International Journal on Advanced Science Engineering Information Technology

Vol.11 (2021) No. 2
ISSN: 2088-5334

Leveraging the 3000 Rice Genome Data for Computational Design of Polymorphic Markers in a Local Rice Variety Lacking Sequence Data

Dani Satyawan[a,*], Ahmad Warsun[a], Ahmad Dadang[a], Muhamad Yunus[a]

[a]*Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development (ICABIOGRAD-IAARD), Bogor,16111, Indonesia*

*Corresponding author: *d.satyawan@gmail.com*

Abstract— **DNA markers can detect DNA sequence variations in the genome, and they are useful for genetic studies, DNA fingerprinting, and genotype-based selection in breeding programs. Rice, as one of the model plants for genetic and genomic studies, has abundant DNA markers stored in various online databases. Selecting markers in rice is not limited by marker availability but rather by their polymorphism in the target population. We developed a computational method to screen millions of single nucleotide polymorphism (SNP) markers listed in IRRI 3000 rice genome database in order to find a subset of markers that are polymorphic in an F2 mapping population created from a cross between a parental line with a known genome sequence and a local Indonesian variety with no genome sequence data. The parental lines were genotyped using an affordable medium-density SNP array. The genotype data was cross-referenced with the rice genome database to perform phylogenetic analysis and identify accessions clusters with the highest genetic similarities to each parental line. The cluster data was then used to identify monomorphic SNP candidates within the cluster but exhibit consistent polymorphism between the two clusters. Using this method, we obtained a SNP marker set for a segment in rice chromosome 8 with 76.19% polymorphism rate, which is much higher than the expected 1.06% polymorphism rate if the SNP markers were chosen randomly. The improved polymorphism rate was also observed when the method was applied to other random chromosome segments and randomly chosen parental candidates.**

Keywords— **Rice; SNP marker; computational marker design; genome sequence utilization.**

Manuscript received 14 Jul. 2020; revised 15 Dec. 2020; accepted 22 Feb. 2021. Date of publication 30 Apr. 2021.

I. INTRODUCTION

Genetic markers are essential tools in genetic studies since they can assess the variable regions in the genome. They have been utilized to find the location of DNA sequences that contribute to superior traits through linkage mapping [1] and association mapping [2], trace crop evolution [3] and migration [4], and develop fingerprinting tools for varietal identification [5]. The underlying technology to visualize DNA variation has evolved from simple morphological markers, isozymes, restriction fragment length polymorphisms to polymerase chain reaction (PCR) based markers such as randomly amplified polymorphic DNA and simple sequence repeats [6].

As genetic studies started to examine finer details of the genome structure, the limited availability of usable markers in various genomic regions became an issue [7]. The emergence of whole-genome sequencing technology has helped to solve this problem, as virtually all DNA sequences in the genome of individuals of interest can now be read and compared. Consequently, most polymorphic loci in the genome can be detected with a high degree of precision. Nevertheless, even though the cost of whole-genome sequencing continues to decline, the technology is not always economically viable for regular uses in genetic studies and breeding programs [8]. Thus, cheaper genetic markers that are sufficiently ubiquitous in the genome are still indispensable for genetic studies.

Single nucleotide polymorphism (SNP) markers fulfill this need, as they are prevalent in the genome and can be assayed economically using various methods [9]. SNP genotyping techniques for simultaneous analysis of a large number of SNPs exist, and simpler methods that utilize PCR amplification followed by visualization in agarose gels are also available. The total cost and SNP genotyping methods can thus be optimized according to the available budget and the number of SNPs to be genotyped.

The declining cost of whole-genome sequencing has enabled many institutions and research groups to sequence the genomes of various crops and share the resulting DNA variation data in publicly accessible databases, which can be used as a rich source of DNA markers. One of those databases is the IRRI SNP-Seek database [10], which was constructed from the whole genome sequence data of 3243 rice accessions in IRRI collection [11]. Various groups sequenced an additional 3982 rice accessions, and the SNPs were also made available to the public [12]. Together, these resources can help rice researchers obtain DNA markers for most regions in the rice genome.

Nevertheless, those abundant DNA markers are unusable if they are monomorphic in the intended sample population. To obtain polymorphic markers, marker candidates are usually screened first in a subset of the target population to eliminate monomorphic markers. This screening stage often requires considerable investments in terms of time, labor, and reagents. Thus, there is a need for inexpensive techniques that facilitate quick identification of polymorphic markers for the target samples in research and breeding programs that utilize DNA markers. Here we report a method that utilizes the SNP data generated by the 3000-rice genome project to design SNP markers for a target region with 76% polymorphism rate in an F2 population. The population was developed to map the gene or quantitative trait loci (QTL) that are responsible for resistance to brown planthopper, by crossing an elite variety and a local variety that has not been sequenced. A major QTL was successfully mapped using 7K Infinium SNP genotyping [13] to a 1.833 Mb chromosomal segment in rice chromosome 8, which contains 236 genes [14].

A fine-mapping study using more markers in the target region to identify the causal gene of the QTL was planned, and we devised a computational screening method to improve the polymorphism rate of the marker selection step. We used the genotypic data of the parental lines from the 7K SNP data to deduce their phylogenetic relationship with the 3000 rice accessions sequenced by IRRI. Accessions with the highest genetic similarity to one of the parents were grouped together to be compared to the other group comprising accessions most similar to the other parent. We demonstrated that by selecting markers in the IRRI SNP database that exhibited consistent polymorphism between groups but are monomorphic among accessions within the same group, we could improve the polymorphism rate of selected markers significantly higher level compared to random marker selections.

II. Materials and Methods

A. Plant Materials

This study's plant materials are a segregating F2 population from a cross between TN1 and Untup Rajab, a local Indonesian rice variety. Leaf samples were taken from both parents and individual F2 plants for DNA extraction and marker analysis. DNA extraction was performed according to the methods of Dellaporta et al. [15]. Initial SNP genotyping of both parents was performed using rice SNP 7K Infinium genotyping assay [16] at IRRI service laboratory using their standard operating procedures.

B. Extraction of 3000 Rice Sequence Data for Phylogenetic Analysis

The 7K SNP coordinate data was extracted from coordinate data supplied by IRRI. To reduce computational burden in the downstream analyses, the coordinates were further filtered to include only those located in chromosome 8, since the QTL interval of interest is located in that chromosome. Those coordinates were then used to extract the SNPs of interest from the 3000 rice whole-genome SNP data, which were downloaded from IRRI [17] and contained 29 million SNPs in PLINK format. SNP extraction was carried out using PLINK software [18], using the *--extract* command.

C. Phylogenetic Analysis

The extracted SNP data from the 3000 sequenced accessions and the two parental lines were converted to DarWin [19] input format. Using DarWin, phylogenetic analysis was carried out by first calculating dissimilarities from allelic data using the following equation:

$$d_{ij} = 1 - \frac{1}{L}\sum_{l=1}^{L}\frac{m_l}{\pi}$$

Where d_{ij} : dissimilarity between units i and j, L : number of loci, π : ploidy, and m_l : number of matching alleles for locus l. The resulting dissimilarity vector was then used to construct a phylogenetic tree using weighted neighbor-joining procedure. The tree was subsequently exported in newick format, to be drawn and annotated using iToL [20].

D. Extracting and Selecting Markers from Closely Related Rice Accessions

By visualizing the resulting phylogenetic tree, six accessions closest to TN1 and six accessions closest to Untup Rajab were identified. All of their DNA variants in the region of interest in chromosome 8 were extracted from the original 29 million SNP data using PLINK *--chr* command along with *--from-kb* and *--to-kb* commands. After filtering for variants located in the region of interest (Rice Chromosome 8, base 24,205,833 to 26,038,950), the SNP data was then loaded to Microsoft Excel. By grouping the accessions into two clusters, i.e., the Untup Rajab group and TN-1 group, variants that were monomorphic within the group but polymorphic between the group were identified using "=IF" function in Microsoft Excel. Each selected SNP's coordinates were then recorded and intersected with genic coordinate data to identify SNPs located in coding regions using Bedtools' intersectBed command [21].

E. Marker Design and Assay

The selected SNP coordinates were used to extract the DNA sequence data from the reference genome. A total of 100 bp of sequences upstream and downstream of the target SNP were extracted using fastaFromBed command in Bedtools, and submitted for KASP genotyping assay [22] to analyze the genotypes of the SNPs in the segregating F2 population. Raw genotype data was viewed using SNPviewer (v. 2, KBioscience), while the graphical representation of the SNP alleles was visualized using GGT 2.0 [23].

F. Performance Verification

Applicability of the SNP selection method was tested on a random chromosome segment chosen using a random number selector [24] and the same phylogenetic tree generated in this study. For this test, IRRI's TN1 and Sossoka Oule accessions were chosen to substitute our TN1 and Untup Rajab since they belong to the same cluster and their whole genome sequences were available to check the polymorphism of selected SNPs. The number of related accessions per cluster was varied from five to two, to observe the effect of cluster size on SNP polymorphism prediction accuracy. False positives were defined as SNPs predicted to be polymorphic but found to be monomorphic in TN1 and Sossoka Oule, while false negatives were predicted to be SNPs monomorphic but are polymorphic in the two accessions. The test was repeated on two randomly selected accessions (Ampipit and ITA 117), but this time the related accessions clusters were chosen based on IRRI's phylogenetic tree [17] to see the effects of higher quality phylogenetic tree on prediction accuracy. The targeted chromosome segment was again chosen using a random number selector.

III. Results and Discussions

A. Phylogenetic Analysis

The IRRI 7k SNP chip contained 7098 markers, 1046 of which were polymorphic between TN-1 and Untup Rajab. Among the 7098 SNPs, 545 were located in chromosome 8, where the target QTL interval was located [14]. There were 511 SNPs in chromosome 8 intersected with whole-genome SNP data extracted from the IRRI 3000 genome project.

Based on those 511 markers from chromosome 8, a neighbor-joining phylogenetic tree was constructed (Fig. 1). According to the tree, the accessions closest to TN-1 belong to ind1A or indx subpopulation and dominated by accessions from the Indian subcontinent. The tree also correctly placed the TN-1 genotyped in this study close to the TN-1 that IRRI sequenced. Slight differences found between the two TN-1 accessions could result from variant-calling errors [25], sequencing errors [26], or newly-arising mutations [27]. Untup Rajab is clustered with accessions from ind2 subpopulation, although one of the closely related accessions belongs to indx subpopulation. All six genetically similar accessions to Untup Rajab originated from Africa, which is surprising since Untup Rajab is classified as a local landrace in Indonesia.

Genetic dissimilarities between Untup Rajab and its most similar accessions were higher than the dissimilarities observed in TN-1 cluster (Table 1). TN-1 was one of the earliest semi-dwarf elite varieties that later brought about the green revolution. It was popular in India in the 1960's and it was commonly used as one of the parents for developing new varieties [28]. This could explain the relatively low dissimilarity observed in the TN-1 cluster. On the other hand, Untup Rajab is an Indonesian landrace. Thus any similarities with other rice accessions from different continents in the cluster could be due to a common ancestor that was exchanged or introduced a long time ago, after which each accession had the chance to mutate and evolve when adapting to a new location [29].

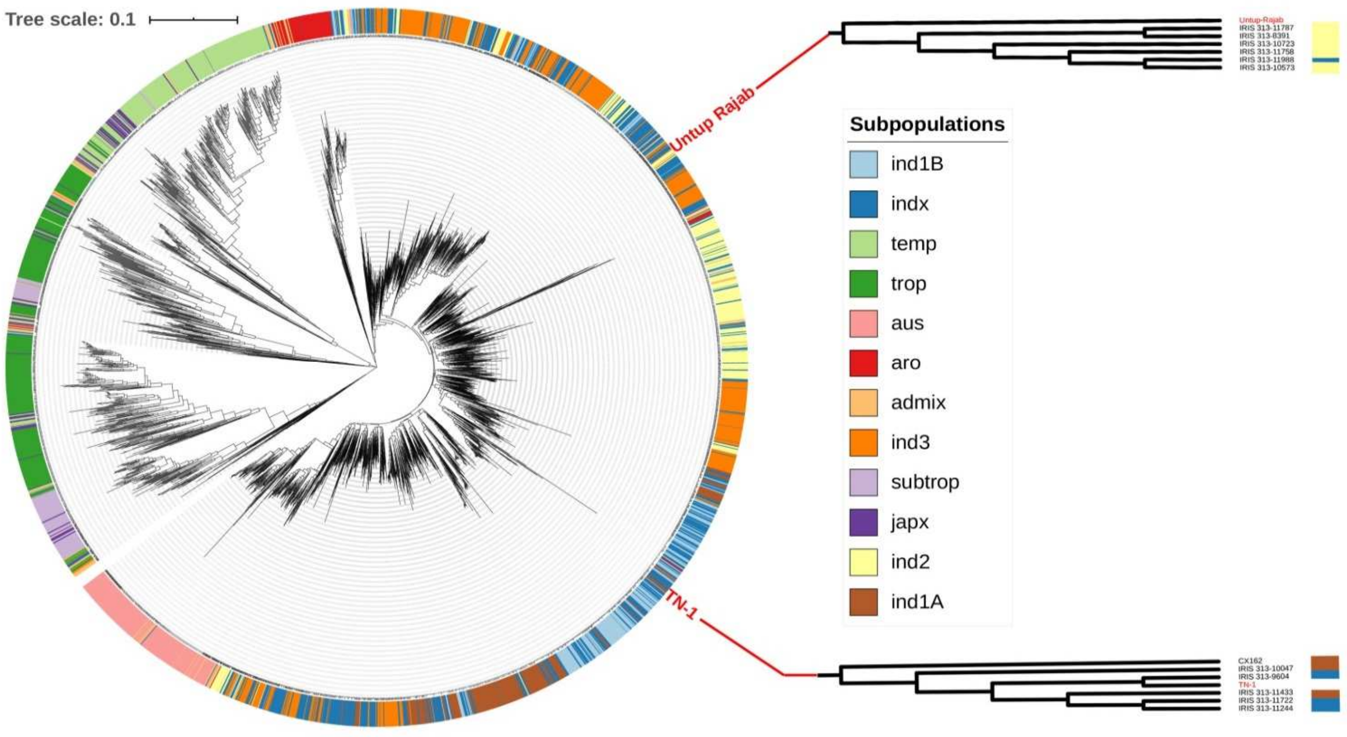

Fig. 1 Location of the two parental lines relative to the 3000 rice accessions sequenced by IRRI in a phylogenetic tree constructed from 511 SNP data in chromosome 8. The colored bars indicate each accession's subpopulation membership specified by IRRI. The twelve accessions that were most similar to either parent were shown in the two zoomed in cladograms in the right, along with their subpopulation membership.

TABLE I

IRRI ACCESSIONS WITH THE HIGHEST GENETIC SIMILARITY TO THE PARENTAL LINES

Accession Name	Accession Code	Cluster	Subpopulation	Origin	Genetic Dissimilarity
TN-1	CX162	TN-1	ind1A		0.0142
EX MARABA GURUKU	IRIS 313-10047	TN-1	ind1A	Nigeria	0.0132
W 1014-162	IRIS 313-11244	TN-1	indx	India	0.0142
MR 136-1	IRIS 313-11433	TN-1	ind1A	India	0.0143
KABERI	IRIS 313-11722	TN-1	indx	Bangladesh	0.0141
RPA 5929 (K 45)	IRIS 313-9604	TN-1	indx	India	0.0142
SOSSOKA OULE	IRIS 313-10573	Untup Rajab	ind2	Mali	0.0478
MOND BA	IRIS 313-10723	Untup Rajab	ind2	Senegal	0.0628
FOFFI	IRIS 313-11758	Untup Rajab	ind2	Cote d'Ivoire	0.0599
TANDAKAY FINGO	IRIS 313-11787	Untup Rajab	ind2	Gambia	0.0702
POTOQOIN	IRIS 313-11988	Untup Rajab	indx	Sierra Leone	0.0569
MARAGBE	IRIS 313-8391	Untup Rajab	ind2	Burkina Faso	0.0546

B. *Marker Selection*

The allelic dissimilarity between TN-1 and Untup Rajab was 0.117. Among the 1022 alleles used to create the phylogenetic tree, 106 were polymorphic between TN-1 and Untup Rajab, while 800 were monomorphic, and the rest had missing data. Therefore, based on these observed alleles, the proportion of polymorphic alleles in chromosome 8 was 11.7%. Assuming that the alleles were representative samples for the whole chromosome 8, selecting random 42 SNPs across the chromosome on average will only result in 5 polymorphic SNPs, which is inadequate for fine-mapping studies.

To obtain SNP marker candidates, we downloaded and extracted all SNP data that are located within the target QTL interval from the 12 accessions listed in Table 1. The targeted interval lies between base 24,205,833 to 26,038,950 in rice chromosome 8. Among the 146,814 biallelic SNPs within the interval, 1,670 were identified as consistently polymorphic between TN-1 cluster and Untup Rajab cluster. Consistent polymorphism is defined as monomorphism within cluster members but polymorphic between different clusters. This chromosome segment's observed polymorphism rate is only 1.06% , much lower than the 11.7% predicted from the 7K SNP chip results. Thus, picking 42 random SNPs across the interval is not likely to produce even a single polymorphic SNP in the marker set.

The lower polymorphism rate could be due to the fact that the markers in the 7K rice SNP chip were chosen based on their likelihood to be polymorphic among different rice subpopulations [30]. Hence, they are not representative samples of the typical SNPs found in most intervals within the rice genome. It is also possible that our target interval contains more conserved regions among indica rice, which typically have a lower-than-average polymorphism rate. However, such an assertion needs to be verified by examining the polymorphism rate across the genome.

The 1,670 consistently polymorphic SNP candidates were then annotated to identify SNPs that intersect with genic regions. They are also binned according to their position to ensure even and equidistant representation within the QTL interval. A final set of 48 SNPs were chosen to represent each bin at roughly similar distance from other SNPs, and whenever possible genic SNPs were prioritized in each bin, since genic SNPs are more likely to affect phenotypes and likely to be conserved within a cluster (Table 2).

C. *KASP Assay*

SNP genotyping for the new set of SNPs was performed based on Competitive Allele-Specific PCR or KASP [31], since it had been verified as a reliable and cost-effective method for genotyping a smaller number of SNPs in small population size [32]. Among the 48 SNPs submitted to the genotyping service provider, 42 SNPs were suitable for KASP primer design and genotyping assays. Genotyping assays of those 42 SNPs revealed that 32 SNPs were polymorphic in the sample F2 population, which consisted of 192 individuals from a cross between Untup Rajab and TN-1. Among the ten monomorphic SNPs, seven SNPs only produced the alleles from Untup Rajab, while the remaining three SNPs only had TN-1 alleles. The monomorphic SNPs also overwhelmingly favor the G/C alleles than A/T alleles, with a ratio of 4:1 (Table 2).

The SNP assay had a high success rate, with only 2% of uncalled genotypes recorded. Some of those uncalled genotypes can be called with high confidence upon consulting the raw signal data (Fig 2A). Only one plant sample had an unusually high proportion uncalled genotype, as 19 of the 42 SNPs could not be genotyped in this sample, while most had better SNP call rate with 85 samples reported 100% SNP calls and 74 samples only had one missing SNP call. Each SNP marker on average had four uncalled genotypes when it was used to assay the 192 samples. Two of the monomorphic SNP markers (Chr8_24848984 and Chr8_25915452) had unusually high uncalled genotypes, which were observed in around 25% of the assayed samples. This indicates that those locus may not be suitable for primer design and KASP assay. One possible reason is that two of the primers used in KASP assays must have 3' ends that contain the SNP of interest [31], hence there are not enough flexibilities in choosing the flanking DNA segment that can be used to design the primers. If the SNP sequences contain some features that inhibit PCR, such as unusual GC content and hairpins [33], it will be difficult to design an effective primer for such locus. It is also possible that primers designed from such locus may be more effective for one allele and less so for the other allele [34], creating an illusion of monomorphism or skewed distribution of alleles.

TABLE II
PROPERTIES OF SELECTED SNP IN THE TARGET INTERVAL

Chromosome	SNP Position	Polymorphism	Notes
Chr8	24255130	(A/G)	Polymorphic
Chr8	24286188	(C/T)	Polymorphic
Chr8	24300914	(C/T)	Polymorphic
Chr8	24324859	(A/G)	Excluded
Chr8	24367765	(A/G)	Polymorphic
Chr8	24417814	(A/G)	Polymorphic
Chr8	24446073	(A/G)	Polymorphic
Chr8	24494660	(A/T)	Polymorphic
Chr8	24525162	(A/G)	Excluded
Chr8	24580101	(G/T)	Polymorphic
Chr8	24626816	(A/G)	Polymorphic
Chr8	24640146	(A/G)	Polymorphic
Chr8	24741960	(A/C)	Monomorphic
Chr8	24769690	(A/G)	Polymorphic
Chr8	24848984	(G/T)	Monomorphic
Chr8	24887322	(A/T)	Polymorphic
Chr8	24937118	(A/G)	Polymorphic
Chr8	24967388	(A/G)	Polymorphic
Chr8	25000189	(A/G)	Polymorphic
Chr8	25048975	(C/T)	Excluded
Chr8	25071564	(C/T)	Monomorphic
Chr8	25116271	(C/T)	Polymorphic
Chr8	25125023	(C/T)	Monomorphic
Chr8	25163997	(C/T)	Polymorphic
Chr8	25208980	(A/G)	Polymorphic
Chr8	25230553	(C/T)	Monomorphic
Chr8	25271327	(C/G)	Polymorphic
Chr8	25287787	(C/T)	Monomorphic
Chr8	25338637	(C/T)	Monomorphic
Chr8	25401336	(A/G)	Polymorphic
Chr8	25408352	(A/G)	Polymorphic
Chr8	25440564	(A/G)	Polymorphic
Chr8	25482037	(G/T)	Polymorphic
Chr8	25505323	(G/T)	Excluded
Chr8	25539507	(C/T)	Excluded
Chr8	25591272	(A/G)	Polymorphic
Chr8	25610820	(A/C)	Excluded
Chr8	25663664	(A/G)	Polymorphic
Chr8	25677006	(A/T)	Polymorphic
Chr8	25713549	(C/T)	Polymorphic
Chr8	25735605	(A/G)	Polymorphic
Chr8	25762477	(A/T)	Polymorphic
Chr8	25835547	(G/T)	Monomorphic
Chr8	25915452	(A/G)	Monomorphic
Chr8	25932039	(C/T)	Polymorphic
Chr8	25972248	(A/T)	Polymorphic
Chr8	25986057	(C/T)	Monomorphic
Chr8	26035154	(C/T)	Polymorphic

To test that hypothesis, we examined the sequences flanking the target SNPs using the PCR Primer Stats utility in the Sequence Manipulation Suite website [35]. The result indicated that one of the unusual monomorphic primers had low GC content and the alternative site had a high melting temperature. However, such problems can also be found in some of the polymorphic markers, which shows that they are not always critical in KASP assays. The other unusual monomorphic primer passed all the tests, which means that there should be no problem should that locus is used to design a PCR primer. Thus, factors other than the common ones known to hinder PCR were responsible for the unusually high fraction of uncalled SNPs in the two monomorphic markers.

Except for the monomorphic SNPs, most SNPs did not significantly deviate from the 1:2:1 Mendelian ratio for segregation in an F2 population when tested using Chi-Squared Tests. Only one polymorphic SNP marker (Chr8_24367765) significantly deviated from the 1:2:1 ratio, and the observed ratio between alleles of Untup Rajab, heterozygotes, and TN-1 was 45:136:11 respectively. Such ratio indicates that some of the homozygotic TN-1 alleles were scored as heterozygotes. Inspection of the original fluorescent data revealed that the heterozygote spots formed a continuous distribution with the TN-1 spots, making even manual allelic determination difficult (Fig. 2B). Whether this is due mainly to suboptimal primer design or other factors is currently unknown. Allelic data from this particular SNP, especially the heterozygotes and TN-1 alleles, should be cross-referenced with data from the adjacent SNPs to ensure the correct genotype call for this SNP.

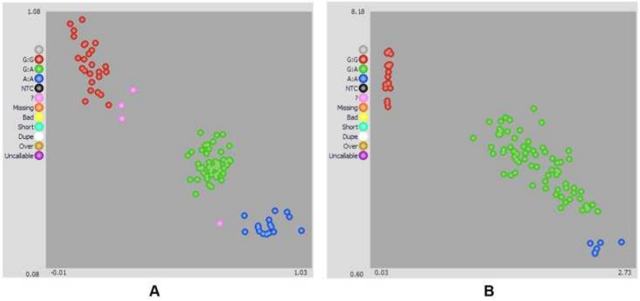

Fig. 2 Fluorescence data of the KASP genotyping assay, with the parental alleles represented by red and blue dots, while the green dots are heterozygotes and pink dots are SNP alleles that could not be resolved by the automatic SNP caller.

After the monomorphic SNPs were excluded and missing data were manually verified from the original fluorescence data, the final genotype calls were used to identify recombinant break points in the region of interest (Fig 3). Individuals with varying allelic distribution within the interval will be included in a brown planthopper resistance test. By correlating an allele's presence, which represents a smaller interval within the QTL, with resistance to BPH, the segment that confers resistance to BPH can be identified. By narrowing down the segment, the number of genes and mutations that need to be considered is also reduced, thus improving the probability of identifying the underlying mutation that confers resistance to BPH.

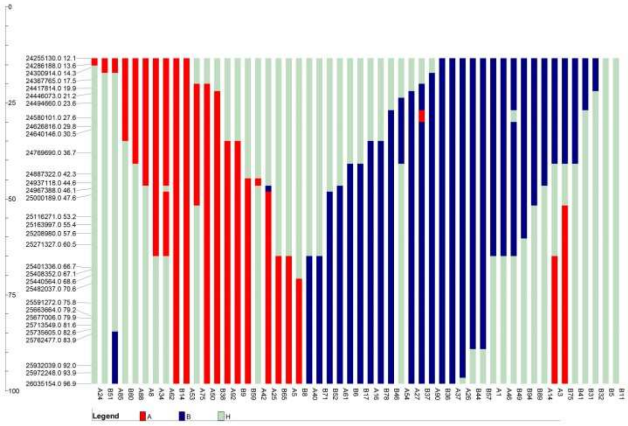

Fig. 3 Patterns of recombination break points in selected F_2 progenies. Green bars denote heterozygous segments, while red bars represent homozygous TN1 alleles and blue bars represent homozygous Untup Rajab alleles.

D. Phylogeny-Based Preliminary Marker Screening Significantly Improved Polymorphic Marker Detection

We have demonstrated that phylogeny-based pre-screening can significantly improve the probability of obtaining polymorphic markers. Traditionally, researchers who utilize DNA markers in their studies or breeding programs collect a large number of DNA markers. Those DNA markers will need to be surveyed using parental lines or a sample of the intended population to select polymorphic markers in the target population [36]. Thus, some of the collected markers may never be used because they are monomorphic in all the populations used in the collector's projects.

As the need for high-density markers like SNP arises, SNPs become more commonly used as they are very common in the genome and can be scaled up economically. However, SNPs also typically have lower polymorphism information content (PIC) compared to the other commonly used DNA marker, the simple sequence repeat (SSR) markers [37]. Thus, the probability of a marker being monomorphic is higher in SNPs than SSRs. The reason could be because SNPs are usually biallelic [38], which means that they only have two possible alleles in each marker.

Without pre-screening, the probability of obtaining a polymorphic marker in our segregating population within the interval of interest is 1.06%. Thus, a random selection of SNP markers within this interval will yield mostly monomorphic markers. The most thorough method to eliminate monomorphic marker candidates is by whole-genome sequencing since the two parents' whole-genome sequencing can identify all monomorphic and polymorphic loci and all mutations that cause the resistance to BPH. However, this would have incurred high cost and still does not eliminate the need for further SNP assays to fine-map recombination breakpoints in the segregating population and their association with resistance. This is because purely computational and prediction analysis of the candidate gene in the QTL interval is challenging since the interval contains 236 genes, and predictions cannot be reliably made if the type of resistance is novel.

Since the two parents, along with their segregating progenies, had been genotyped with a high-density SNP chip assay for the initial QTL analysis, we hypothesized that there was sufficient genotypic data from the 7K SNP chip assay to deduce the genetic similarities between the parents and the 3000 rice accessions sequenced by IRRI. Using 511 SNP data from the chip, we identified six genetically most similar accessions to TN-1 and six accessions most similar to Untup Rajab. The polymorphism between those accessions and the two parents ranged from 3.77% to 6.58%, with an average of 5.54%. By selecting for SNPs that are polymorphic between clusters but monomorphic within clusters, we obtained a SNP set with a polymorphism rate of 76.19%, much higher than the 1.06% detected inside the QTL region.

We also tested the efficacy of this approach in other parts of the genome and in another population. A 400,000 base pairs segment in chromosome 11 beginning at base 7,968,253 was chosen as the target region using a random number generator to determine a random spot in the genome. Since complete SNP data for Untup Rajab and our TN-1 line were not available for this region to check the accuracy of the polymorphism predictions, we used the genotype data of the 10 high-similarity accessions from both clusters to predict the genotypes of sequenced TN-1 from IRRI and Sossoka Oule from the Untup Rajab cluster. This segment has low DNA variability, since we only obtained 5694 SNPs from the 3k filtered SNP data set in the IRRI database. Among those, 712 were polymorphic between TN-1 and Sossoka Oule, but only nine were predicted to be polymorphic if they also have to be monomorphic in both TN-1 and Untup Rajab clusters (Table 3). Although there were no false positives among the nine SNP candidates, the number of SNP candidates were too low and concentrated only in a 100 kb segment. The number of SNP candidates can be improved if we reduce each cluster's representatives to two accessions. However, the number of false positives (SNPs that will be monomorphic between TN-1 and Sossoka Oule) increased to 19%, which is comparable to what we obtained in our study.

A similar trend was observed when we randomly picked an Indica 1B subgroup and an Indica 3 subgroup to find polymorphic SNP candidates from a random segment in the chromosome. For this, we used the phylogenetic tree constructed by IRRI using a higher number of SNPs than ours [17]. Using polymorphism data from five accessions closest to Ampipit (Indica 3) and five accessions closest to ITA 117 (Indica 1B), we failed to obtain even one SNP candidate from 3726 SNPs in the target interval, even though 591 SNPs were actually polymorphic between Ampipit and ITA 117 (Table 3). Reducing the number of accessions for pre-screening to two accessions from each cluster improved the identification of SNP candidates to 563, at the cost of increased false positives of 3.38% among the SNP candidates. Thus, when a high-quality phylogenetic tree is used, the number of accessions required to predict SNP polymorphism can be safely reduced to merely two accessions from each cluster.

TABLE III

Performance of Phylogeny-Based Polymorphic SNP Selection in a Randomly-Selected Chromosome Interval and Population

Parents	Chromosome Interval (chromosome: start-stop)	SNP Number	Total Poly	Poly in C2	False + in C2	False – in C2	Poly in C5	False + in C5	False – in C5
TN-1 Sossoka Oule	CHR11:7968253-8068252	1600	251	4	2	247	0	0	251
	CHR11:7868253-7968252	1432	397	184	28	213	9	0	397
	CHR11:7768253-7868252	1494	59	25	7	34	0	0	59
	CHR11:7689253-7768252	1168	5	3	5	2	0	0	5
	Total	**5694**	**712**	**216**	**42**	**496**	**9**	**0**	**712**
Ampipit ITA 117	CHR7:5635847-5735846	953	551	532	14	19	0	0	551
	CHR7:5535847-5635846	839	30	29	4	1	0	0	30
	CHR7:5435847-5535846	1114	9	2	2	7	0	0	9
	CHR7:5335847-5435846	820	1	0	0	1	0	0	1
	Total	**3726**	**591**	**563**	**20**	**28**	**0**	**0**	**591**

Codes: Poly = polymorphic SNPs ; C2 = cluster where each parental cluster has 2 accessions; C5 = cluster where each parental cluster has 5 accessions

E. Possible Improvements and Automation

Although the SNP pre-screening method can reduce our fine-mapping program's cost and time, we also noted that this approach currently has some undesirable characteristics that may hinder its wider adoption by other researchers. Firstly, this approach relies on the abundant availability of genomic data in rice, where high-quality reference genome and whole-genome sequence data are already available for thousands of rice accessions that represent the worldwide genetic diversity of rice [39]. Researchers working on other crops may not have such luxury [40], making this approach difficult to apply for such crops.

Another factor that needs to be considered is that this approach requires basic bioinformatics skills to extract the large volume of data provided by IRRI. For this study, we worked with raw data from IRRI using various tools that could be daunting to learn for many researchers. However, a large portion of the procedures can be performed on IRRI's online SNP database, such as the steps of extracting the 7K SNP chip data from 3000 rice accessions and extraction of SNPs from a given interval from a list of rice accessions. The phylogenetic analysis and SNP collection can be easily done in regular windows-based computers commonly used by the general public. It is also possible to develop a web-based software solution, similar to the ones developed by Ha et al. [41] to identify the closest relative by submitting DNA variation data. Screening of common SNPs in a cluster that is polymorphic in a different cluster is similarly amenable to software automation. Thus, it is possible to develop a completely automatic toolset that is easy to use, where users will only need to submit the parental genotype data and the desired genomic location where new SNPs need to be generated to obtain the desired SNP candidates.

Another data component that could be unfamiliar to other researchers is the 7K SNP chip data. The chip contains 7098 SNPs can be used for various purposes such as diversity analysis, mapping, fingerprinting, and genotyping of some known traits. It is offered as a genotyping service by IRRI, and we found that the savings in time and labor during the QTL mapping stage could justify the cost of the service. Among the 7098 SNPs, 1046 were polymorphic between the parents and 885 had proper mendelian segregation and acceptable level of missing data to produce sufficiently dense genetic and QTL map and for phylogenetic analysis in this study. It is possible to use other genotyping methods such as SSR to be used as the initial data for phylogenetic analysis, provided that sufficient markers are assayed to obtain a reasonably accurate phylogenetic tree. However, it will be more difficult to cross-reference the allelic data with the 3000 rice genome data, as they need to be converted to insertion/deletion data using indel data convention used by IRRI [42].

The type of SNP assays that will be used for the selected SNPs should also be considered to obtain optimal results. For example, our assay of choice was KASP, which is based on PCR. Consequently, SNP candidates' choice should account for whether the flanking sequences are suitable for primer design. Therefore, extreme GC content, self-complementary sequences, and duplication in the genome must be avoided when the final SNP set is selected among the candidates. Other assay types such as Infinium and axiom arrays are not suitable for G/C or A/T polymorphism [43], [44], so they need to be eliminated from the pool of SNP candidates if the resulting SNPs will be assayed using such methods.

IV. Conclusion

This study has validated the efficacy of candidate SNP identification based on phylogenetic and whole-genome sequence data to increase the likelihood of obtaining polymorphisms in segregating populations significantly. The screening system produced 76.19% observed polymorphism in our F2 population, which was much higher than 1.06% expected polymorphisms from random SNP selection. The use of higher quality phylogenetic trees, which are influenced by the number of markers to generate the tree, can improve the screening process since fewer accessions can be used for SNP screening, and the resulting candidates have less monomorphic SNPs. The screening method does not require high computing power or advanced bioinformatics skills. It is amenable to automation so a more user-friendly software can be developed to accommodate more users interested in utilizing genomic information for genetic studies plant breeding.

Acknowledgment

The authors are grateful to Fajar Suryawan and Riko Harmando for their assistance in the sample collection stage. This study is funded by DIPA BB Biogen in the 2018 fiscal year.

References

[1] X. Zhang *et al.*, "Combining QTL-seq and linkage mapping to fine map a wild soybean allele characteristic of greater plant height," *BMC Genomics*, vol. 19, no. 1, p. 226, Mar. 2018.

[2] M. P. M. Thoen *et al.*, "Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping," *New Phytol.*, vol. 213, no. 3, pp. 1346–1362, Feb. 2017.

[3] M. Schreiber, N. Stein, and M. Mascher, "Genomic approaches for studying crop evolution," *Genome Biology*, vol. 19, no. 1. BioMed Central Ltd., pp. 1–15, 21-Sep-2018.

[4] B. M. Sharif *et al.*, "Genome-wide genotyping elucidates the geographical diversification and dispersal of the polyploid and clonally propagated yam (Dioscorea alata)," *Ann. Bot.*, vol. 126, no. 6, pp. 1029–1038, Nov. 2020.

[5] S. Dreisigacker *et al.*, "Tracking the adoption of bread wheat varieties in Afghanistan using DNA fingerprinting," *BMC Genomics*, vol. 20, no. 1, pp. 1–13, Aug. 2019.

[6] M. A. Nadeem *et al.*, "DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing," *Biotechnol. Biotechnol. Equip.*, vol. 32, no. 2, pp. 261–285, Mar. 2018.

[7] N. Qureshi *et al.*, "Fine mapping of the chromosome 5B region carrying closely linked rust resistance genes Yr47 and Lr52 in wheat," *Theor. Appl. Genet.*, vol. 130, no. 3, pp. 495–504, 2017.

[8] A. Rasheed *et al.*, "Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives," *Molecular Plant*, vol. 10, no. 8. Cell Press, pp. 1047–1064, 07-Aug-2017.

[9] H. Ayalew *et al.*, "Comparison of TaqMan, KASP and rhAmp SNP genotyping platforms in hexaploid wheat," *PLoS One*, vol. 14, no. 5, p. e0217222, May 2019.

[10] L. Mansueto *et al.*, "Rice SNP-seek database update: New SNPs, indels, and queries," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D1075–D1081, Jan. 2017.

[11] Z. Li *et al.*, "The 3,000 rice genomes project," *Gigascience*, vol. 3, no. 1, p. 7, Dec. 2014.

[12] H. Peng *et al.*, "MBKbase for rice: An integrated omics knowledgebase for molecular breeding in rice," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D1085–D1092, 2020.

[13] K. Y. Morales *et al.*, "An improved 7K SNP array, the C7AIR, provides a wealth of validated SNP markers for rice breeding and genetics studies," *PLoS One*, vol. 15, no. 5, p. e0232479, May 2020.

[14] M. Yunus *et al.*, "Mapping of Resistance Genes to Brown Planthopper in Untup Rajab, an Indonesian Local Rice Variety," *J. AgroBiogen*, vol. 14, no. 2, p. 75, Dec. 2018.

[15] S. Dellaporta, J. Wood, and J. Hicks, "A plant {DNA} mini-preparation: version {III}.," *Plant Mol. Biol. Report.*, vol. 41, no. 4, pp. 19–21, 1983.

[16] "7k Infinium SNP genotyping - Genotyping Services Laboratory." [Online]. Available: https://sites.google.com/a/irri.org/snp-genotyping-mmal/genotyping/infinium-7k?overridemobile=true. [Accessed: 04-Jun-2020].

[17] "Rice SNP-Seek Database." [Online]. Available: https://snp-seek.irri.org/. [Accessed: 04-Jun-2020].

[18] S. Purcell *et al.*, "PLINK: A tool set for whole-genome association and population-based linkage analyses," *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, Sep. 2007.

[19] Perrier X. and J.-C. J.P., "DARwin - Dissimilarity Analysis and Representation for Windows," 2006. [Online]. Available: https://darwin.cirad.fr/. [Accessed: 04-Jun-2020].

[20] I. Letunic and P. Bork, "Interactive Tree of Life (iTOL) v4: recent updates and new developments," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W256–W259, Apr. 2019.

[21] A. R. Quinlan, "BEDTools: The Swiss-Army tool for genome feature analysis," *Curr. Protoc. Bioinforma.*, vol. 2014, no. 1, pp. 11.12.1-11.12.34, Sep. 2014.

[22] "KASP genotyping chemistry | LGC Biosearch Technologies." [Online]. Available: https://www.biosearchtech.com/products/pcr-kits-and-reagents/genotyping-assays/kasp-genotyping-chemistry. [Accessed: 04-Jun-2020].

[23] R. Van Berloo, "GGT 2.0: Versatile Software for Visualization and Analysis of Genetic Data," *J. Hered.*, no. 2, pp. 232–236, 2008.

[24] "Random number." [Online]. Available: https://www.google.com/search?q=random+number.

[25] S. Sandmann *et al.*, "Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data," *Sci. Rep.*, vol. 7, no. 43169, pp. 1–12, 2017.

[26] F. Pfeiffer *et al.*, "Systematic evaluation of error rates and causes in short samples in next-generation sequencing," *Sci. Rep.*, vol. 8, no. 10950, pp. 1–14, 2018.

[27] X. Tang *et al.*, "A large-scale whole-genome sequencing analysis reveals highly specific genome editing by both Cas9 and Cpf1 (Cas12a) nucleases in rice," *Genome Biol.*, vol. 19, no. 1, p. 84, Jul. 2018.

[28] T. R. Hargrove, W. R. Coffman, and V. L. Cabanilla, "Genetic interrelationships of improved rice varieties in Asia," *IRRI Res. Pap. Ser.*, vol. 23, p. 34p, 1978.

[29] M. Exposito-Alonso *et al.*, "The rate and potential relevance of new mutations in a colonizing plant lineage," *PLoS Genet.*, vol. 14, no. 2, p. e1007155, Feb. 2018.

[30] M. J. Thomson *et al.*, "Large-scale deployment of a rice 6 K SNP array for genetics and breeding applications," *Rice*, vol. 10, no. 1, p. 40, Dec. 2017.

[31] C. He, J. Holme, and J. Anthony, "SNP genotyping: The KASP assay," *Methods Mol. Biol.*, vol. 1145, pp. 75–86, 2014.

[32] S. Yang *et al.*, "An extended KASP-SNP resource for molecular breeding in Chinese cabbage(Brassica rapa L. ssp. pekinensis)," *PLoS One*, vol. 15, no. 10, p. e0240042, Oct. 2020.

[33] S. Bustin and J. Huggett, "qPCR primer design revisited," *Biomolecular Detection and Quantification*, vol. 14. Elsevier GmbH, pp. 19–28, 01-Dec-2017.

[34] F. C. Silva, G. T. Torrezan, R. C. Brianese, R. Stabellini, and D. M. Carraro, "Pitfalls in genetic testing: a case of a SNP in primer-annealing region leading to allele dropout in *BRCA1*," *Mol. Genet. Genomic Med.*, vol. 5, no. 4, pp. 443–447, Jul. 2017.

[35] P. Stothard, "Sequence Manipulation Suite: PCR Primer Stats." [Online]. Available: https://www.bioinformatics.org/sms2/pcr_primer_stats.html. [Accessed: 30-Mar-2020].

[36] B. C. Colburn, S. A. Mehlenbacher, and V. R. Sathuvalli, "Development and mapping of microsatellite markers from transcriptome sequences of European hazelnut (Corylus avellana L.) and use for germplasm characterization," *Mol. Breed.*, vol. 37, no. 2, pp. 1–14, Feb. 2017.

[37] P. Gramazio, J. Prohens, D. Borràs, M. Plazas, F. J. Herraiz, and S. Vilanova, "Comparison of transcriptome-derived simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers for genetic fingerprinting, diversity evaluation, and establishment of relationships in eggplants," *Euphytica*, vol. 213, no. 12, pp. 1–18, Dec. 2017.

[38] S. A. Kaiser, S. A. Taylor, N. Chen, T. S. Sillett, E. R. Bondra, and M. S. Webster, "A comparative assessment of SNP and microsatellite markers for assigning parentage in a socially monogamous bird," *Mol. Ecol. Resour.*, vol. 17, no. 2, pp. 183–193, Mar. 2017.

[39] R. Kamboj, B. Singh, T. K. Mondal, and D. S. Bisht, "Current status of genomic resources on wild relatives of rice," *Breed. Sci.*, vol. 70, no. 2, pp. 135–144, 2020.

[40] J. A. Labate, J. C. Glaubitz, and M. J. Havey, "Genotyping by sequencing for SNP marker development in onion," *Genome*, vol. 63, no. 12, pp. 607–613, 2020.

[41] J. Ha *et al.*, "Soybean-VCF2Genomes: A database to identify the closest accession in soybean germplasm collection," *BMC Bioinformatics*, vol. 20, no. S13, p. 384, Jul. 2019.

[42] L. Li *et al.*, "An accurate and efficient method for large-scale SSR genotyping and applications," *Nucleic Acids Res.*, vol. 45, no. 10, p. e88, 2017.

[43] D. Iamartino *et al.*, "Design and validation of a 90K SNP genotyping assay for the water buffalo (Bubalus bubalis)," *PLoS One*, vol. 12, no. 10, p. e0185220, 2017.

[44] Q. You, X. Yang, Z. Peng, L. Xu, and J. Wang, "Development and applications of a high throughput genotyping tool for polyploid crops: Single nucleotide polymorphism (SNP) array," *Frontiers in Plant Science*, vol. 9. Frontiers Media S.A., p. 104, 06-Feb-2018.