Plant Breed. Biotech. 2014 (September) 2(3):195~212
http://dx.doi.org/10.9787/PBB.2014.2.3.195
Online ISSN: 2287-9366
Print ISSN: 2287-9358

REVIEW ARTICLE

High-Throughput SNP Genotyping to Accelerate Crop Improvement

Michael J. Thomson*

Plant Breeding, Genetics and Biotechnology Division, International Rice Research Institute, DAPO Box 7777, Metro Manila 1301, Philippines

ABSTRACT Recent advances in next-generation sequencing (NGS) and single nucleotide polymorphism (SNP) genotyping promise to greatly accelerate crop improvement if properly deployed. High-throughput SNP genotyping offers a number of advantages over previous marker systems, including an abundance of markers, rapid processing of large populations, a variety of genotyping systems to meet different needs, and straightforward allele calling and database storage due to the bi-allelic nature of SNP markers. NGS technologies have enabled rapid whole genome sequencing, providing extensive SNP discovery pools to select informative markers for different sets of germplasm. Highly multiplexed fixed array platforms have enabled powerful approaches such as genome-wide association studies. On the other hand, routine deployment of trait-specific SNP markers requires flexible, low-cost systems for genotyping smaller numbers of SNPs across large breeding populations, using platforms such as Fluidigm's Dynamic Arrays™, Douglas Scientific's Array Tape™, and LGC's automated systems for running KASP™ markers. At the same time, genotyping by sequencing (GBS) is rapidly becoming popular for low-cost high-density genome-wide scans through multiplexed sequencing. This review will discuss the range of options available to modern breeders for integrating SNP markers into their programs, whether by outsourcing to service providers or setting up in-house genotyping facilities, and will provide an example of SNP deployment for rice research and breeding as demonstrated by the Genotyping Services Lab at the International Rice Research Institute.

Keywords Genotyping by sequencing, Molecular breeding, Rice, SNP marker deployment

INTRODUCTION

Over the past few decades, there has been a large investment around the world in basic plant science research, from trait characterization to functional genomics, as well as the infrastructure to collect, store, and characterize the genetic resources of important crop species. Although some of these efforts are embarked upon purely for scientific discovery, the underlying justification of many of these initiatives is that the advances in genomics and germplasm collections will prove essential to make future gains in crop improvement to feed a growing world (Delmer 2005; McCouch *et al.* 2012; Alfred *et al.* 2014; Varshney *et al.* 2014). Towards this end, approaches employing molecular markers are now being pursued to translate these discoveries into tangible products to accelerate progress in plant breeding. There has been a gap, however, between the discovery of useful genes and QTLs and their deployment in breeding programs—to date few examples show the successful release of marker-assisted selection (MAS) products having a significant impact in farmers' fields. But this is about to change—the field is now poised for a rapid acceleration of progress in crop improvement, made possible by the large-scale deployment of new technologies for next-generation sequencing (NGS) and single-nucleotide polymorphism (SNP) genotyping (McCouch *et al.* 2010; Davey *et al.* 2011; Feuillet *et al.* 2011; Morrell *et al.* 2012; Poland and Rife 2012; Chen *et al.* 2013b; Huang *et al.* 2013).

The foundation for this opportunity is based on two main developments: the accumulated knowledge of useful genetic diversity, genes, and QTLs, and the technical

Received September 19, 2014; Revised September 22, 2014; Accepted September 23, 2014; Published September 30, 2014
***Corresponding author** Michael J. Thomson, m.thomson@irri.org, Tel: +82-63-2-580-5600, Fax: +82-63-2-580-5699

advances in sequencing, genotyping and bioinformatics that have enabled rapid, high-throughput molecular marker approaches. Since the shift to simple sequence repeat (SSR) markers around 15 years ago and subsequently to SNP markers, excellent progress has been made to characterize the genetic diversity of major crop species, to map QTLs for key traits, and to clone genes important for crop improvement. For major crop species such as rice, maize and wheat, there are a large number of fine-mapped and cloned genes with associated functional markers that now provide breeders with a molecular marker toolkit for transferring traits into new varieties (Lübberstedt *et al.* 2005; Van Damme *et al.* 2011; Miura *et al.* 2011; Liu *et al.* 2012). This provides an opportunity for breeders to use a targeted approach to select and combine beneficial alleles at known major genes controlling traits of interest. Likewise, advances in NGS have provided an in depth view into SNP haplotypes at key regions of identity by descent (IBD)—allowing these important chromosome segments to be tracked with greater precision (Chia *et al.* 2012). At the same time, ultra high-throughput DNA extraction and SNP genotyping techniques have brought marker deployment to the point of being able to handle tens of thousands of samples per season corresponding to the scale of lines evaluated by field breeders. In addition to targeted methods, genome-wide selection approaches based on genomic estimated breeding values (GEBVs) now promise greater gains of selection through the use of high-density genome-wide data, rather than being limited to a few known loci (Heffner *et al.* 2009; Bernardo 2010; Jannink *et al.* 2010). When combined with professional sample tracking and quality control (QC) measures, these resources now provide an unprecedented opportunity to accelerate gains of selection and push yield improvement past the plateaus plant breeders have recently encountered. The advantages of NGS and SNP genotyping have quickly been embraced by large multinational seed companies to boost their breeding programs—it's time that public breeding programs start making similar investments.

This review will discuss the advantages that have led to the recent shift to SNP genotyping, show how NGS has provided valuable SNP discovery data sets, review the various SNP genotyping platforms including fixed arrays, flexible low-cost approaches, and genotyping by sequencing (GBS), and finally describe the key issues to consider when integrating SNP markers into a breeding program. While these techniques are relevant to all crop species, the examples provided will focus on rice research and breeding, with a case study of the high-throughput SNP genotyping facility at the International Rice Research Institute (IRRI).

Advantages of SNP genotyping

To understand the shift to single nucleotide polymorphism (SNP) markers, we must first look into the limitations of SSR markers. First, there are limited numbers of SSR motifs in the genome—which becomes a constraint when trying to saturate a region with markers or when trying to identify gene-based markers. In addition, one of the main advantages of SSRs—their high information content from multiple alleles per locus—also presents difficulties when merging SSR data from different platforms and curating allele sizes in databases. In addition, gel-based SSRs are labor intensive and automated fragment sizing systems have limited scope for multiplexing. Therefore, SSR genotyping quickly hits a point where the low throughput and higher cost becomes a limiting factor—which is in contrast to recent SNP genotyping techniques.

The main advantages of SNP markers relate to their ease of data management along with their flexibility, speed, and cost-effectiveness. Bi-allelic SNP markers are straightforward to merge data across groups and create large databases of marker information, since there are only two alleles per locus and different genotyping platforms will provide the same allele calls once proper data QC has been performed. Although it is important to have a bioinformatics data management and curation team to convert SNP markers from different platforms to be on the same DNA strand, that is less challenging than trying to harmonize SSR allele sizes from different systems. With the help of a high quality reference genome, merging sequence and SNP data also enables more powerful analyses of the complete SNP catalog or "SNP universe" for each species. As the most common type of DNA polymorphism, SNPs are also flexible in the selection of SNP variants at target loci, as well as the large numbers of genome-wide loci available to choose from when selecting sets of informative markers for

specific germplasm pools.

A major factor in the advantages of SNP markers for flexibility, speed and cost-effectiveness is the range of genotyping platforms available to address a variety of needs for different marker densities and costs per sample. Whereas early SNP genotyping techniques relied on gel-based methods such as cleaved amplified polymorphic sequence (CAPS) markers (Thiel *et al.* 2004; Komori and Nitta 2005) or allele-specific amplification methods (Drenkard *et al.* 2000), the expansion of the field has led to large-scale commercial investment by life science companies to develop sophisticated sequencing and genotyping platforms that leverage recent advances in nanotechnology, computer science, and automation. These range from high-marker-throughput technologies, such as highly multiplexed fixed arrays providing over 1 million SNP loci per run, to high-sample-throughput technologies that enable running of hundreds of samples per day with low-cost SNP assays. These genotyping platforms have been highly optimized for speed, efficiency, robustness and cost-effectiveness. Although many of these systems require a large initial capital investment, the end result is that the cost per sample has decreased to the point where it is significantly cheaper to genotype a breeding line than to phenotype it. Another benefit is that many of these approaches are automated or can be easily outsourced to a service provider—which has freed up staff from the monotonous routine of hands-on lab work to more effective functions in analyzing and managing genotype data, integrating phenotype data, and applying new tools for breeding applications.

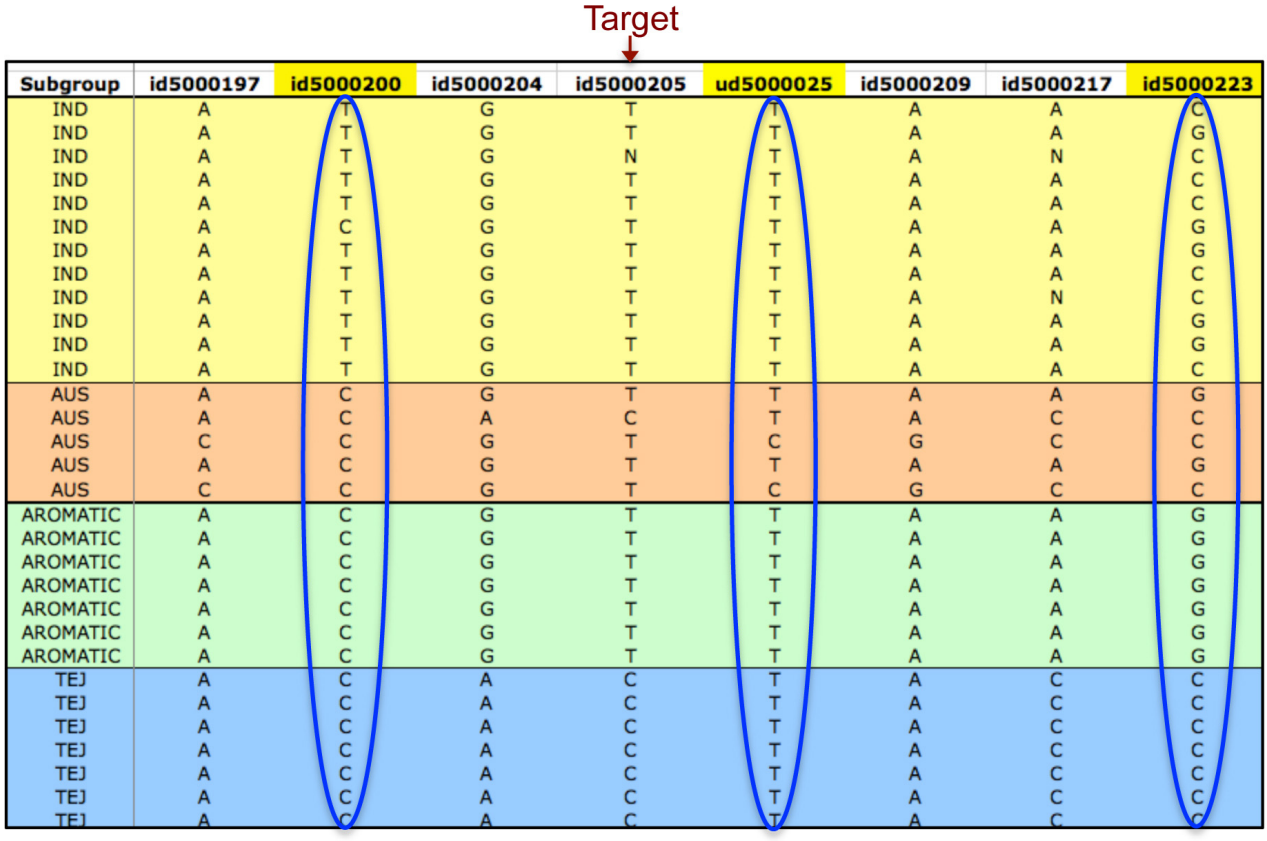
Next generation sequence data and SNP haplotypes

The rise of NGS has led to a flood of sequence data for most agriculturally relevant plant and animal species (Rounsley *et al.* 2009; Feuillet *et al.* 2011; Morrell *et al.* 2012; Edwards *et al.* 2013; Huang *et al.* 2013; Bolger *et al.* 2014; Li *et al.* 2014). Massively parallel short read technologies from Illumina and Ion Torrent have enabled routine re-sequencing of genomes, while long read technologies, such as from Pacific Biosciences, are making it possible to more rapidly develop high quality reference genomes (www.illumina.com; www.lifetechnologies.com; www.pacificbiosciences.com). While it has been increasingly easy to obtain whole genome re-sequence data, the need for high quality reference genomes is essential for accurate annotation of gene sequences and to enable rapid alignment of re-sequence data. As the cost per base of sequencing rapidly declines, the bottleneck has now shifted to the need for bioinformatics expertise to analyze large amounts of sequence data. The difficulty is how to extract useful information from NGS data—whether to define a subset of informative SNP loci across the genome or to identify promising SNP haplotypes at key genes and alleles for crop improvement.

One major challenge has been to identify and validate sets of informative genome-wide SNP loci from large sequence data sets that will function well as SNP markers. The first steps are to filter the data to ensure that the specific SNP variant in question has been observed multiple times, is single copy in the genome, and has no nearby variant that might interfere with the assay design. This can be further refined with population-based filtering across accessions to ensure that the SNP has a minor allele frequency (MAF) above a certain threshold within and between target germplasm groups, which will eliminate sequence errors and rare SNPs, while maximizing chances SNP markers will be polymorphic and informative (Fig. 1). Correlation between SNPs is also used to select tagging SNPs that represent all of the linkage disequilibrium (LD) blocks across the genome (Zhao *et al.* 2011; Chen *et al.* 2013c), in some cases also incorporating information of allelic variation at targeted functional genes for breeding (Yu *et al.* 2014). The SNP selection process is especially important in designing fixed arrays, since poor-performing and monomorphic SNPs will impact the performance of the array. One caveat, however, is that highly selected SNP chip designs will inherently have an ascertainment bias that will affect diversity analysis results. Even with the best informatics approaches to SNP selection, it is still recommended to go through several steps of SNP validation to optimize sets of high quality SNP markers. This process was followed during the development of several fixed arrays in rice: the high quality Nipponbare reference sequence provided the starting point for an array-based re-sequencing project across 20 varieties that identified 160,000 SNP loci (IRGSP 2005; McNally *et al.* 2009). That

data was combined with BAC-end sequences to select a set of 44,100 SNP loci, out of which 36,901 were high-performing across 413 *O. sativa* accessions (Zhao *et al.* 2011). Subsequently, that set of 36,901 validated SNPs was used to select informative 384-SNP sets for multiple germplasm groups (Thomson *et al.* 2012). At each step, the SNP validation efforts of the previous generation helped to eliminate problematic SNP markers and narrow down sets of robust, evenly spaced, high quality SNPs for diversity analysis and genome-wide scans (Tung *et al.* 2010).

Large pools of NGS data are also valuable for characterizing SNP haplotypes that enable precise tracking of beneficial alleles for breeding applications. Now that many important genes and QTLs have been cloned, there has been progress to identify functional SNPs and gene-based SNP haplotypes to improve selection of target alleles. This process requires knowledge of the genetic donor providing the gene or QTL, along with some idea of the LD of the markers linked to the allele of interest. For marker-assisted backcrossing (MABC) applications, this is fairly straightforward since the donor introgression will usually be at least 1 Mb in size and easily tracked with flanking SNP markers polymorphic between the donor and recurrent parent. However, the availability of high density SNP genotyping and genome-wide association studies (GWAS) across sets of diverse germplasm has led to the identification of important chromosome segments with



Subgroup	id5000197	id5000200	id5000204	id5000205	ud5000025	id5000209	id5000217	id5000223
IND	A	T	G	T	T	A	A	C
IND	A	T	G	T	T	A	A	G
IND	A	T	G	N	T	A	N	C
IND	A	T	G	T	T	A	A	C
IND	A	T	G	T	T	A	A	C
IND	A	C	G	T	T	A	A	G
IND	A	T	G	T	T	A	A	G
IND	A	T	G	T	T	A	A	C
IND	A	T	G	T	T	A	N	C
IND	A	T	G	T	T	A	A	G
IND	A	T	G	T	T	A	A	G
IND	A	T	G	T	T	A	A	C
AUS	A	C	G	T	T	A	A	G
AUS	A	C	A	C	T	A	C	C
AUS	C	C	G	T	C	G	C	C
AUS	A	C	G	T	T	A	A	G
AUS	C	C	G	T	C	G	C	C
AROMATIC	A	C	G	T	T	A	A	G
AROMATIC	A	C	G	T	T	A	A	G
AROMATIC	A	C	G	T	T	A	A	G
AROMATIC	A	C	G	T	T	A	A	G
AROMATIC	A	C	G	T	T	A	A	G
AROMATIC	A	C	G	T	T	A	A	G
AROMATIC	A	C	G	T	T	A	A	G
TEJ	A	C	A	C	T	A	C	C
TEJ	A	C	A	C	T	A	C	C
TEJ	A	C	A	C	T	A	C	C
TEJ	A	C	A	C	T	A	C	C
TEJ	A	C	A	C	T	A	C	C
TEJ	A	C	A	C	T	A	C	C
TEJ	A	C	A	C	T	A	C	C

Fig. 1. Example of patterns of informative SNPs within and between subgroups. A subset of the rice 44K SNP data from Zhao *et al*. 2011 is shown for representative accessions from four subgroups: *indica* (IND), *aus* (AUS), *aromatic* (Aromatic), and *temperate japonica* (TEJ). Eight SNP loci are shown flanking a gene target, with three SNPs outlined: id5000200 is an example of a SNP mostly monomorphic within subgroups, but polymorphic between *indica* the others; ud5000025 is an example of a SNP monomorphic within all groups except for two *aus* accessions; and id5000223 is segregating within *indica* and aus, and polymorphic between aromatic and *temperate japonica*. In practice, the minor allele frequencies (MAF) within and between subgroups will be used as criteria during the SNP selection process.

smaller LD blocks (on the order of 50-100 kb for inbred crops such as rice, but much smaller for outcrossing species such as maize), along with an interest in predicting target alleles across breeding lines and varieties having unknown relationships with the original genetic donors. The challenge then becomes characterizing the desired chromosomal segments or IBD blocks predictive for the trait of interest at that particular gene. In a few cases the functional nucleotide polymorphism (FNP) can be assayed directly, especially if the causal variant is a SNP, but more commonly a set of closely linked SNPs will define the unique haplotype that is associated with the targeted IBD block.

This is where NGS data becomes important, since a haplotype map (HapMap) of the genome can help define the extent of local LD decay, along with common haplotype block segments, as has been recently characterized for the maize genome (Chia *et al.* 2012). Haplotype blocks have also been used to track the origin of agronomically important alleles, as can be seen by recent examples of haplotype analyses for cloned genes in rice (Sweeney *et al.* 2007; Kovach *et al.* 2007; Famoso *et al.* 2011; Shao *et al.* 2013). Note that the term "haplotype" is used to describe a set of SNPs on a single chromosome that are statistically associated (derived from "haploid genotype"). For outcrossing species, a fair amount of effort is needed to define the haplotype phasing for long heterozygous chromosome segments, information which is not normally directly observed from NGS data. However, for inbred species, such as rice, haplotype phasing is not a factor since most of the genome will be homozygous, so the SNP haplotype merely refers to closely linked SNP markers along the chromosome that are inherited together. With the massive output of short read sequence technologies, re-sequencing studies using large NGS data sets have become commonplace for most crop species, providing the raw data that will enable haplotype analysis at key loci across the genome.

High-throughput SNP genotyping: fixed array SNP platforms

The early successes with high-throughput SNP genotyping relied on fixed sets of SNP markers assayed using microarrays. For example, the first phase of the human HapMap project employed whole-genome SNP arrays from Illumina and Affymetrix for large scale SNP genotyping (International HapMap Consortium, 2005). The Illumina BeadArray technology uses beads covered with specific oligos that fit into patterned microwells allowing for highly multiplexed SNP detection, initially employing the GoldenGate assay that incorporates locus and allele-specific oligos for hybridization followed by allele-specific extension and fluorescent scanning (Shen *et al.* 2005). The BeadArray technology was expanded to higher density arrays with Infinium assays, which are based on two-color single base extension from a single hybridization probe per SNP marker (Steemers *et al.* 2006; Steemers and Gunderson 2007). GoldenGate assays were later deployed using VeraCode microbeads with the BeadExpress Reader to genotype up to 384 SNPs using a fluidic system instead of printed arrays (Lin *et al.* 2009). In contrast, Affymetrix implemented their GeneChip arrays using photolithographic printing of oligos on an array, followed by hybridization to overlapping allele-specific oligos consisting of perfect match and mis-match probes for SNP calling (Matsuzaki *et al.* 2004). More recently the Affymetrix Axiom technology, based on a two color, ligation-based assay with 30-mer probes, allows simultaneous genotyping of 384 samples with 50K SNPs or 96 samples x 650K SNPs (Hoffmann *et al.* 2011; www.affymetrix.com).

The advantages of fixed array SNP platforms include a range of multiplex levels providing rapid high-density genome scans, robust allele calling with high call rates, and cost-effectiveness per data point when genotyping large numbers of SNPs. Between the different Illumina and Affymetrix technologies, a wide range of options is available for custom genotyping of various numbers of samples x SNPs, such as running 24 samples by 3K up to 700K SNPs or 384 samples by 50K SNPs. By carefully selecting informative, evenly spaced SNPs across the genome, these arrays are powerful tools for GWAS and diversity analysis, as has been achieved in rice using Illumina 1,536 and 50K SNP arrays (Zhao *et al.* 2010; Chen *et al.* 2013c) and an Affymetrix 44K SNP chip (Zhao *et al.* 2011). For running fixed SNP sets with lower costs per sample, GoldenGate 384-SNP sets have been implemented for diversity analysis and QTL mapping in rice (Chen *et al.* 2011; Thomson *et al.* 2012; Wang *et al.*

2013; Baltazar *et al.* 2014; Chen *et al.* 2014). More recently, two independent Infinium rice 6K chips were designed to achieve high-density genome-wide scans at a reasonable cost per sample: one designed to target functional genes in addition to genome-wide loci (Yu *et al.* 2013) and another from Cornell University designed to be informative within and between *O. sativa* subgroups and between *O. sativa* and *O. rufipogon* (M. Wright and S. McCouch, pers. comm.). Since these arrays are based on two-color fluorescent dyes, they require allele-calling algorithms that use clustering of the three genotype groups, as provided by Illumina's GenomeStudio software, or statistical modeling of the raw intensity data, as provided by the Alchemy software (Wright *et al.* 2010). With proper data analysis QC steps, SNP arrays can provide very high quality data with low missing data rates. Fixed arrays also make it easier to build up SNP fingerprinting databases, since the same set of SNPs are genotyped across samples, helping to ensure more complete data sets for comparisons between accessions.

While fixed arrays have been the SNP genotyping workhorses over the past decade, they have several disadvantages. First, it is expensive to design a custom SNP array, which also limits the number of re-designs that can be used to optimize the chip, in addition to needing a large initial commitment to get volume discounts to make them more cost-effective. For this reason, fixed arrays are best used when a "universal" design can be employed to make them widely usable across a broad range of germplasm—thus allowing the development cost to be spread across a large number of users, as can be implemented through a consortium model for designing custom SNP chips useful to the larger community. However, this presents a key challenge: to cover rare SNPs across multiple germplasm groups, a universal design can quickly become too large and expensive (and will result in large numbers of monomorphic loci for non-target germplasm groups), while using multiple population-specific chips adds to the development costs and limits the number of users needing any particular chip. In addition, the process of selecting informative SNPs for different germplasm groups will introduce ascertainment bias: these SNP variants no longer represent a set of random, neutral loci, but instead present a biased view of genetic relationships depending on what selection criteria were used to select the SNPs (Moragues *et al.* 2010). Moreover, the costs per sample for fixed arrays have not decreased as quickly as competing technologies. That being said, fixed arrays will likely play a useful role for many years to come for those requiring high quality, easy-to-analyze data of a set of stable SNP loci across large numbers of samples, and who can afford to pay a higher cost for the convenience of using SNP chips.

High-throughput SNP genotyping: flexible SNP platforms

In addition to genotyping systems employing fixed SNP arrays, there are a number of high-throughput technologies available to run flexible sets of SNP markers. At the low range of the spectrum, PCR-based fluorescently-labeled SNP assays, such as TaqMan® and KASP™ markers, can be run one marker at a time and scanned on real-time PCR machines or fluorescent plate readers. For these methods, the cost is determined by the size of the PCR reaction volume—since fewer reagents are needed for smaller volumes. Thus, a 5 uL reaction in a 384-well PCR plate is more cost-effective than running 15 uL reactions in 96-well PCR plates. Moving to 1,536-well PCR plates can further reduce the cost, but at this point automation becomes necessary. The 5′-nuclease TaqMan® assay, which combines PCR with competitive hybridization, has been considered the gold standard in SNP genotyping since it was introduced almost 20 years ago (Livak *et al.* 1995; Ranade *et al.* 2001). More recently, the KASP™ competitive allele-specific PCR system has gained popularity for lower cost single-plex genotyping of crop species, for example with uptake in wheat and maize (Neelam *et al.* 2013; Semagn *et al.* 2014). The KASP™ assay, using two competing allele-specific forward primers, one reverse primer, and a master mix with a FRET cassette and Taq polymerase, is available for ordering reagents to run in house as well as outsourcing to LGC's automated high-throughput genotyping facility (www.lgcgroup.com).

The concept of miniaturizing the reaction volumes for reducing the PCR reagent costs has been further advanced in a number of flexible high-throughput SNP systems, including Array Tape™ by Douglas Scientific, the OpenArray®

system from Life Technologies, and Dynamic Arrays™ from Fluidigm. The Douglas Scientific technology uses a production line of automated modules to process spools of Array Tape™ that contain the equivalent of 200 microplates with 800 nL – 1.6 μL reaction volumes through automated assay setup, PCR, and fluorescent detection—genotyping up to 150,000 data points per day using either TaqMan or KASP assays (www.douglasscientific. com). In addition to saving with small reaction volumes, Array Tape has the advantage of low-cost consumables and an automated production line that can handle very large numbers of sample by SNP combinations. Another flexible system is the OpenArray® platform from Life Technologies, where custom TaqMan® assays can be purchased pre-loaded onto OpenArray® plates with different combinations of samples and assays, providing up to 3,072 reactions at 33 nL volumes for genotyping up to 70,000 data points per day (www.lifetechnologies.com). In addition, Fluidigm has several formats of Dynamic Array™ Integrated Fluidic Circuits (IFCs), such as 96 samples x 96 SNPs or 192 samples x 24 SNPs, that bring the reaction volumes even smaller—down to 7 to 10 nL in size, with the option to run TaqMan®, KASP™, or Fluidigm's own SNPtype™ assays (Wang *et al.* 2009; www.fluidigm.com). Another PCR-based fluorescently-labeled SNP genotyping technique uses high-resolution melting (HRM), which has been shown to have similar sensitivity and accuracy as TaqMan assays once the HRM conditions are optimized (Martino *et al.* 2010). On the other hand, other SNP systems use alternate techniques, such as the Sequenom MassARRAY® iPLEX system that employs MALDI-TOF mass spectrometry to identify SNP allele across up to 36 multiplexed products (Gabriel *et al.* 2009). Lastly, targeted amplicon sequencing approaches can provide a flexible genotyping system, which will be discussed in the section on genotyping by sequencing (GBS) below.

Each of these genotyping systems has their own pros and cons, since no single system is most efficient for every application; each breeding program, institute, or research community should select the platform that best addresses their specific needs. The flexible systems described above share the key advantage of being able to mix and match different SNPs for each set of samples—which reduces the wasted resources from genotyping a proportion of monomorphic loci as occurs with fixed SNP sets, especially when genotyping mapping populations. For diversity and fingerprinting, different subsets of informative SNPs for various germplasm groups can be optimized to enable running smaller sets of SNPs than a universal fixed array would require. Flexible SNP systems are also ideal for targeted SNPs, including functional SNPs and trait-specific haplotypes, since they have a very low cost per data point; however, for genome-wide scans they will quickly reach a threshold where fixed arrays or GBS will be more efficient. That threshold is determined by the cost per data point x number of SNP loci for a single-plex system versus the cost per sample of a multiplex system; i.e. it may still be reasonable to run 200-300 genome-wide SNPs on Fluidigm or Array Tape, but for a higher SNP density it might be more cost-effective to run a 6K SNP chip or GBS instead. However, these systems vary greatly in the initial capital investment required to purchase the equipment—often, with greater capital investments needed to reach the very low costs per data point (Table 1). This is a major factor to consider between setting up several small labs versus having centralized genotyping facilities, as will be discussed further below.

There are a number of examples of flexible SNP systems being used successfully for crop research and breeding. A major effort was initiated by the Generation Challenge Program as part of their Integrated Breeding Platform to validate KASP markers across globally-important field crops, ranging from 96 SNPs for groundnut to over 1,000 SNPs for ten other crops, including 1,250 for maize, 1,864 for wheat, and 2,015 for rice (www.integratedbreeding.net/snp-marker-conversion; Khera *et al.* 2013; He *et al.* 2014; Pariasca-Tanaka *et al.* 2014). The 1,250 KASP assays converted in maize have been successfully deployed in breeding programs at the International Maize and Wheat Improvement Center (CIMMYT) for QC analysis, QTL mapping, marker-assisted recurrent selection (MARS) and allele mining (Semagn *et al.* 2014). In cotton, KASP assays were used to validate 1,052 SNP markers, of which 367 SNPs were later run on 96.96 Dynamic Arrays™ on a Fluidigm EP1™ for genetic linkage mapping (Byers *et al.* 2012). A comparison in maize has also been performed to

Table 1. Examples of high-throughput SNP genotyping technologies.

Genotyping Platform	Technology	SNP x sample combinations	Capital investment	Cost per sample	Advantages
Illumina Infinium iSelect HD	Fixed array	3,072 − 700K SNPs x 24 samples	High (iScan)	Moderate to high	Highly multiplexed
Affymetrix Axiom	Fixed array	50K SNPs x 384 samples; 650K SNPs x 96 samples	High (GeneTitan)	Moderate to high	Highly multiplexed
Douglas Array Tape	Flexible, PCR-based	1 SNP/sample x 76,800 reactions/reel	Very High (Nexar, Soellex, Araya)	Very low	Ultra high-throughput
Fluidigm Dynamic Arrays	Flexible, PCR-based	96 SNPs x 96 samples; 24 SNPs x 192 samples	Moderate (IFC Controller, FC1, EP1)	Low	High-throughput
RE-based GBS	Genotyping by sequencing	~10K-100K SNPs x 96 or 384 samples	Low to moderate (NGS outsourced or in-house)	Low to moderate	Lots of data relative to the cost
Amplicon sequencing	Genotyping by sequencing	Variable (e.g. 20-500 SNPs x 48-384 samples)	Low to moderate (NGS outsourced or in-house)	Low to moderate	Multiple targeted loci at once

identify 162 versatile SNP markers that could be successfully converted across four genotyping platforms: GoldenGate, Infinium, KASPar and TaqMan (Mammadov *et al.* 2012). Likewise, approximately 300 SNPs previously run on the rice 44K SNP chip and 384-SNP GoldenGate assays have been converted into Fluidigm SNPtype assays for running on an EP1 Reader at IRRI (unpublished data).

High-throughput SNP genotyping: genotyping by sequencing (GBS)

In addition to fixed arrays and flexible methods, the approach of using NGS for low-cost genotyping, called "genotyping by sequencing" (GBS), has become increasingly popular. While whole genome sequence data can be used to call SNP variants, for most crop species it is still too expensive to obtain deep sequence data merely for genotyping purposes. Thus, a number of approaches have been developed to bring down the cost of NGS to a level where it can be used for routine genotyping—which entails lower coverage sequencing, often by running multiple barcoded DNA samples in a single lane of an NGS machine. So at the simplest level, genotyping by sequencing can be achieved by low coverage "skim" sequencing, as has been used recently in rice with 0.02X-0.13X sequence coverage for mapping populations and approximately 1X sequence coverage for diverse germplasm (Huang *et al.* 2009; Huang *et al.* 2010; Xu *et al.* 2010; Huang *et al.* 2012). Skim sequencing, however, spreads out the sequence reads across the whole genome, making SNP calling more difficult due to the low coverage at any particular locus. For this reason, most genotyping by sequencing approaches employ reduced representation techniques that will focus the sequencing reads on a subset of the genome, preferably at discrete loci that will be used for SNP calling.

Most genotyping by sequencing techniques make use of restriction enzyme (RE) digestion, followed by adapter ligation, PCR and sequencing. The first of these used NGS at restriction-site associated DNA (RAD) tags by restriction digestion and ligation of adapters containing unique barcode sequences for sample multiplexing (Baird *et al.* 2008). Thus the RAD sequencing method takes advantage of focusing sequencing reads only on the tags flanking a restriction site, which allows higher levels of multiplexing while maintaining deep enough sequence coverage at the RE cut sites for effective SNP calling, as has been demonstrated in barley (Chutimanitsakun *et al.* 2011). An

improved RE-based GBS technique was subsequently developed that simplified the library preparation steps for 96 and 384-sample multiplexed GBS and provided a large number of genome-wide SNPs based on Illumina sequencing (Elshire *et al.* 2011). This technique has been successfully used in a number of species, generating 24,186 genome-wide markers in barley (Elshire *et al.* 2011), between 56,807 and 63,388 SNPs in conifers (Chen *et al.* 2013a), from 60-100K polymorphic SNPs filtered down to 30,984 and 17,387 high quality SNPs in rice (Spindel *et al.* 2013; Bandillo *et al.* 2013), from 54,455 and 97,190 informative SNPs in maize mapping populations (Guimaraes *et al.* 2014; Ogugo *et al.* 2014) and 681,257 genome-wide SNP markers in a diverse collection of maize germplasm (Romay *et al.* 2013). A GBS service has been set up at the Genomic Diversity Facility at Cornell University, offering 96 up to 384-plex GBS for a large number of different species (http://www.biotech.cornell.edu/brc/genomic-diversity).

While the original GBS protocol employed a single-enzyme protocol, a two-enzyme modification has been successfully employed in barley, wheat and oat (Poland *et al.* 2012a; Poland *et al.* 2012b; Huang *et al.* 2014). Two-enzyme GBS was also modified for use with the Ion Torrent PGM and Proton sequencing platforms (Mascher *et al.* 2013). At the same time, similar techniques have been described for Sequence-Based Genotyping (SBG; Truong *et al.* 2012; Poecke *et al.* 2013), Diversity Array Technology sequencing (DArTseq; Cruz *et al.* 2013); RESTriction Fragment SEQuencing (RESTseq; Stolle and Moritz 2013), and Restriction Enzyme Site Comparative Analysis (RESCAN; Kim and Tai 2013).

In contrast to the random, genome-wide SNP loci produced by RE-based GBS approaches, there are also several targeted re-sequencing approaches that can be used for genotyping. In the past, Sanger sequencing of PCR amplicons has been used for SNP variant detection, but it is too expensive for large-scale genotyping projects. Thus, recent efforts have focused on taking advantage of the power of NGS while maximizing the number of amplicons and samples that can be pooled into a single NGS run. For example, the Targeted Amplicon Sequencing (TAS) method uses a two-step PCR process to amplify specific targets across the genome and then add a barcode multiplex identifier across multiple individuals before pooling and sequencing (Bybee *et al.* 2011). This method was further modified with a one-step PCR method for preparing amplicon tags for sequencing (Clarke *et al.* 2014), and a related technique for Genome-Tagged Amplification (GTA) has been described for preparing sets of 96 samples x 192 amplicons (Ho *et al.* 2014). There are also commercial kits available to enable amplicon sequencing, such as Illumina's TruSeq Custom Amplicon and the Ion AmpliSeq custom DNA panels; however, these have so far been limited to a small number of model species. In addition, the Access Array™ from Fluidigm can efficiently prepare pools of 48 samples x 48 amplicons for next-generation sequencing. Amplicon sequencing approaches are ideal for genotyping known and novel variants at key genes for traits of interest, but are still too expensive for routine genotyping for breeding.

One major challenge with GBS approaches is the considerable investment needed for bioinformatics support to properly analyze, curate and store the massive amounts of sequence data obtained from running GBS on large populations. GBS analysis pipelines are required to group the sequence tags, align to a reference genome (if available), call SNP variants, and assign calls to individual samples. For example, the *Stacks* software package was developed to analyze and call SNPs from sequenced RAD-tags, whether *de novo* or by comparison to a reference genome (Catchen *et al.* 2011). Likewise, the TASSEL-GBS pipeline was developed primarily for species with a reference genome available, and has been optimized for dealing with large data sets (Glaubitz *et al.* 2014). The TASSEL-GBS software has a Discovery Pipeline aimed at merging all GBS data obtained for a particular species and calling all possible SNPs across the genome; for example, this pipeline was used on 46.8 billion sequence reads obtained from 31,978 samples to distill 97.5 million GBS tags, resulting in the calling of 955,690 useful SNPs in maize (Glaubitz *et al.* 2014). Once the SNP catalog from the discovery build is obtained, this is used by the TASSEL-GBS Production Pipeline to routinely call SNPs on new samples with minimal computational time required (Glaubitz *et al.* 2014). Although the TASSEL-GBS pipeline currently has a command line interface, the graphical user interface

version of TASSEL can be used for downstream processing of the SNP matrix, including filtering SNPs based on missing data and minor allele frequency, pulling out subsets of SNPs, and performing genome-wide association studies (Bradbury *et al.* 2007). For GBS calling in species without a reference genome, the Universal Network-Enabled Analysis Kit (UNEAK) software can sort through networks of repeats, paralogs and error tags to identify reciprocal tag pairs that can be used for SNP calling (Lu *et al.* 2013).

GBS has a number of advantages that has led to its rapid uptake (Poland and Rife, 2012). First, GBS performs SNP discovery and genotyping simultaneously, without the ascertainment bias that occurs when selecting sets of SNPs for fixed arrays, and without any prior information needed. It also has a low entry cost to establish a manual GBS library preparation workflow, while at the same time is it amenable to setting up an automated workflow using liquid handling workstations. The greatest advantage, however, is that GBS leverages the rapidly falling costs of NGS to provide an excellent balance of low costs per sample and high-density genome-wide SNP data. With tweaking of the choice of restriction enzymes, the number of samples multiplexed per run, and the sequencing platform, GBS can be further fine-tuned to provide a wide range of SNP densities at varying costs per sample (Beissinger *et al.* 2013). There is still an issue with relatively high rates of missing data from GBS; however, this can be alleviated with imputation—where missing SNP calls are imputed to fill in the gaps. For example, the software BEAGLE is optimized for use on diverse heterozygous populations, and uses patterns of haplotypic variation in a reference panel to infer genotypes at missing loci (Browning and Browning 2009). For inbred lines and breeding populations, two methods were recently introduced: Full-Sib Family Haplotype (FSFHap) and Fast Inbred Line Library ImputatioN (FILLIN), which are implemented in TASSEL 5.0 and are optimized for imputing missing genotypes in GBS data without requiring knowledge of the parental genotypes (Swarts *et al.* 2014). As more whole genome sequence and GBS data is accumulated for crop species, imputation will become more accurate and will enable higher levels of multiplexed GBS to provide even lower cost, high-resolution SNP data.

Issues to consider when deciding on SNP genotyping options

One of the main issues to consider when evaluating options for SNP genotyping is whether to develop in-house facilities or outsource to a service provider. In most cases, the new genotyping technologies require a large capital investment in order to provide very low costs per sample; moreover, these platforms are most efficient when they run very large numbers of samples, due to discounts for high volume purchases of reagents and consumables. Thus there has been a shift for smaller labs to outsource their genotyping needs to commercial service providers or for core facilities or "genotyping hubs" to be set up to serve the needs of local or regional communities of researchers and breeders. Although it can be convenient to outsource to a service provider who takes on the risk of upgrading infrastructure when equipment becomes obsolete, there becomes a point when having a core facility in-house becomes more efficient—especially if there is enough demand to keep the genotyping platforms running at full capacity. In these cases, the advantages of having a centralized core facility in-house include: faster turnaround times, being able to optimize protocols and markers to a few target crops, and avoiding the hassle of shipping seeds, leaf tissue or DNA samples out of the country. On the other hand, having service providers and core facilities available to accept DNA samples from anywhere in the world allows for unprecedented flexibility for smaller labs and breeding programs. In either case, it is essential to have professional level sample tracking, along with solid QA/QC measures, to ensure that reliable and accurate data is provided.

Another issue to consider is how much effort should be spent for MAS and MABC of targeted, trait-specific SNPs for known genes and QTLs versus employing more high-density genome-wide SNP scans (Fig. 2). This will depend on several factors, including the crop species, the genetic architecture of the trait of interest, and the number of breeding-relevant, large-effect genes and QTLs that are fine-mapped and cloned. As was discussed earlier in this review, one important aspect of targeted selection is knowledge on the size of the LD block and IBD status of the

target—whether it's included in a large introgression from a known, recent donor, or selected from a smaller LD block across a set of diverse germplasm. "Diagnostic" SNP markers, whether functional SNPs or gene-based haplotypes, can be used to profile diverse sets of germplasm with unknown pedigrees for the specific allele of interest, while flanking SNPs are best used to transfer introgressions from a known genetic donor from a recent cross. On the other hand, some traits do not lend themselves to a targeted approach and are better suited to genome-wide prediction or genomic selection methods that use precise phenotyping and high-density genotyping on a training population to calculate genome estimated breeding values (GEBVs), which are then used on breeding populations for rapid cycles of selection with the genotype data alone (Heffner *et al.* 2009; Bernardo 2010; Jannink *et al.* 2010; Poland *et al.* 2012b). Ideally targeted selection of major loci can be combined with a genomic selection approach. This will also affect the choice of SNP genotyping platform as well, whether to invest in fixed arrays, flexible systems, GBS, or a combination of systems.

Above all these factors, bioinformatics plays an essential role behind any SNP genotyping program. Whether analyzing clusters of two-color fluorescent intensities or complex sequence data from GBS, robust pipelines need to be set up for routine allele calling, preferably in relation to a high quality reference genome. Moreover, many labs and breeding programs need to merge data across platforms, such as whole genome sequence data, fixed arrays, GBS and targeted single-plex assays—which requires careful attention to the DNA strand used to design the SNP assay, in addition to the relation of the SNP to the reference genome. At the same time, lower density data can be imputed using NGS data, whether from related lines or from a global HapMap. Once data is compiled and imputed, it needs to be analyzed for quality control and stored in a database structure that allows for user-friendly queries and downstream data analysis. The final step is then enabling access and decision support tools for breeders to integrate SNP markers into their selections to accelerate the progress in their breeding programs, such as the breeding information management systems being developed at IRRI

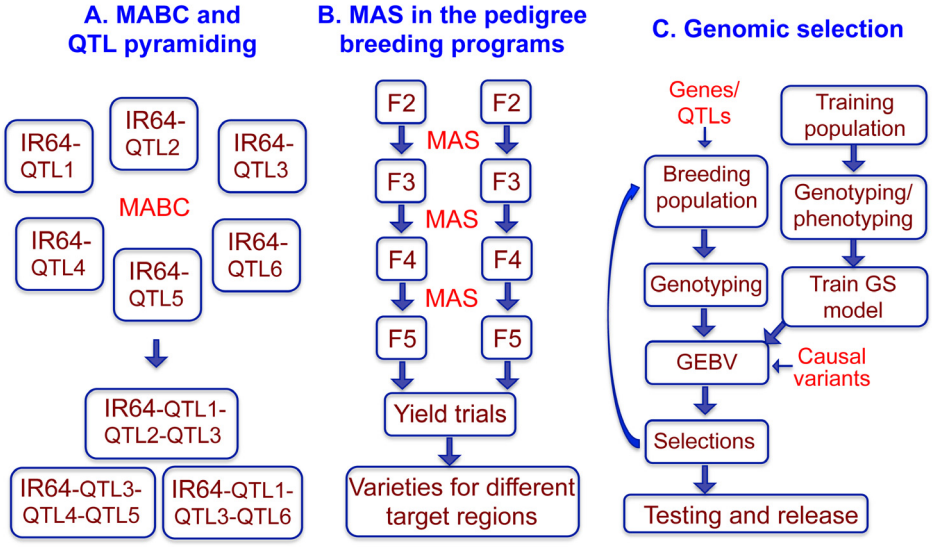


Fig. 2. Breeding schemes for integrating SNP genotyping into molecular breeding programs. (A) Marker-assisted backcrossing (MABC) and QTL pyramiding are used to rapidly transfer major genes and QTLs into existing high-yielding varieties, as in the example of rice variety IR64 with different combinations of added QTLs; (B) marker-assisted selection (MAS) can be used during pedigree and bulk selection to eliminate undesirable lines and fix major genes in the breeding populations; and (C) genomic selection uses genome-wide scans to calculate genomic estimated breeding values (GEBVs), which can be combined with selection for major genes and QTLs as fixed effects to improve the model.

(E. Nissilä, pers. comm.) and by the Integrated Breeding Platform (www.integratedbreeding.net).

Setting up an in-house genotyping facility: IRRI's Genotyping Services Lab

An example of a core facility for SNP genotyping is IRRI's Genotyping Services Laboratory (GSL), which was recently set up to provide rapid and cost-effective marker services to research and breeding groups at IRRI and the larger rice community. GSL currently has 12 full time staff divided into teams for marker validation, optimizing lab operations, running the routine genotyping services, and interfacing with IRRI's bioinformatics group. A sample processing workflow is being optimized to efficiently move leaf tissue in the greenhouse and field into the DNA extraction and SNP genotyping pipelines. Leaf tissue is sampled using a Brooks PlantTrak Hx™ handheld plant sampling and barcoding device, which allows up to 12 leaf punches per sample and 100 samples per plastic magazine cartridge, and reduces issues of human error during the sampling process (www.brooks.com). The leaf tissue is lyophilized in the plastic cartridges before being transferred into a deep 96-well plate using the PlantTrak Sx™ benchtop unit, which also provides a plate layout with the identity of each sample based on the barcodes. Steel balls are added to each well and the tissue is ground using a TissueLyser II (www.qiagen.com) or Geno/Grinder® (www.spexsampleprep.com). The 96-well plate then moves into our DNA extraction protocol, which is currently based on the LGC sbeadex™ magnetic bead kit that uses a two-step binding mechanism to provide high quality DNA for downstream SNP and NGS protocols (www.lgcgroup.com). After incubation with lysis buffer, the lysate is transferred either to a KingFisher Flex 96 from Thermo Scientific (www.fishersci.com) or with an oKtopure™ robotic system (www.lgcgroup.com) for automated DNA extraction. The DNA samples are then checked on a Nanodrop 8000 (www.nanodrop.com) or with the PicoGreen® dsDNA quantitation assay (www.lifetechnologies.com) before moving to SNP genotyping.

As of late 2014, the GSL genotyping platforms were focused on using a Fluidigm EP1™ Reader for targeted SNP markers and an Illumina Infinium rice 6K chip for genome-wide scans. For targeted genotyping, 96.96 Dynamic Array IFCs are used for diversity analysis, QTL mapping and background selection, while 192.24 sample x SNP format IFCs are used for running trait-specific SNPs across large populations. Targeted SNPs have been selected as flanking key gene and QTL positions from the rice 44K SNP chip (Zhao *et al.* 2011) or using informative markers from GoldenGate 384-SNP sets (Thomson *et al.* 2012), and have been converted to Fluidigm SNPtype™ assays. Additional functional SNPs from publications and promising SNPs from GWAS studies and GBS data are also being validated on Fluidigm. For the past five years, the workhorse for genome-wide SNP scans in the lab was the Illumina BeadXpress Reader running GoldenGate assays, with over 20,000 samples run on 384-SNP sets at IRRI; however, recently we prefer running the Infinium rice 6K chip developed by Susan McCouch at Cornell University. The 6K chip provides approximately 4,500 high quality SNP markers for diversity analysis, SNP fingerprinting, QTL studies, and characterizing donor introgressions in specialized genetic stocks, with over 4,400 samples genotyped at IRRI to date (unpublished data). For future high-density genome wide scans, GSL is also testing several GBS protocols, including 96 and 384-plex GBS based on Elshire *et al.* 2011 for outsourcing to an Illumina MiSeq, NextSeq, or HiSeq, and 96-plex GBS based on Mascher *et al.* 2014 for running on an Ion Proton machine recently set up at the Genomic Institute of Asia (GINA) facility on the IRRI campus. Lastly, legacy SSRs and functional indel markers are being genotyped on a 96-capillary Fragment Analyzer™ from Advanced Analytical (http://aati-us.com).

Recent efforts at GSL have also aimed towards improving sample tracking, SNP analysis, and data management in the lab. An integrated laboratory information management system (LIMS) is being optimized for GSL's operations using the web-based, cloud-hosted Biotracker™ LIMS from Ocimum Biosolutions (http://lims.ocimumbio.com). The LIMS is being configured to handle GSL's customer requests, user access, sample tracking, inventory management, and workflow management, while being accessible from any web browser. At the same time, barcoding is being implemented along the entire workflow, from leaf sampling

in the field, to plates of DNA samples, and for tracking SNP assays with a VisionMate 2D barcode reader (www.thermoscientific.com). Moreover, web-based SNP data analysis tools have been deployed through the IRRI Galaxy workbench to speed up SNP data filtering and formatting for downstream applications (R. Mauleon, pers. comm.).

CONCLUSIONS

Recent advances in molecular marker technology have enabled rapid high-throughput genotyping for pre-breeding discovery research as well as SNP deployment in breeding programs. Research and breeding groups now have a large number of options, including outsourcing to genotyping service providers or setting up a core facility based on one of the many genotyping platforms. With the rapid decrease in NGS costs, genotyping by sequencing (GBS) will become increasingly attractive to handle high-density genome-wide marker scans, as long as adequate bioinformatics support is available. Future prospects to increase the efficiency and impact of SNP genotyping will come on several fronts, including improved DNA extraction, more predictive SNP markers, more efficient GBS, and improved bioinformatics tools for SNP data analysis, management, and integration with breeders' selection decisions. While techniques for DNA extraction from leaf tissue can be further improved, a larger gain can be made by switching to automated seed chipping, which saves the embryo for germination while extracting DNA from the remainder of the seed, allowing genotyping to screen out unwanted individuals before going to the field (see Monsanto patent EP1869961B1). At the same time, further progress in cloning important QTLs and characterizing functional SNPs and allele-specific haplotypes will continue to provide improved predictive markers for targeted selection. Moreover, as whole genome sequence and GBS data accumulates, it will become more feasible to impute functional variants with genome-wide data. Alternatively, it may be increasingly possible to use a smaller number of low-cost markers for genome-wide scans and then impute back to the whole genome sequence data of the parental lines. In either case, having improved bioinformatics and SNP data management tools will be essential—the gains of the future will largely rest on the bioinformatics teams who are optimizing allele calling pipelines, building infrastructure for managing massive GBS data sets, and developing the tools that will seamlessly link SNP data with downstream applications for calculating GEBVs, tracking haplotypes, and assisting the breeders in making selections. The tsunami of sequence and SNP data has arrived; we should be prepared to take advantage of the data to accelerate progress in trait development, gene discovery, and increasing the rate of genetic gain for crop improvement.

ACKNOWLEDGMENTS

I am grateful to Susan McCouch and Mark Wright for the rice 384-SNP and 6K chip designs, and for Susan's vision for developing SNP resources for the rice community. I would like to thank my colleagues at IRRI for their discussions and support for using SNP markers to implement molecular breeding in rice: Joong Hyoun Chin, Ramil Mauleon, Eero Nissilä, Hei Leung, Bertrand Collard, R. K. Singh, Endang Septiningsih, Casiana Vera Cruz, Bo Zhou, Guoyou Ye, John Damien Platten, Nickolai Alexandrov, Kenneth McNally, K. K. Jena, Glenn Gregorio, and Abdelbagi Ismail. Lastly, I am very grateful for all of the hard work from my team at IRRI in the Genotyping Services Lab for validating targeted SNP markers, optimizing lab operations, processing samples in the lab, and assisting with data analysis and QC: Maria Ymber Reveche, Christine Jade Dilla-Ermita, Nadia Vieira Castañeda, Maria S. Dwiyanti, Geraldine Ann M. Layaoen, Geisha Sanchez, Venice Juanillas, Erwin Tandayu, Socorro Carandang, Grace Cariño, Annalhea Jarana, Cristomo Dizon, Krizzel Llantada, and Maria Elena Gamo. This work was supported by IRRI, the Global Rice Science Partnership (GRiSP) product 2.1.3, and the following grants: the Transforming Rice Breeding project from the Bill and Melinda Gates Foundation, the IRRI Scientific Know-How and Exchange Project with Syngenta, and the Japan Rice Breeding Project under the support of the Japan Ministry of Finance under the Policy and Human Resources Development Project of the World Bank.

REFERENCES

Alfred J, Dangl JL, Kamoun S, McCouch SR. 2014. New horizons for plant translational research. PLoS Biol. 12: e1001880.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS One 3: e3376.

Baltazar MD, Ignacio JCI, Thomson MJ, Ismail AM, Mendioro MS, Septiningsih EM. 2014. QTL mapping for tolerance of anaerobic germination from IR64 and the aus landrace Nanhi using SNP genotyping. Euphytica 197: 251-260.

Bandillo N, Raghavan C, Muyco PA, Sevilla MAL, Lobina IT, Dilla-Ermita CJ, Tung CW, McCouch S, Thomson M, Mauleon R, Singh RK, Gregorio G, Redona E, Leung H. 2013. Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. Rice 6: 1-15.

Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Johnson JM, Muttoni G, Vaillancourt B, Buell CR, Kaeppler SM, de Leon N. 2013. Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics 193: 1073-1081.

Bernardo R. 2010. Genomewide selection with minimal crossing in self-pollinated crops. Crop Sci. 50: 624-627.

Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B, Mayer KF. 2014. Plant genome sequencing—applications for crop improvement. Curr. Opin. Biotech. 26: 31-37.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635.

Browning BL, Browning SR. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Amer. J. Human Genet. 84: 210-223.

Bybee SM, Bracken-Grissom H, Haynes BD, Hermansen RA, Byers RL, Clement MJ, Udall JA, Wilcox ER, Crandall KA. 2011. Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. Genome Biol. Evol. 3: 1312-1323.

Byers RL, Harker DB, Yourstone SM, Maughan PJ, Udall JA. 2012. Development and mapping of SNP assays in allotetraploid cotton. Theor. Appl. Genet. 124: 1201-1214.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. 2011. Stacks: building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics, 1: 171-182.

Chen C, Mitchell SE, Elshire RJ, Buckler ES, El-Kassaby YA. 2013a. Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. Tree Genet. Genomes 9: 1537-1544.

Chen H, He H, Zhou F, Yu H, Deng XW. 2013b. Development of genomics-based genotyping platforms and their applications in rice breeding. Curr. Opin. Plant Biol. 16: 247-254.

Chen H, He H, Zou Y, Chen W, Yu R, Liu X, Yang Y, Gao YM, Xu JL, Fan LM, Li ZK, Deng XW. 2011. Development and application of a set of breeder-friendly SNP markers for genetic analyses and molecular breeding of rice (*Oryza sativa* L.). Theor. Appl. Genet. 123: 869-879.

Chen H, Xie W, He H, Yu H, Chen W, Li J, Yu R, Yao Y, Zhang W, He Y, Tang X, Zhou F, Deng XW, Zhang Q. 2013c. A high-density SNP genotyping array for rice biology and molecular breeding. Mol. Plant 7: 541-553.

Chen W, Chen H, Zheng T, Yu R, Terzaghi WB, Li Z, Deng XW, Xu J, He H. 2014. Highly efficient genotyping of rice biparental populations by GoldenGate assays based on parental re-sequencing. Theor. Appl. Genet. 127: 297-307.

Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz J, Gore M, Guill KE, Holland J, Hufford MB, Lai J, Li M, Liu X, Lu Y, McCombie R, Nelson R, Poland J, Prasanna BM, Pyhajarvi R, Rong T, Sekhon RS, Sun Q, Tenaillon MI, Tian F, Wang J, Xu X, Zhang Z, Kaeppler SM, Ross-Ibarra J, McMullen, MD, Buckler ES, Zhang G, Xu Y, Ware D. 2012. Maize HapMap2 identifies extant variation from a genome in flux. Nature Genet. 44: 803-807.

Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A, Cistué L, Corey A, Filichkina T, Johnson EA, Hayes PM. 2011. Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. BMC Genomics 12: 4.

Clarke LJ, Czechowski P, Soubrier J, Stevens MI, Cooper A.

2014. Modular tagging of amplicons using a single PCR for high-throughput sequencing. Mol. Ecol. Resour. 14: 117-121.

Cruz VMV, Kilian A, Dierig DA. 2013. Development of DArT Marker Platforms and Genetic Diversity Assessment of the US Collection of the New Oilseed Crop Lesquerella and Related Species. PloS One 8: e64062.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Rev. Genet. 12: 499-510.

Delmer DP. 2005. Agriculture in the developing world: connecting innovations in plant research to downstream applications. Proc. Nat. Acad. Sci. USA 102: 15739-15746.

Drenkard E, Richter BG, Rozen S, Stutius LM, Angell NA, Mindrinos M, Cho RJ, Oefner PJ, Davis RW, Ausubel FM. 2000. A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in Arabidopsis. Plant Phys. 124: 1483-1492.

Edwards D, Batley J, Snowdon RJ. 2013. Accessing complex crop genomes with next-generation sequencing. Theor. Appl. Genet. 126: 1-11.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS One 6: e19379.

Famoso AN, Zhao K, Clark RT, Tung CW, Wright MH, Bustamante C, Kochian LV, McCouch SR. 2011. Genetic architecture of aluminum tolerance in rice (*Oryza sativa*) determined through genome-wide association analysis and QTL mapping. PLoS Genet. 7: e1002221.

Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K. 2011. Crop genome sequencing: lessons and rationales. Trends Plant Sci. 16: 77-88.

Gabriel S, Ziaugra L, Tabbaa D. 2009. SNP genotyping using the Sequenom MassARRAY iPLEX platform. Curr. Protocols Human Genet. 60:2.12.

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PloS One 9: e90346.

Guimaraes CT, Simoes CC, Pastina MM, Maron LG, Magalhaes JV, Vasconcellos RC, Guimaraes LJM, Lana UGP, Tinoco CFS, Noda RW, Jardim-Belicuas SN, Kochian LV, Alves VMC, Parentoni SN. 2014. Genetic dissection of Al tolerance QTLs in the maize genome by high density SNP scan. BMC Genomics 15: 153.

He C, Holme J, Anthony J. 2014. SNP Genotyping: The KASP Assay, p. 75-86. In: F. Delphine, R Whitford (eds.). Crop Breed. Springer, New York.

Heffner EL, Sorrells ME, Jannink JL. 2009. Genomic selection for crop improvement. Crop Sci. 49: 1-12.

Ho T, Cardle L, Xu X, Bayer M, Prince KSJ, Mutava RN, Marshall DF, Syed N. 2014. Genome-Tagged Amplification (GTA): a PCR-based method to prepare sample-tagged amplicons from hundreds of individuals for next generation sequencing. Mol. Breed. 34: 977-988.

Hoffmann TJ, Kvale MN, Hesselson SE, Zhan Y, Aquino C, Cao Y, Cawley S, Chung E, Connell S, Eshragh J, Ewing M, Gollub J, Henderson M *et al.* 2011. Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Genomics 98: 79-89.

Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, Guan J, Fan D, Weng Q, Huang T, Dong G, Sang T, Han B. 2009. High-throughput genotyping by whole-genome re-sequencing. Genome Res. 19: 1068-1076.

Huang X, Lu T, Han B. 2013. Re-sequencing rice genomes: an emerging new era of rice genomics. Trends Genet. 29: 225-232.

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang Q, Li J, Han B. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nature Genet. 42: 961-967.

Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, Li W, Guo Y, Deng L, Zhu C, Fan D, Lu Y, Weng Q, Liu K, Zhou T, Jing Y, Si L, Dong G, Huang T, Lu T, Feng Q, Qian Q, Li J, Han B. 2012. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nature Genet. 44: 32-39.

Huang YF, Poland JA, Wight CP, Jackson EW, Tinker NA. 2014. Using genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat. PloS One 9: e102448.

International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437: 1299-1320.

International Rice Genome Sequencing Project (IRGSP).

2005. The map-based sequence of the rice genome. Nature 436: 793-800.

Jannink JL, Lorenz AJ, Iwata H. 2010. Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9: 166-177.

Khera P, Upadhyaya HD, Pandey MK, Roorkiwal M, Sriswathi M, Janila P, Guo Y, McKain MR, Nagy ED, Knapp SJ, Leebens-Mack J, Conner JA, Ozias-Akins P, Varshney RK. 2013. Single nucleotide polymorphism–based genetic diversity in the reference set of Peanut (spp.) by developing and applying cost-effective Kompetitive Allele Specific Polymerase chain reaction genotyping assays. Plant Genome 6: doi:10.3835/plantgenome2013.06.0019.

Kim SI, Tai TH. 2013. Identification of SNPs in closely related *Temperate Japonica* rice cultivars using restriction enzyme-phased sequencing. PloS One 8: e60176.

Kovach MJ, Sweeney MT, McCouch SR. 2007. New insights into the history of rice domestication. Trends Genet. 23: 578-587.

Komori T, Nitta N. 2005. Utilization of the CAPS/dCAPS method to convert rice SNPs into PCR-based markers. Breeding Sci. 55: 93-98.

Li JY, Wang J, Zeigler RS. 2014. The 3,000 rice genomes project: new opportunities and challenges for future rice research. GigaScience 3: 1-3.

Lin CH, Yeakley JM, McDaniel TK, Shen R. 2009. Medium-to high-throughput SNP genotyping using VeraCode microbeads, p. 129-142. In: P. Bugert (ed.) DNA and RNA Profiling in Human Blood. Humana Press, New York.

Liu Y, He Z, Appels R, Xia X. 2012. Functional markers in wheat: current status and future prospects. Theor. Appl. Genet. 125: 1-10.

Livak KJ, Marmaro J, Todd JA. 1995. Towards fully automated genome-wide polymorphism screening. Nature Genet. 9: 341-342.

Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, Buckler ES, Costich DE. 2013. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet. 9: e1003215.

Lübberstedt T, Zein I, Andersen JR, Wenzel G, Krützfeldt B, Eder J, Ouzunova M, Chun S. 2005. Development and application of functional markers in maize. Euphytica 146: 101-108.

Mammadov J, Chen W, Mingus J, Thompson S, Kumpatla S. 2012. Development of versatile gene-based SNP assays in maize (*Zea mays* L.). Mol. Breed. 29: 779-790.

Martino A, Mancuso T, Rossi AM. 2010. Application of high-resolution melting to large-scale, high-throughput SNP genotyping a comparison with the TaqMan® method. J. Biomol. Screen. 15: 623-629.

Mascher M, Wu S, Amand PS, Stein N, Poland J. 2013. Application of genotyping-by-sequencing on semiconductor sequencing platforms: a comparison of genetic and reference-based marker ordering in barley. PloS One 8: e76925.

Matsuzaki H, Dong S, Loi H, Di X, Liu G, Hubbell E, Law J, Berntsen T, Chadha M, Hui H, Yang G, Kennedy GC, Webster TA, Cawley S, Walsh PS, Jones KW, Fodor SP, Mei R. 2004. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nature Methods 1: 109-111.

McCouch SR, McNally KL, Wang W, Hamilton RS. 2012. Genomics of gene banks: A case study in rice. Amer. J. Botany 99: 407-423.

McCouch SR, Zhao K, Wright M, Tung CW, Ebana K, Thomson M, Reynolds A, Wang D, DeClerck G, Ali ML, McClung A, Eizenga G, Bustamante C. 2010. Development of genome-wide SNP assays for rice. Breeding Sci. 60: 524-535.

McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureasu RE, Stokowski R, Ballinger DG, Frazer KA, Cox DR, Padhukasahasram B, Bustamanate CD, Weigel D, Mackill DJ, Bruskiewich RM, Ratsch G, Buell CR, Leung H, Leach JE. 2009. Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proc. Natl. Acad. Sci. USA 106: 12273-12278.

Miura K, Ashikari M, Matsuoka M. 2011. The role of QTLs in the breeding of high-yielding rice. Trends Plant Sci. 16: 319-326.

Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, Russell JR. 2010. Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. Theor. Appl. Genet. 120: 1525-1534.

Morrell PL, Buckler ES, Ross-Ibarra J. 2012. Crop genomics: advances and applications. Nature Rev. Genet. 13: 85-96.

Neelam K, Brown-Guedira G, Huang L. 2013. Development

and validation of a breeder-friendly KASPar marker for wheat leaf rust resistance locus Lr21. Mol. Breed. 31: 233-237.

Ogugo V, Semagn K, Beyene Y, Runo S, Olsen M, Warburton ML. 2014. Parental genome contribution in maize DH lines derived from six backcross populations using genotyping by sequencing. Euphytica doi: 10.1007/s10681-014-1238-6

Pariasca-Tanaka J, Lorieux M, He C, McCouch S, Thomson MJ, Wissuwa M. 2014. Development of a SNP genotyping panel for detecting polymorphisms in *Oryza glaberrima*/*O. sativa* interspecific crosses. Euphytica doi: 10.1007/s10681-014-1183-4

Poecke RM, Maccaferri M, Tang J, Truong HT, Janssen A, Orsouw NJ, Salvi S, Sanguineti MC, Tuberosa R, Vossen EA. 2013. Sequence-based SNP genotyping in durum wheat. Plant Biotech. J. 11: 809-817.

Poland JA, Brown PJ, Sorrells ME, Jannink JL. 2012a. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PloS One 7: e32253.

Poland JA, Rife TW. 2012. Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5: 92-102.

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigackere S, Crossae J, Sánchez-Villedae H, Sorrells M, Jannink JL. 2012b. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5: 103-113.

Ranade K, Chang MS, Ting CT, Pei D, Hsiao CF, Olivier M, Pesich R, Hebert J, Chen YI, Dzau VJ, Curb D, Olshen R, Risch N, Cox DR, Botstein D. 2001. High-throughput genotyping with single nucleotide polymorphisms. Genome Res. 11: 1262-1268.

Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA, McMullen MD, Holland JB, Buckler ES, Gardner CA. 2013. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 14:R55.

Rounsley S, Marri PR, Yu Y, He R, Sisneros N, Goicoechea JL, Lee SJ, Angelova A, Kudrna D, Luo M, Affourtit J, Desany B, Knight J, Niazi F, Egholm M, Wing RA. 2009. De novo next generation sequencing of plant genomes. Rice 2: 35-43.

Semagn K, Babu R, Hearne S, Olsen M. 2014. Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. Mol. Breed. 33: 1-14.

Shao G, Tang S, Chen M, Wei X, He J, Luo J, Jiaoa G, Huc Y, Xiea L, Hu P. 2013. Haplotype variation at *Badh2*, the gene determining fragrance in rice. Genomics 101: 157-162.

Shen R, Fan JB, Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Garcia EW, McBride C, Steemers F, Garcia F, Kermani BG, Gunderson K, Oliphant A. 2005. High-throughput SNP genotyping on universal bead arrays. Mutat. Res. 573: 70-82.

Spindel J, Wright M, Chen C, Cobb J, Gage J, Harrington S, Lorieux M, Ahmadi N, McCouch S. 2013. Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theor. Appl. Genet. 126: 2699-2716.

Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL. 2006. Whole-genome genotyping with the single-base extension assay. Nat. methods 3: 31-33.

Steemers FJ, Gunderson KL. 2007. Whole genome genotyping technologies on the BeadArray™ platform. Biotechnol. J. 2: 41-49.

Stolle E, Moritz RF. 2013. RESTseq−efficient benchtop population genomics with RESTriction Fragment SEQ-uencing. PloS One 8: e63960.

Swarts K, Li H, Navarro JAR, An D, Romay MC, Hearne S, Acharya C, Glaubitz JC, Mitchell S, Elshire RJ, Buckler ES, Bradbury PJ. 2014. FSFHap (Full-Sib Family Haplotype Imputation) and FILLIN (Fast, Inbred Line Library ImputatioN) optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome doi: 10.3835/plantgenome2014.05.0023

Sweeney MT, Thomson MJ, Cho YG, Park YJ, Williamson SH, Bustamante CD, McCouch SR. 2007. Global dissemination of a single mutation conferring white pericarp in rice. PLoS Genet. 3: e133.

Thiel T, Kota R, Grosse I, Stein N, Graner A. 2004. SNP2CAPS: a SNP and INDEL analysis tool for CAPS marker development. Nucleic Acids Res. 32: e5.

Thomson MJ, Zhao K, Wright M, McNally KL, Rey J, Tung CW, Reynolds A, Scheffler B, Eizenga G, McClung A, Kim H, Ismail AM, de Ocampo M, Mojica C, Reveche

MY, Dilla-Ermita CJ, Mauleon R, Leung H, Bustamante C, McCouch SR. 2012. High-throughput single nucleotide polymorphism genotyping for breeding applications in rice using the BeadXpress platform. Mol. Breed. 29: 875-886.

Truong HT, Ramos AM, Yalcin F, de Ruiter M, van der Poel HJ, Huvenaars KH, Hogers RCJ, van Enckevort LJG, Janssen A, van Orsouw NJ, van Eijk MJ. 2012. Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. PLoS One 7: e37565.

Tung CW, Zhao K, Wright MH, Ali ML, Jung J, Kimball J, Tyagi W, Thomson MJ, McNally K, Leung H, Kim H, Ahn SN, Reynolds A, Scheffler B, Eizenga G, McClung A, Bustamante C, McCouch SR. 2010. Development of a research platform for dissecting phenotype–genotype associations in rice (*Oryza* spp.). Rice 3:205-217.

Van Damme V, Gómez-Paniagua H, de Vicente MC. 2011. The GCP molecular marker toolkit, an instrument for use in breeding food security crops. Mol. Breed. 28:597-610.

Varshney RK, Terauchi R, McCouch SR. 2014. Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. PLoS Biol. 12: e1001883.

Wang J, Lin M, Crenshaw A, Hutchinson A, Hicks B, Yeager M, Berndt S, Huang W, Hayes RB, Chanock SJ, Jones RC, Ramakrishnan R. 2009. High-throughput single nucleotide polymorphism genotyping using nanofluidic Dynamic Arrays. BMC Genomics 10: 561.

Wang K, Qiu F, Dela Paz MA, Zhuang J, Xie F. 2013. Genetic diversity and structure of improved *indica* rice germplasm. Plant Genet. Resour. 12: 248-254.

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

Xu J, Zhao Q, Du P, Xu C, Wang B, Feng Q, Liu Q, Tang S, Gu M, Han B, Liang G. 2010. Developing high throughput genotyped chromosome segment substitution lines based on population whole-genome re-sequencing in rice (*Oryza sativa* L.). BMC Genomics 11: 656.

Yu H, Xie W, Li J, Zhou F, Zhang Q. 2014. A whole-genome SNP array (RICE6K) for genomic breeding in rice. Plant Biotech. J. 12: 28-37.

Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR. 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat. Commun. 2:467.

Zhao K, Wright M, Kimball J, Eizenga G, McClung A, Kovach M, Tyagi W, Ali ML, Tung CW, Reynolds A, Bustamante C, McCouch SR. 2010. Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. PLoS One 5: e10780.