assistance with the BeadXpress Reader at IRRI, and to Chih-Wei Tung for providing control DNA from Cornell University. This research was supported in part by grants to IRRI by the German Federal Ministry of Development (BMZ/GTZ) and the Government of Japan, and by grants from the US National Science Foundation to Susan McCouch and Carlos Bustamante at Cornell University.

References

Andersen J and Lübberstedt T. 2003. Functional markers in plants. *Trends in Plant Science* 8: 554–560.

Collard B and Mackill D. 2008. Marker-assisted selection: An approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B* 363: 557–572.

Collard B, Vera Cruz C, McNally K, Virk P and Mackill D. 2008. Rice molecular breeding laboratories in the genomics era: Current status and future considerations. *International Journal of Plant Genomics* 2008, Article ID 524847, doi:10.1155/2008/524847.

Heuer S, Lu X, Chin J-H, Tanaka JP, Kanamori H, Matsumoto T, De Leon T, Ulat VJ, Ismail AM and Wissuwa M. 2009. Comparative sequence analysis of the major quantitative trait locus phosphorus uptake 1 (*Pup1*) reveal a complex genetic structure. *Plant Biotechnology Journal* 7: 456–471.

Matsumoto T, Wu JZ, Kanamori H *et al*. 2005. The map-based sequence of the rice genome. *Nature* 436: 793–800.

McNally K, Childs K, Bohnert R, Davidson R, Zhao K, Ulat V, Zeller G, Clark R, Hoen D, Bureau T, Stokowski R, Ballinger D, Frazer K, Cox D, Padhukasahasram B, Bustamante C, Weigel D, Mackill D, Bruskiewich R, Rätsch G, Buell C, Leung H and Leach J. 2009. Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences USA* 106: 12 273–12 278.

Neeraja C, Maghirang-Rodriguez R, Pamplona A, Heuer S, Collard B, Septiningsih E, Vergara G, Sanchez D, Xu K, Ismail A and Mackill D. 2007. A marker-assisted backcross approach for developing submergence-tolerant rice cultivars. *Theoretical and Applied Genetics* 115: 767–776.

Septiningsih E, Pamplona A, Sanchez D, Maghirang-Rodriguez R, Neeraja C, Vergara G, Heuer S, Ismail A and Mackill D. 2009. Development of submergence-tolerant rice cultivars: The *Sub1* locus and beyond. *Annals of Botany* 103: 151–160.

Wright MH, Tung CW, Zhao K, Reynolds A, McCouch SR and Bustamante CD. 2010. ALCHEMY: A reliable method for automated SNP genotype calling for small batch sizes and highly homozygous populations. *Bioinformatics* 26: 2952–2960.