Accelerating public sector rice breeding with high-density KASP markers derived from whole genome sequencing of indica rice

Steele, Katherine; Quinton-Tulloch, Mark; Amgai, Resham B. ; Dhakal, Rajeev ; Khatiwada, Shambhu P. ; Vyas, Darshna; Heine, Martin ; Witcombe, John

Molecular Breeding

DOI:
10.1007/s11032-018-0777-2

Published: 01/04/2018

Peer reviewed version

Research Article

Accelerating public sector rice breeding with high-density KASP markers derived from whole genome sequencing of *indica* rice

Katherine A. Steele · Mark J. Quinton-Tulloch · Resham B. Amgai · Rajeev Dhakal · Shambhu P. Khatiwada · Darshna Vyas · Martin Heine · John R. Witcombe

Electronic supplementary material The online version of this article (doi:) contains supplementary material, which is available to authorised users.

K.A.Steele (corresponding author: email k.a.steele@bangor.ac.uk, Tel 00 44 1248 388655, ORCID 0000-0003-4896-8857) M.J. Quinton-Tulloch (ORCID 0000-0001-5713-0736) · J.R. Witcombe
School of the Environment, Natural Resources and Geography, SENRGY, Bangor University, Bangor, Gwynedd, LL57 2UW, UK

R.B. Amgai · S.P. Khatiwada
Nepal Agricultural Research Council, Biotechnology Division, PO Box No. 1135 Kathmandu, Nepal

R. Dhakal
Anamolbiu Private Ltd., P.O. Box 28, Jagritichok, Bharatpur-11, Chitwan, Nepal. Current address: LI-BIRD, Head Office: PO Box 324, Gairapatan, Pokhara, Kaski, Nepal

D. Vyas
LGC Genomics, Units 1 & 2, Trident Industrial Estate, Pindar Road, Hoddesdon, Herts, EN11 0WZ, UK

M. Heine
LGC Genomics LGC Genomics TGS Haus 8, Ostendstr. 25, 12459 Berlin, Germany: Current address: NuGEN Technologies Inc.201 Industrial Road, Suite 310 San Carlos, CA 94070, USA

Abstract

Few public sector rice breeders have the capacity to use NGS-derived markers in their breeding programmes despite rapidly expanding repositories of rice genome sequence data. They rely on >18,000 mapped microsatellites (SSRs) for marker-assisted selection (MAS) using gel analysis. A lack of knowledge about target SNP and InDel variant loci has hampered their uptake of KASP, a proprietary technology of LGC genomics. KASP is a cost-effective single-step genotyping technology, cheaper than SSRs and more flexible than genotyping by sequencing (GBS) or array based genotyping when used in selection programmes. Before this study there were 2,015 rice KASP in the public domain, mainly identified by array-based screening leaving large proportions of the rice genome with no KASP marker coverage. Here we have addressed the urgent need for a wide choice of appropriate rice KASP markers, and demonstrated that NGS can provide full genome marker coverage. Through resequencing of nine *indica* rice breeding lines or released varieties, this study has identified 2.5 million variant sites. Stringent filtering of variants generated 1.3 million potential KASP assay designs, including 92,500 potential functional markers. This strategy delivers a 650-fold increase in potential selectable KASP markers at a density of 3.1 marker per 1 kb in the *indica* crosses analysed with 377,178 polymorphic KASP marker design sites on average per cross. This knowledge is available to breeders and has been utilised to improve the efficiency of public sector breeding in Nepal, enabling identification of polymorphic KASP at any mapped trait or QTL in relevant crosses. Validation of 39 new KASP was carried out by genotyping progeny from a range of crosses and detecting segregating alleles to aid trait selection during marker-assisted backcrossing, where target traits included rice blast and BLB resistance. Furthermore, we provide the software for plant breeders to generate KASP designs from their own datasets.

Keywords Bacterial blight · genomic selection (GS) · kompetitive allele-specific PCR (KASP) · marker-assisted selection (MAS) · next generation sequencing (NGS) · physical mapping · rice blast · single-nucleotide polymorphism (SNP) · allele mining software

Introduction

Cost is a major factor that determines whether or not marker assisted selection (MAS) is a viable breeding method for national programmes and smaller breeders. Despite advantages such as improved reliability, MAS will rarely be used if it is more expensive than phenotyping. Reducing the costs of markers increases the frequency of cases where MAS is more cost effective than phenotyping. KASP is a cost effective and flexible proprietary technology of LGC Genomics, however, public sector rice breeders have been slow to adopt it because KASP assays have not been widely published in linkage maps to the same extent as SSRs. Where costs permit, SSRs are still the marker technology most commonly used by most public sector breeders, especially for marker-assisted rice breeding (Miah et al., 2013) because they alone provide a sufficient choice of mapped markers. Breeders can choose from over 18,000 SSRs (Narshimulu et al., 2011) while the use of KASP markers is limited by the number publically available and these offer limited options in crosses between *indica* lines.

Prior to this study, 2,015 KASP assays were made publically available for rice (Pariasca-Tanaka et al., 2015) that were developed in rice using a arrray-based Illumina GoldenGate technology by the Generation Challenge Program of the Consultative Group for International Agricultural Research (CGIAR) to analyse crosses between *O. sativa indica* and *O. glaberrima*. The original 2,015 SNPs had been identified from the OryzaSNP project (McNally et al., 2009) and Sanger sequencing. OryzaSNP used 20 genetically diverse genotypes to discover SNPs via long range PCR and re-sequencing of microarrays. To date, and to our knowledge, no large scale SNP and InDel discovery effort has been published for rice where NGS whole genome re-sequencing was used specifically to identify potential KASP, yet there is an urgent need for large numbers of KASP markers in rice.

KASP is a single-step genotyping technology that reveals, via fluorescence resonance energy transfer (FRET), pre-identified co-codominant alleles for both SNP and InDel variations between parents and progeny in segregating crosses for MAS. KASP has the major advantage of improved cost-effectiveness because it is both cheaper and more reliable than other marker technologies, including other sequence-based markers, such as TaqMan (Patil et al, 2017). A resource of available genome-wide variations would facilitate KASP to be used for whole genome coverage in genomic selection (GS) which has been pioneered using and array-based technology. Array-based genotyping and NGS-based genotyping technologies (such as Genotyping by Sequencing) are not being taken up by public sector breeders because they lack the flexibility and ease afforded by SSRs. KASP offer the benefits of SSRs plus the added ability of being able to detect functional markers within target genes and

KASP are easier to use: either LGC Genomics can provide a full KASP genotyping service or the KASP reagents can be ordered from them for carrying out assays in a basic molecular laboratory. KASP technology is more rapid than SSRs and it has scalability that makes it suitable for a wide range of experimental designs with greatly varying target loci and sample numbers (He et al., 2014). These can range from only a single marker, such as a selectable marker for a specific gene, through to several thousands of markers for applications such GS. The effectiveness of KASP has been demonstrated in plant-breeding applications, including quality control analysis of germplasm (Semagn et al., 2012; Ertiro et al., 2015), screening for candidate alleles and genotyping (Mideros et al., 2013; Pham et al., 2015), bulk segregant analysis and genetic mapping (Ramirez-Gonzalez et al., 2014; Mackay et al., 2014), and MAS (Cabral et al., 2014; Leal-Bertioli et al., 2015).

Marker assisted breeding has been introduced in Nepal's national programmes, mainly based on SSRs but recently incorporating existing KASP for background selection. However, few of the existing rice KASP were suitable for selection at the breeders' targets of BLB and blast resistance genes and aroma QTLs. Therefore, the objective of the work reported here was to identify appropriate SNPs and InDels, for this purpose, in order to facilitate the uptake of KASP for greater efficiency of rice breeding. At current rates the KASP genotyping service is estimated to be 60% cheaper than running SSRs in-house at NARC's laboratories in Kathmandhu, Nepal: Genotyping 475 samples with 10 assays costs $0.20 per data point with KASP (full genotyping service, including shipping costs), $0.39 with in-house KASP and $0.53 with in-house SSRs.

This study used whole genome NGS specifically to identify large numbers of SNP and InDel variations and used bioinformatics filtering of NGS reads to discover potential KASP assays throughout the rice genome. We re-sequenced nine *indica* rice lines and aligned the sequences to the *indica* reference genome to maximise the identification of applicable loci. The study provides new evidence on the effectiveness of using NGS sequence data from a limited number of lines and makes comparisons between the new potential KASP and those that were available prior to this work for density and genomic distribution throughout the physical map in a range of crosses.

Materials and methods

Sequence data generation

Plant materials and DNA extraction for NGS

Nine *indica* rice lines (Table S1) were selected for sequencing. Three (Sunaulo Sugandha, Anamol Masuli and Sugandha-1) were from a breeding programme in Nepal (Witcombe et al., 2013) and one (Khumal-4) is a widely grown mid-hill variety in Nepal. They are all being used as recurrent parents for rice breeding in Nepal. Sunaulo Sugandha and Sugandha-1 are aromatic. Four (IR64, IR71033, IR65482, IRBB60 and Loktantra) were chosen as donors of resistance to the diseases bacteria blight (caused by *Xanthomonas oryzae* pv. *oryzae*) and blast (caused by *Magnaporthe oryzae*). Seedlings were grown in a controlled environment room at Bangor University (BU) and DNA extracted at BU from the leaves of one representative seedling per variety using Qiagen DNEasy kits (Qiagen, Manchester, UK). The plants were grown to maturity and visually checked for phenotypic uniformity within each variety.

NGS, read processing and read alignment

Paired-end sequencing, using the Illumina HiSeq 2000 platform, and read processing was carried out at LGC Genomics (Berlin, Germany). For bioinformatics analysis Illumina adaptor sequences were removed and quality trimming of adaptor-clipped reads was performed, removing reads containing Ns, and 3'-end trimming reads to get a minimum average Phred quality score of 20 over a window of ten bases. Reads with a final length of less than 20 bases were discarded. The sequences have been submitted to the NCBI Sequence Read Archive under BioProject accession PRJNA395505 (available at www.ncbi.nlm.nih.gov/bioproject/395505).

The reference genome sequence used was cultivar 93-11 of *Oryza sativa* ssp. *indica*. The Read Assembly version ASM465v1 of 93-11, sequenced and annotated by the Beijing Genome Institute (Yu et al., 2002; Zhao et al., 2004) was downloaded from EnsemblPlants (http://plants.ensembl.org). Sequencing reads were aligned against this reference using Bowtie2 (Langmead and Salzberg 2012). Discordant or mixed paired-read alignments were not permitted, with all other alignment parameters kept as default. Only read pairs with both reads aligning in the expected orientation were used in subsequent analyses.

Variant calling

SAMtools (Li et al., 2009) was used to calculate genotype likelihoods and identify single nucleotide polymorphisms (SNPs) and InDels between the aligned sequencing reads and the *O. sativa* ssp. *indica* reference. SNPs or insertions with a read depth higher than 200 were filtered out (using vcfutils) due to likelihood of variable copy number repeats influencing read mapping. Also, those with a read depth of less than five were removed. Custom Perl scripts were used to identify variants between all pairwise combinations of the nine rice lines, based on the variant calls made for each variety against the *indica* reference. The positions of the variants were compared against the annotated gene and coding sequence positions to test whether they corresponded to functional mutations.

Variant filtering for suitability as KASP markers

Variant Call Format (VCF) files generated by SAMtools (see above) were parsed using a custom Perl script (Supplementary File S1) to retrieve the flanking sequences 50 bp either side of each variation site, and identify variants suitable for KASP markers following a stepwise identification process (Figure. S1). The criteria for selection were that the flanking sequences a) did not contain any InDels; b) contained a maximum of four ambiguous bases; c) had a base coverage of at least five at any position; and d) had no more than four consecutive repeats of any 1-5 nucleotide sequence. Variants that passed this filtering were defined as potential KASP markers. The SNP positions of the potential KASP markers were used in the diversity analysis of potential KASP assays below.

In-silico analysis of diversity using the new and existing KASP markers for the nine rice lines.

The sequence variants of each of the 1,329,325 potential KASP that passed the filtering (Figure. S1) were used to make 45 comparisons - the 36 possible pairwise comparisons between these nine lines and the nine comparisons to the *indica* reference cultivar. For the 2,015 existing KASP markers based on rice SNPs that had previously been developed (Pariasca-Tanaka et al. 2015), the KASP primer sequences were aligned against the *indica* reference using BLAST (Altschul et al., 1990) to determine if the sequence reliably aligned to *indica* (those with at least 95% identity). This eliminated 205 KASPs specific to *japonica*. A further 731 KASP were at sites where no polymorphism was detected between any of the nine lines and the *indica* reference. This left 1,159 existing KASP markers that were used for the same 45 comparisons. The density and distribution of potential and existing KASP were analysed for the nine lines compared tp *indica* reference genome, and for pairwise comparisons between the resequenced lines, using the genome locations of the SNPs and InDels of the KASP marker targets.

Validation of KASP assays for genotyping in segregating populations

Plant materials and DNA extraction for genotyping

For KASP genotyping, plants representing the nine sequenced parental lines and progeny lines (at F_1 and BC_1) derived from fifteen crosses between pairs of parents were grown in the field or polyhouse in Nepal, in October 2015 and October 2016. Leaf samples were collected from each plant using the LGC Genomics' Plant Sample Collection Kits and delivered to LGC Genomics (Hoddesdon, Herts.,UK) for DNA extraction and KASP genotyping (full service). All plants were from the marker-assisted breeding programmes of either Anamolbiu or NARC and parental lines were used as controls for MAS. The first three plates were screened in the first round (69 KASP including 21 new ones) and third round (with a further 5 new KASP). Five plates were screened with in the second round with 86 KASP. Two plates containing only BC_1 material were screened with 40 KASP (39 new) in the fourth round.

Selection of 46 variants and development of new KASP assays

The SNPs or InDels selected for validation in this study were either located near to/within target resistant gene alleles (for BLB or blast) or to known fragrance QTLs, or they were useful as background markers in regions where no existing KASP were suitable. They included 35 variants that passed the filtering criteria and 11 variants that did not pass. All 46 new KASP assays gave *in-silico* validated primers in LGC's Kranken Software and KASP primers were produced by LGC and used in their standard protocol for KASP validation. Here, we define validation as where the KASP assay was successfully used for genotyping in at least one cross. In total, four separate rounds of genotyping were carried out on different sets of segregating lines, each round having a different combination of new and existing KASP assays.

Marker-level, cross-level, and assay-level validation of the KASP assays was carried out using bioinformatics on genotype results from all four rounds. KASP markers were considered to be validated if they successfully genotyped any of the tested progeny lines and identified both predicted parental alleles. Cross-level validation assessed whether a marker could be validated at the marker level using only progeny lines originating from a specific pair of parental lines. Assay-level validation tested whether or not each individual KASP assay had produced genotyping results. Genotyping results from parental lines were not used for validation as they

would be expected to be homozygous for the tested alleles and thus the genotyping results could not be used to validate successful binding of both of the KASP allele-specific primers.

Identification and subsequent filtering of 'background' markers

From the existing 1,159 KASP that reliably aligned to the *indica* reference genome we identified those that were polymorphic *in-silico* in at least three of the bi-parental crosses used for this study. Of these, 75 were selected as 'background' markers for genotyping because they were distributed in genomic regions required for recurrent parent selection. Of the 75 existing KASP, 48 met our filtering criteria (Figure. S1) for selecting variants appropriate for marker generation. These existing KASP were used for genotyping in parental and progeny lines by LGC Genomics (Hoddesdon, Herts., UK).

Results

Sequencing read alignment and identification of variants

More sequencing reads of all of the nine re-sequenced rice lines aligned in the expected orientation to the *indica* reference (mean of 92.1% ± 0.96) than to the *japonica* reference (mean of 88% ± 0.69). Mean *indica* genome coverage was 89% with a mean sequencing depth of 59 for the nine lines (Table S2). We identified variations between the *indica* reference and at least one of the nine lines at 2,561,351 unique sites. For over half (56.5%) of these sites two or more lines were polymorphic against the reference genome and for 3.4% of sites all nine were polymorphic against the reference, whereas more than one-million variant sites were found in only a single line (Figure. S2). There was an average of 0.96 million homozygous variations (SNPs and InDels) between each of the nine rice lines compared with the *indica* reference variety 93-11 (Figure S3). IR71033 was the most similar line to the reference (0.78 million variations) and Sunaulo Sugandha the least similar (1.1 million variations).

Identification of potential KASP markers and functional markers

To identify KASP markers that would be informative for crosses between the nine lines and the *indica* reference, *in silico* filtering of the 2,561,351 variation sites was carried out, based on the composition of their flanking sequences (Figure. S1). The KASP marker sequences were determined for the 1,329,325 sites that passed the filtering criteria, i.e., a conversion rate of 51.9% of the total variation sites.

For each of the nine lines, those variations that were suitable for KASP markers were categorised according to the nature of the polymorphism against the *indica* reference (Table 1), determined according to the annotated gene and coding sequence positions. The majority of potential KASP were situated in non-coding portions of the genome, with 78% located in intergenic regions and 11% in introns. Of the remaining 11 % of variations located in the exons, 68% are predicted to result in functional differences due to changes in the amino acids encoded.

[Table 1 about here]

Comparing diversity in nine *indica* lines at new KASP

This new approach of pair-wise comparisons for each of the nine resequenced lines against each other and against the *indica* reference genome identified many more potential new KASP than previously existed for rice (Table 2). The highest diversity in the pairwise comparisons was >511,000 in the new set (IR65482 with Sunaulo Sugandha) but only 522 in the existing set (Loktantra with Sunaulo Sugandha). The least informative number of KASP markers in the pairwise comparisons was >245,000 in the new set (IR64 with IR71033) compared with 361 in the existing set (IR64 with IR71033). A similar pattern was seen for comparisons with the *indica* reference where the average number of informative KASP markers was 388,540 in the new set and 451 in the existing set. The highest number of new markers against the reference genome was 459,229 for Sunaulo Sugandha, compared with a maximum of 496 for Loktantra with the existing markers.

[Table 2 about here]

The new KASP markers were distributed throughout the entire genome with high levels of marker density (Figure. 1). In a great majority of cases (86.9%) the distance between consecutive informative markers was less than 1 kb with a median distance of 127 bp in all pairwise combinations. Chromosomal distribution plots of markers informative for each pairwise combination of the sequenced lines show very few regions with no markers (Figure. S4).

[Figures 1 & 2 about here on two whole consecutive pages, use B&W for print and colour for online]

Comparing diversity in nine *indica* lines at existing KASP

Of the 1,890 existing KASP markers that could be aligned against the *indica* reference, 1,159 (61%) were polymorphic between at least one of the sequenced lines and the *indica* reference genome. However, they were not evenly distributed throughout the genome nor across all lines (Figure. 2). In pairwise comparisons between the lines there were between 345 and 520 informative polymorphic markers for each cross combination (Table 2 and Figure. S5). There were some areas of the genome that had polymorphisms in all of the crosses (e.g, between 0.5 Mbp and 10 Mbp on Chromosome 6) but many regions had polymorphisms only in specific pairs of crosses. There were also many regions lacking any polymorphisms (e.g. on Chromosome 7 between 9 Mbp and 16 Mbp

there is only one polymorphic marker and it is only in crosses with Loktantra). Consecutive informative existing KASP markers were not often close together, in only 1.1% of cases were they closer than 1 kb. The median distance between markers of 353 kb across all pairwise combinations of lines was over 2,700 times longer than that found for the new markers (Figure S6 and Tables S3 and S4).

The positions of the 1,159 markers that aligned to the indica reference and corresponded to polymorphic sites in our lines were compared with the positions of the new KASP markers. Matches were found for 727 (62.7%) of the existing markers, with new markers not being identified at the other genomic positions due to the filtering criteria applied by the marker detection algorithm (Figure. S1). The filtering method excluded 37% (432 of 1,159) existing KASP markers because they had InDels or repeats of five or more bases in their flanking regions.

KASP validation for use in genotyping

KASP genotyping was carried out on F_1 and BC_1 progeny of fifteen crosses between pairs of the nine re-sequenced lines. For the purposes of KASP validation, genotyped progeny of different generations were grouped according to the parental lines initially crossed, with a KASP assay being considered validated for a particular cross group if genotyping was successful in showing segregation of alleles for one or more progeny lines from any generation of the cross. Eighty-three markers (35 new and 48 existing KASP) that passed our filtering criteria (Figure. S1) were tested on at least one cross, with a total of 412 unique marker-cross combinations. Successful genotyping results were obtained for 78 (94.0%) of these markers including 30 of the new markers, with 394 of 412 (95.6%) marker-cross combinations being successful (Tables S5 and S6).

Genotyping was also carried out with 38 markers (11 new and 27 existing KASP) that did not meet our filtering criteria, the 11 new markers were designed manually through visualisation of the aligned sequencing reads at sequences for target traits. 31 (81.6%) of these markers gave genotyping results in at least one of the progeny tested, including 9 of the new markers. 232 marker-cross combinations were tested, with 201 (86.6%) being successful (Tables S5 and S6).

Parental lines were genotyped with the KASP markers as controls and the results confirmed the presence of the predicted alleles in the parents but also revealed within-line genetic variation for some of the parents at some loci (data not shown). Expected allelic ratios were detected in segregating progeny for all successfully genotyped crosses (data not shown) and the results informed selection of donor alleles and recurrent (background)

alleles for 70 existing KASP and 39 newly validated KASP (Table 3 and Table S7), of which 30 were discovered from filtering and 9 identified by manual design.

Discussion

SNPs provide the highest genome-wide density of genetic variants and occur in both coding and non-coding genomic regions. Due to their bi-allelic nature not all SNPs and InDels will be polymorphic for all cross combinations. We showed that, for the existing 2,015 rice KASP markers (all SNPs) published by Pariasca-Tanaka et al., 2015, in all cross combinations there were very large gaps between markers across the rice genome (Figure. 2). Only 1,890 were applicable to *indica* and the number that were informative between any pair of nine *indica* lines studied here varied from as few as 361 to, at most, 522. It is unsurprising that the existing set is insufficient to meet all rice-breeding challenges because, apart from being less numerous than available SSRs, they were derived from chip-based technologies based on SNPs nominated by the rice community to address particular breeding targets. Hence, a much higher density of SNPs or InDel variants are needed in order to identify suitable markers for selection in a broader range of specific crosses.

Thousands of SNPs have previously been employed in array-based platforms such as those used in the Illumina Bead Array and the Affymetrix GeneChip (Thomson, 2014). However, unlike KASP, these fixed sets of SNPs do not meet the need of breeders that wish to assay a small number of polymorphic markers known to be linked to traits of interest in their breeding populations, and to have the opportunity to change the set of markers used in subsequent generation. Next-generation sequencing (NGS) technologies have been used to re-sequence of diverse rice genomes or directly for genotyping in technologies such as genotyping by sequencing (GBS) (McCouch et al., 2010; Kumar et al., 2012), but most variants have only been made available as array-based platforms.

Here, NGS was used for re-sequencing nine *indica* breeding lines, chosen with no deliberate effort to select for high diversity, and it identified an average of 1.05 million SNP or InDel variants between any one of the individual rice lines and the *indica* reference genome, out of a total of 2.5 million variants across the whole set of lines (available at www.ncbi.nlm.nih.gov/bioproject/395505). By mining this data using bioinformatics filtering we discovered hundreds of thousands of potential new KASP markers giving high resolution coverage over the entire genome (Figure. 1, Table 1). This has vastly reduced the number of regions with no selectable markers (compare Figures 1 and 2) and offers breeders access to over 1.3 million informative KASP with a minimum of more than 245,000 for any paired combination of the 9 rice lines (Table 2) and has produced over 650 times more KASP marker sequences than were available in rice to date. Approximately 92,500 (7%) were located in exons and altered the amino acid sequence encoded, so could be used as functional markers (Table 1). For all pairwise comparisons between lines, over 98% of consecutive informative markers were less than 10 kb

apart, with over 85% being less than 1 kb apart (Figure S6). Some of these comparisons were between lines having common recent ancestors (Table S1) so this data set should provide a high density of polymorphic KASP assays across the genome in almost any cross. Moreover, these estimates are conservative, as many more KASP markers would be identified if the filtering criteria were relaxed slightly to allow the detection of KASP markers in gaps at target genomic regions. Relaxing the criteria is a practical option as they were very stringent; they provided a 52% conversion rate for new markers from identified variations but excluded 37% of the 1,159 existing KASP.

Early rice genome sequencing of *indica* and *japonica* revealed about 1 SNP per kb (Feltus et al., 2004) and the material that is subsequently selected to be re-sequenced determines the density of NGS based markers identified. Re-sequencing of 12 cultivated and wild accessions of *indica* that were chosen for gave an average of 5.7 nucleotide differences per kb diversity (Xu et al., 2012). Here, we found an average of 3.1 variations per kb in *indica* lines used for breeding in Nepal. All lines were adapted to lowland or medium land and we made no attempt to include diverse lines adapted to greatly different rice ecosystems. They were simply chosen on the basis of being in current breeding programmes in Nepal and seven of the lines were either bred at IRRI or had IRRI lines in their recent ancestry. Hence, the high frequency of KASP markers (Figure 1) we have discovered should also apply to most, or all, other *indica* material of interest to breeders. The total of 2.5 million SNPs in the nine lines compares favourably with the total of 18.9 million found in the 3000 Rice Genomes Project where the lines included were highly diverse across all the *O. sativa* cultivated groups (Li et al., 2014).

We have demonstrated how high-throughput sequencing data can be used to identify so many new KASP markers that they will be useful for many traits across many parental combinations. A set of 39 fully validated marker designs are given here (Table 3). These design sequences can be submitted directly to LGC Genomics for purchase of KASP primers through their KASP by Design (KBD) or KASP on Demand (KOD) services, or for their full genotyping service. This allows breeders with no bioinformatics expertise to utilise these markers in their breeding programs. The software provided (Supplementary File S1) enables breeders to easily generate KASP marker designs using their own, or publicly available, NGS datasets – for any species. In addition, the sequencing reads for the nine resequenced lines is a valuable resource containing suitable variants for numerous breeding targets.

The work has led to suitable KASP assays for NARC and Anamolbiou (Nepal) and many more assays are being rolled out to rice breeders in India (SKUAST) and Pakistan (NIBGE) with the support of LGC Genomics. Work is currently underway, using data from the 3000 Rice Genomes Project, to generate over 20,000 KASP marker designs, of which 4,000 will be fully validated, that will be applicable to a diverse range of rice

varieties. These will be made available on the LGC Genomics website, allowing breeders to purchase KASP markers close to existing SSR markers or in a region of interest, without the need for any bioinformatics analysis. In the meantime, the paper authors can be contacted for details of KASP marker designs based on the nine resequenced lines, for any particular region of interest of the rice genome. This increase in the number of usable KASP markers has great practical benefits to public sector plant breeders who can use the knowledge derived from this project to incorporate KASP into MAS to accelerate selection of new varieties. These KASP assays are new tools that can complement other innovations introduced to accelerate varietal adoption by farmers in developing nations (Witcombe et al 2016) to expedite yield improvement and increase food security. By increasing the number of available KASP markers this work is expected to remove the barriers to their adoption so they can accelerate progress in rice breeding for future generations.

Acknowledgements This study was co-funded by Innovate UK (Grant number 131781).

374 **References**

375 Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. J Mol Biol
376 215:403-410

377 Cabral AL, Jordan MC, McCartney CA, You FM, Humphreys DG, MacLachlan R, Pozniak CJ (2014)
378 Identification of candidate genes, regions and markers for pre-harvest sprouting resistance in wheat (*Triticum*
379 *aestivum* L.). BMC Plant Biol 14:340

380 Ertiro BT, Ogugo V, Worku M, Das B, Olsen M, Labuschagne M, Semagn K (2015) Comparison of kompetitive
381 allele specific PCR (KASP) and genotyping by sequencing (GBS) for quality control analysis in maize. BMC
382 Genomics 16:908

383 Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) An SNP resource for rice genetics and
384 breeding based on subspecies indica and japonica genome alignments. Genome research 14:1812-1819

385 He C, Holme J, Anthony J (2014) SNP genotyping: the KASP assay. Methods in Mol Biol 1145:75–86

386 Huang N, Angeles ER, Domingo J, Magpantay G, Singh S, Zhang G, Kumaravadivel N, Bennett J, Khush
387 GS (1997) Pyramiding of bacterial blight resistance genes in rice: Marker-assisted selection using RFLP and
388 PCR. Theor Appl Genet 95:313–320

389 Kumar S, Banks TW, Cloutier S (2012) SNP discovery through next-generation sequencing and its applications.
390 Int J of Plant Genomics 831460

391 Kush GS (2005) IR varieties and their impact. Int Rice Res Inst, Los Baños, Phillipines. p. 139

392 Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357–359

393 Leal-Bertioli SCM, Cavalcante U, Gouvea EG, Ball´en-Taborda C, Shirasawa K, Guimar˜aes PM, Jackson SA,
394 Moretzsohn MC (2015) Identification of QTLs for rust resistance in the peanut wild species Arachis magna
395 and the development of KASP markers for marker-assisted selection. G3 Genes|Genomes|Genetics 5:1403–
396 1413

397 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The
398 sequence alignment/map format and SAMtools. Bioinformatics, 25:2078-2079

399 Li, J.Y., Wang, J., Zeigler, R.S. (2014) The 3,000 rice genomes project: new opportunities and challenges for
400 future rice research. GigaScience, 3:8.

401 Mackay IJ, Bansept-Basler P, Barber T, Bentley AR, Cockram J, Gosman N, Greenland AJ, Horsnell R, Howells
402 R, O'Sullivan DM, Rose GA (2014) An eight-parent multiparent advanced generation inter-cross population
403 for winter-sown wheat: creation, properties, and validation. G3: Genes|Genomes| Genetics 4:1603-1610

McCouch SR, Zhao K, Wright M, Tung C-W, Ebana K, Thomsom 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, Bureau TE, Stokowski R (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proceedings of the National Academy of Sciences106:12273-8.

Miah G, Raffi MY, Ismail MR, Puteh AB, Rahim HA, Islam KN, Latif MA (2013) A review of microsattelite markers and their applications in rice breeding programs to improve blast disease resistance. Int J Mol Sci 14:22499-22528.

Mideros SX, Warburton ML, Jamann TM, Windham GL, Williams WP, Nelson RJ (2013) Quantitative trait loci influencing mycotoxin contamination of maize: analysis by linkage mapping, characterization of near-isogenic lines, and meta-analysis. Crop Sci 54:127–142

Narshimulu G, Jamaloddin M, Vemireddy LR, Anuradha G, Siddiq E (2011) Potentiality of evenly distributed hypervariable microsatellite markers in marker-assisted breeding of rice. Plant Breeding 130:314-320.

Pariasca-Tanaka J, Lorieux M, He C, McCouch S, Thomson MJ, Wissuwa M, (2015). Development of a SNP genotyping panel for detecting polymorphisms in *Oryza glaberrima*/*O. sativa* interspecific crosses. Euphytica 201:67-78.

Patil G, Chaudhary J, Vuong TD, Jenkins B, Qiu D, Kadam S, Shannon GJ, Nguyen, H. T. (2017). Development of SNP Genotyping Assays for Seed Composition Traits in Soybean. International Journal of Plant Genomics, 2017, 6572969. http://doi.org/10.1155/2017/6572969

Pham A-T, Harris DK, Buck J, Hoskins A, Serrano J, Abdel-Haleem H, Cregan P, Song Q, Boerma HR, and Li Z (2015) Fine mapping and characterization of candidate genes that control resistance to *Cercospora sojina* K. Hara in two soybean germplasm accessions. PLoS One, 10:e0126753

Ramirez-Gonzalez RH, Segovia V, Bird N, Fenwick P, Holdgate S, Berry S, Jack P, Caccamo M, Uauy C (2014) RNA-seq bulked segregant analysis enables the identification of high-resolution genetic markers for breeding in hexaploid wheat. Plant Biotech J 13:613–624

Rasheed A, Wen W, Gao F, Zhai S, Jin H, Liu J, Guo Q, Zhang Y, Dreisigacker S, Xia X, He Z (2016) Development and validation of KASP assays for genes underpinning key economic traits in bread wheat. Theor Appl Genet 10:1843-1860

433 Semagn K, Babu R, Hearne S, Olsen M (2014) Single nucelotide polymorphism genotyping using kompetitive
434 allele specific PCR (KASP): overview of the technology and its application in crop improvement. Molecular
435 Breeding 33:1–14
436 Semagn K, Beyene Y, Makumbi D, Mugo S, Prasanna BM, Magorokosho C, Atlin G (2012). Quality control
437 genotyping for assessment of genetic identity and purity in diverse tropical maize inbred lines. Theor Appl
438 Genet 125:1487–1501.
439 Thomson MJ (2014) High throughput SNP genotyping to accelerate crop improvement. Mol Breed 2:195-212.
440 Witcombe JR, Gyawali S, Subedi M, Virk DS, Joshi KD, (2013) Plant breeding can be made more efficient by
441 having fewer, better crosses. BMC Plant Biol 13:22
442 Witcombe J, Khadka K, Puri R, Khanal N, Sapkota A, Joshi K. (2016) Adoption of rice varieties. 2.
443 Accelerating Uptake. Experimental Agriculture.:1-7.
444 Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J (2012) Resequencing
445 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature
446 Biotech 30:105-111
447 Yang H, Li C, Lam HM, Clements J, Yan G, Zhao S (2015). Sequencing consolidates molecular markers with
448 plant breeding practice. Theor Appl Genet 128:779-795
449 Yu J, Hu S, Wang J et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science 296:79–
450 92
451 Zhao W, Wang J, He X, Huang X, Jiao Y, Dai M, Wei S, Fu J, Chen Y, Ren X, Zhang Y, Ni P, Zhang J, Li S,
452 Wang J, Wong GK, Zhao H, Yu J, Yang H, Wang J (2004) BGI-RIS: an integrated information resource and
453 comparative analysis workbench for rice genomics. Nucleic Acids Res 32: D377–382

Figures

Figure. 1 Distribution of potential new KASP markers polymorphic between each rice line and the indica reference. Rows represent the chromosomes, subdivided into the different lines in the order indicated on chromosome 12 (from top to bottom: IR64, IR71033, IR65482, Sunaulo Sugandha, Anamol Masuli, Khumal-4, IRBB-60, Loktantra, Sugandha-1), and columns the physical position. Each cell represents an interval of 0.5 Mbp.

Figure. 2 Distribution of previously existing rice KASP markers polymorphic between each rice line and the indica reference genome. Rows represent the chromosomes, subdivided into the different lines in the order indicated on chromosome 12 (from top to bottom: IR64, IR71033, IR65482, Sunaulo Sugandha, Anamol Masuli, Khumal-4, IRBB-60, Loktantra, Sugandha-1), and columns the physical position. Each cell represents an interval of 0.5 Mbp.

Tables

Table 1 Categorisation of variations suitable as KASP markers identified between each of the nine sequenced rice lines and the indica reference genotype.

Line	SNPs						InDels				
	Intergenic	Intron	Exon				Intergenic	Intron	Exon		
			Nonsynonymous*	Synonymous	Unknown**	Ratio of Nonsyn/syn			Frameshift*	Inframe*	Ratio of FS/non-FS
IR64	276,103	36,791	25,280	11,946	1,174	2.12	28,831	5,632	1,808	782	2.31
IR71033	214,507	29,360	20,848	9,703	995	2.15	26,527	5,007	1,744	678	2.57
IR65482	316,846	41,673	29,910	14,158	1,554	2.11	34,287	6,527	2,000	949	2.11
Sunulo-Sugandha	326,995	43,021	29,492	14,084	1,336	2.09	32,242	6,267	1,868	924	2.02
Anmol-Masuli	306,934	40,889	28,462	13,469	1,375	2.11	33,241	6,486	1,955	958	2.04
Khumal-4	260,116	34,685	23,594	11,057	1,175	2.13	30,475	5,803	1,842	825	2.23
IRBB-60	217,103	30,887	21,796	10,245	992	2.13	27,830	5,358	1,796	750	2.39
Loktantra	306,922	39,801	27,752	13,149	1,307	2.11	35,428	6,622	2,079	941	2.21
Sugandha-1	233,949	31,956	22,994	10,718	1,143	2.15	29,016	5,483	1,842	815	2.26
Mean of nine lines	273,275 (70.4%)	36,563 (9.4%)	25,570 (6.6%)	12,059 (3.1%)	1,228 (0.3%)	2.12	30,875 (8.0%)	5,909 (1.5%)	1,882 (0.5%)	847 (0.2%)	2.24

*Nonsynonymous SNPs and all InDels within exons are assumed to be functional markers.

**SNPs within the coding regions of annotated genes were categorised as unknown if the corresponding amino acid could not be determined with certainty due to the presence of ambiguous bases

Table 2 Number of informative markers for each pairwise comparison of the nine sequenced rice lines and the *indica* reference genotype.

	IR64	IR71033	IR65482	Sunaulo Sugandha	Anamol Masuli	Khumal-4	IRBB-60	Loktantra	Sugandha-1	Indica
IR64		361	413	456	377	511	382	492	441	480
IR71033	245,367 (7.8%)		419	453	470	434	345	453	386	377
IR65482	355,518 (7.4%)	322,602 (7.5%)		503	488	473	430	442	469	490
Sunaulo Sugandha	444,337 (7.2%)	418,294 (7.3%)	511,006 (7.1%)		520	497	440	522	485	486
Anamol Masuli	286,304 (7.5%)	342,841 (7.5%)	403,027 (7.2%)	493,297 (7.1%)		474	503	391	428	491
Khumal-4	376,321 (7.4%)	323,346 (7.5%)	397,553 (7.2%)	481,381 (7.2%)	387,264 (7.3%)		467	473	426	433
IRBB-60	328,293 (7.4%)	273,578 (7.5%)	397,538 (7.2%)	407,849 (7.3%)	404,343 (7.2%)	369,498 (7.4%)		452	441	392
Loktantra	362,689 (7.7%)	346,699 (7.7%)	385,651 (7.5%)	460,348 (7.4%)	332,649 (7.6%)	391,608 (7.5%)	378,459 (7.5%)		407	496
Sugandha-1	328,829 (7.5%)	274,529 (7.7%)	385,646 (7.2%)	465,745 (7.1%)	356,552 (7.2%)	330,285 (7.4%)	345,187 (7.5%)	361,699 (7.4%)		421
Indica	388,347 (9.5%)	309,369 (11.0%)	447,904 (9.8%)	459,229 (9.1%)	433,769 (9.8%)	369,572 (10.5%)	316,757 (11.3%)	434,001 (10.4%)	337,913 (11.0%)	

Numbers in the **lower-left diagonal (shaded)** correspond to counts of potential new informative KASP markers identified in this study based on SNPs, with percent of InDels shown in brackets.

Numbers in the **upper-right diagonal** correspond to counts of informative markers from the existing set of 1,890 KASP markers that could be aligned against the *indica* reference. All existing informative markers are SNP-based.

Table 3 New validated KASP assays available from LGC genomics (for sequences see Table S7).

							Expected alleles								
ID	Indica position	Japonica position	Variation type	Met filtering criteria?	Target	Reference allele	IR64	IR71033	IR65482	IRBB-60	Loktantra	Sunaulo-Sugandha	Anamol-Masuli	Sugandha-1	Khumal-4
bu0000001	1:17107691	11:26052781	Non-synonymous SNP	Yes	Background	T	C	C	C	C	C	T	T	T	T
bu0000002	1:43712357	11:25968298	Intergenic SNP	Yes	Background	A	G	G	G	G	G	A	A	A	A
bu0000003	2:7181457	11:28006481	Non-synonymous SNP	Yes	Xa resistance	A	G	A	G	A	A	G	A	A	G
bu0000004	2:13037890	2:12216235	Intergenic SNP	Yes	RM301	C	C	C	C	C	C	T	C	C	C
bu0000005	3:21352744	3:18993558	Intergenic SNP	Yes	Background	G	A	A	A	A	A	G	G	G	G
bu0000006	4:19577243	4:21625135	Non-synonymous SNP	Yes	Fragrance QTL	G	A	A	A	A	G	A	G	G	G
bu0000007	5:415717	5:437057	Intergenic SNP	Yes	Xa resistance	T	T	T	T	G	T	T	T	T	T
bu0000008	5:416155	5:437499	Unknown SNP	Yes	Xa resistance	T	T	T	T	A	T	T	T	T	T
bu0000009	5:417389	5:438733	Intron SNP	No	Xa resistance	C	C	C	C	T	C	C	C	C	C
bu0000010	5:417820	5:439189	Intron SNP	No	Xa resistance	T	T	T	T	C	T	T	T	T	T
bu0000011	5:7701715	5:7362881	Intergenic SNP	Yes	Background	A	G	G	G	G	G	A	A	A	A
bu0000012	5:19380178	5:18345193	Intergenic SNP	Yes	Background	A	T	T	T	A	T	A	A	A	A
bu0000013	6:11281447	6:10388210	Non-synonymous SNP	Yes	Pi resistance	G	G	A	G	G	G	G	G	G	G
bu0000014	6:11283253	6:10390022	Non-synonymous SNP	No	Pi resistance	T	T	T	G	T	T	T	T	T	T
bu0000015	6:17240520	6:16386769	Intergenic SNP	No	Pi resistance	C	C	C	T	C	T	C	C	C	C
bu0000016	6:17241451	6:16387700	Intergenic SNP	No	Pi resistance	A	A	A	G	A	G	A	A	A	A
bu0000017	6:17243954	6:16391179	Intergenic insertion	No	Pi resistance	-	-	-	CACAATGGAAG	-	CACAATGGAAG	-	-	-	-
bu0000018	6:17961172	11:23639531	Intergenic SNP	Yes	Background	T	C	C	C	T	C	T	T	T	T
bu0000019	7:3573045	7:3678922	Intergenic SNP	No	Xa resistance	G	A	G	G	G	G	A	G	G	G
bu0000020	7:3588154	7:3694327	Intergenic SNP	No	Xa resistance	T	T	T	T	T	C	T	C	C	T
bu0000021	7:14382827	7:15929199	Synonymous SNP	Yes	Xa resistance	C	C	C	C	C	T	C	C	C	C
bu0000022	7:14384019	7:15930391	Non-synonymous SNP	No	Xa resistance	A	A	A	A	A	G	A	G	A	G
bu0000023	7:14384210	7:15930582	Synonymous SNP	Yes	Xa resistance	A	A	A	A	A	T	A	T	A	T
bu0000024	8:5379548	8:5115025	Synonymous SNP	Yes	Pi resistance	A	G	G	G	G	A	G	A	G	G
bu0000025	8:8486583	8:7832567	Intergenic SNP	Yes	Background	C	T	T	T	T	T	C	C	C	C
bu0000026	8:11193818	8:18168439	Intergenic SNP	Yes	Background	C	T	T	T	T	T	C	C	C	C
bu0000027	8:21701896	8:20380804	Intron deletion	Yes	Fragrance QTL	TG	TG	TG	TG	TG	TG	-	TG	TG	TG
bu0000028	8:21701975	8:20380883	Intron SNP	Yes	Fragrance QTL	T	T	T	T	T	T	C	C	C	C
bu0000029	8:21704520	8:20383435	Intron SNP	Yes	Fragrance QTL	C	C	C	C	C	C	T	T	T	T
bu0000030	8:28422597	8:26729241	Intergenic SNP	Yes	Xa resistance	T	G	G	T	T	T	G	G	G	G
bu0000031	9:7018065	9:7513604	Intergenic SNP	Yes	Background	C	C	C	C	C	C	G	G	G	G
bu0000032	9:7725492	11:24474192	Intergenic SNP	Yes	Background	C	T	T	C	T	T	C	C	C	C
bu0000033	10:15236821	10:16682028	Intergenic insertion	Yes	Background	-	G	G	G	G	G	-	-	-	-
bu0000034	11:5950201	11:6605583	Synonymous SNP	Yes	Xa resistance	A	A	A	A	A	T	A	A	A	A
bu0000035	11:6033817	11:6658350	Intron SNP	Yes	Xa resistance	G	G	G	G	G	A	G	G	G	G
bu0000036	11:17838940	11:21047256	Intergenic SNP	Yes	Xa resistance	T	A	A	T	T	T	T	A	T	T
bu0000037	11:19964533	11:24664749	Synonymous SNP	Yes	Xa resistance	G	G	G	G	G	A	G	G	G	G
bu0000038	11:20939753	11:24249679	Intergenic SNP	Yes	Background	C	T	T	T	T	T	C	C	C	C
bu0000039	12:8644297	12:10835433	Intergenic SNP	Yes	Background	G	C	G	C	G	C	G	G	G	G

Positions are based on the *indica* ASM4565v1 and *japonica* IRGSP-1.0 reference genomes. Linkage analysis is underway to assign linkage to traits in relevant crosses; preliminary data for IR64 x Jumli Marshi shows that novel 24 is associated with field resistance to BLB locus Pi33 (χ^2 = 29.6, P <0.01). Shading shows an example of a cross in which the KASP is being used for selection.

Supplementary Files

File. S1 KASP marker design sequence generation software.

Supplementary Figures

Figure. S1 Overview of the criteria used for identification of potential KASP markers from variations identified using SAMtools.

Figure. S2 Number of variations identified at the same positions relative to the indica reference in all sequenced rice lines (maximum nine).

Figure. S3 Number of variations (homozygous SNPs, insertions and deletions) identified between each of the nine sequenced rice lines and the indica reference genome.

Figure. S4 Distribution of potential new rice KASP markers polymorphic between each rice line pair. Rows represent the chromosomes, subdivided into the different lines (ordered as indicated on chromosome 12), and columns the physical position. SubFigures show the distribution of markers informative for crosses against (a) IR64 (b) IR71033 (c) IR65482 (d) Sunaulo Sugandha (e) Anamol Masuli (f) Khumal-4 (g) IRBB-60 (h) Loktantra (i) Sugandha-1.

Figure. S5 Distribution of existing rice KASP markers polymorphic between each rice line pair. Rows represent the chromosomes, subdivided into the different lines (ordered as indicated on chromosome 12), and columns the physical position. SubFigures show the distribution of markers informative for crosses against (a) IR64 (b) IR71033 (c) IR65482 (d) Sunaulo Sugandha (e) Anamol Masuli (f) Khumal-4 (g) IRBB-60 (h) Loktantra (i) Sugandha-1.

Figure. S6 Distribution of distances between consecutive informative KASP markers, for new and existing markers, for all combined pairwise combinations of rice lines used in this study. Vertical bars represent the medians with boxes extending from the 25th to 75th percentiles. Whiskers extend from the 5th to 95th percentiles, dots represent the minimum and maximum distances.

Supplementary Tables

Table S1 Details of nine *indica* rice cultivars and breeding lines used for whole genome NGS.

Table S2 Mapping rates, genome coverage and read depth of the nine sequenced rice lines, for all mapped reads and for uniquely-mapped reads only. Only those reads with both pairs aligning in the expected orientation were included.

Table S3 Distribution of distances (bp) between consecutive informative existing KASP markers (previously developed using chip-based technology) for all rice line pairings.

Table S4 Distribution of distances (bp) between consecutive informative potential new KASP markers (developed here using NGS data) for all rice line pairings.

Table S5 Summary of KASP assay validation results. Counts are given of the number of unique marker-cross combinations that have been validated (or otherwise), split according to whether or not the markers met our bioinformatics filtering criteria, and between new and existing markers.

Table S6 Details of marker-cross combinations tested for validation of 46 new and 75 existing KASP assays (N.B. some failed assays may be due to lack of polymorphism.)

Table S7 Sequences for new validated KASP assays available from LGC genomics.