Pak. J. Agri. Sci., Vol. 59(3), 405-413; 2022
ISSN (Print) 0552-9034, ISSN (Online) 2076-0906
DOI:10.21162/PAKJAS/22.06
http://www.pakjas.com.pk

Basmati genome recovery and functional annotation using tunable-genotyping by sequencing

Umer Maqsood[1, 2] Jauhar Ali[2], Allah Ditta Babar[1], Ma. Anna Lynn Sevilla[2], Shahid Masood Shah[3] and Muhammad Arif[1,*]

[1]Agricultural Biotechnology Division, National Institute for Biotechnology and Genetic Engineering, Constituent College of Pakistan Institute of Engineering and Applied Sciences, Faisalabad, Pakistan; [2] International Rice Research Institute (IRRI), Los Baños, Laguna Philippines; [3]Department of Biotechnology, COMSATS University Islamabad, Abbottabad Campus, Pakistan
*Corresponding author's e-mail: marif_nibge@yahoo.com

Basmati rice is famous for its aroma and cooking quality around the world. Many biotic and abiotic factors have role in Basmati yield reduction. The development of high-yielding, stress-resilient varieties, using molecular breeding tools is the aim of different breeding programs. Among available genotyping techniques, Tunable Genotyping By Sequencing (tGBS) has an advantage by giving a fewer missing positions and more accurate SNPs. In the current study, we used tGBS for Basmati genome recovery and functional SNP annotation among Super Basmati (recurrent parent), IRBB57 and IR55419-04 (donor parents), with four progeny lines. Among these lines NIBGE-BR1 and NIBGE-BR18 are BLB resistant lines developed by the cross of Super Basmati (recurrent parent) and IRBB57 as donor parent. Similarly, NIBGE-DT11 and NIBGE-DT12 are drought resistant lines developed through cross of Super Basmati (recurrent parent) with drought resistant IRRI bred IR55419-04. The Ion Proton run yielded 2050620 raw reads with 84.3% of reads aligning to the reference genome. 61% reads were retained as uniquely aligned reads with average read depth of 15.4. SNPs were filtered maintaining minimum call rate of 50% and a final set of 4206 MCR50 SNPs was obtained. Phylogenetic analysis showed that selected progeny lines were genetically closer to recurrent parent. Background genome recovery analysis also confirmed that progeny lines carrying 88% to 90% of Super Basmati genome. Functional annotation showed 20 non-synonymous SNPs having a deleterious effect. Out of these seven highly deleterious SNPs were further probed for their functional importance and the status of these SNPs in rice lines currently being used has been discussed. Further studies of these SNPs may be helpful in the development of new stress related genomic markers.

Keywords: tGBS, SNP identification, functional annotation, background genome recovery.

INTRODUCTION

Rice is among major cereal crops of the world. Basmati rice is one of the superior types of rice native to plains of the north-west Himalayan ranges. Pakistan is famous for its Basmati rice, and Basmati is known for its cooking quality, long grain and aroma. It expresses its aromatic and grain quality traits more aggressively when cultivated in kallar tract of Pakistan (Sheikh *et al*, 2006). Basmati rice faces a lot of challenges for its production that includes multiple biotic and abiotic stresses (Zafar *et al.*, 2020). In last couple of decades, the breeders adopted molecular marker assisted breeding (MAS) as a tool to reduce considerable time for crop improvement. Molecular markers are the polymorphism in nucleotide sequences. Different individuals and species have different level of genomic polymorphism, it could be in the form of insertions or deletions (INDELs), Point mutation at single position known as single nucleotide polymorphism (SNPs) and duplication, translocation and replication error based polymorphism like Single Sequence Repeats (SSR), Randomly Amplified Polymorphic DNA (RAPD) (Ijaz and Khan 2009; Ijaz 2011). A suitable molecular marker for breeding programs should be co-dominant, distributed throughout the genome, highly polymorphic and accurate enough to reproduce the same results. Effectiveness of any marker assisted backcross breeding (MABB) scheme is confirmed by repeated background recovery analysis.

Maqsood, U., J. Ali, A.D. Babar, M.A.L. Sevilla[3], S.M. Shah[4] and M. Arif. 2022. Basmati genome recovery and functional annotation using tunable-genotyping by sequencing. Pakistan Journal of Agricultural Science. 58:405-413

[Received 6 Jan 2022; Accepted 27 Mar 2021; Published 27 Jun 2022]

The general objective of the backcross breeding methods is to maintain the gene(s) of interest and eliminate the unwanted genomic chunks of donor parents (Miah *et al.*, 2015). In post-genomic era a rapid advancement has been observed in different aspects of plant genomics research that includes whole-genome sequencing, transcriptome analysis, and single nucleotide polymorphism (SNP) discovery (Harper *et al.*, 2012, Li *et al.*, 2014; Ijaz *et al.*, 2020).

SNP-based markers are most accurate marker replacing the SSR markers for MAS or MAB. Multiple approaches and protocols have been developed for SNP discovery and genotyping in several crop species (Baird *et al.*, 2008, Elshire *et al.*, 2011, Peterson *et al.*, 2012, Van Tassell *et al.*, 2008). Multiplexing enables the GBS techniques to yield high-density data at rather very low cost (Chung *et al.*, 2017). Major draw-backs of most of the GBS approaches include high amount of missing data across the genome and fewer numbers of reads per site. Low read depth causes the problem of inaccurate base calling specially at the heterozygous sites. The tunable GBS (tGBS) offers the solution of both problems by reduction in number of sites being called with increased number of reads per site, yields fewer but more confident SNP calling. A good read depth at selected positions also helps in genotyping heterozygous sites accurately and reduces the chances of false SNP calling and help in discovering rare alleles with a higher confidence level (Ott *et al.*, 2017). Ali *et al.* (2018) have used tGBS for diversity analysis and SNP genotyping of 11 early backcross inbred populations (BC_1F_5). Very low heterozygosity confirmed the effectiveness of a stringent selection strategy under different abiotic stresses. Moreover, functional annotation and SNP typing of this panel also revealed the presence of 426 nsSNPs in 102 genes. In another study, the tGBS information has been applied in mapping QTLs playing important role during growth at different nutrient regimes (Mahender *et al.*, 2019).

Objectives of the current study include analysis of the genome recovery of Basmati rice in two different crosses developed for enhancement of Basmati rice production having gene against BLB and drought, (major problems faced by Basmati rice-growing areas of Pakistan) and functional gene annotation of SNPs using tGBS.

MATERIALS AND METHODS

***Plant material*:** The current study includes the genome analysis of seven rice lines. Among these lines, NIBGE-BR1 and NIBGE-BR18 are the selections from the cross of Super Basmati (a BLB sensitive parent) as recipient, while IRBB57 (BLB tolerant) as donor parent (Arshad *et al.*, 2016, Nasir *et al.*, 2019). Both of these lines were at BC_4F_6 generation, While NIBGE-DT11 and NIBGE-DT12 were the offspring of Super Basmati as drought-sensitive and IR55419-04 as drought-tolerant at BC_2F_3 generation (Sabar *et al.*, 2019). Seeds of IRBB57 and IR55419-04 were obtained from International Rice Research Institute (IRRI) gene bank, the seed source of Super Basmati was Rice Research Institute, Kala Shah Kaku, and Pakistan. NIBGE-BR1, NIBGE-BR18, NIBGE-DT11 and NIBGE-DT12 were developed and maintained at National Institute for Biotechnology and Genetic Engineering (NIBGE), Faisalabad, Pakistan.

***tGBS and sequence reads filtration*:** All seven lines were grown in glass house facility of International Rice Research Institute (IRRI) Philippines during dry season of 2018. Leaf samples were collected from 21 days old seedlings. Library preparation and sequencing run were carried out by Data2Bio using their standard protocol with Ion Proton sequencing machine (Fig. 1**)**. Threshold for PHRED quality score and error rate were set <15 and ≤3% respectively. Trimmed reads were aligned to reference genome using GSNAP (Islam et al.2015, Wu and Nacu 2010). SNP calling was performed after alignment with the reference genome (Ostive_204_v7.0.fa). Mapped reads with good confidence were used for SNP calling (≤2 mismatches for every 36 bp and <5bases for every 75 bp as unaligned tails). Every selected homozygous SNP had a PHRED score of 20 and its major allele was supported by at least three reads. SNPs were finally filtered by considering the following parameters. Missing data rate ≤80%, allele number = 2, number of genotypes in which an SNP is called ≥2, minor allele frequency ≥0.1, and heterozygosity range 0–10% (Islam *et al.*, 2015; Leiboff *et al.*, 2015). Finally, a set of MCR50 (Missing Call Rate of 50%) SNPs was generated based on minor allele frequency.

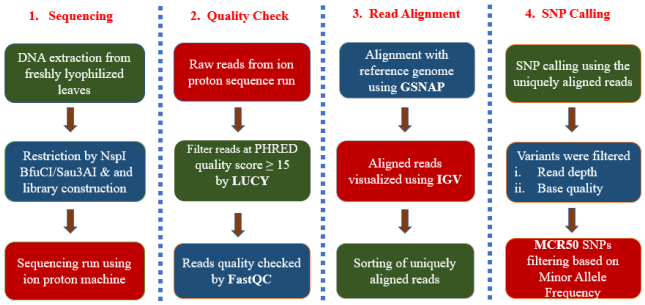

Figure 1. Schematic illustration of tGBS pipeline and downstream filtering steps to generate MCR50 genotype file

***Phylogenetic and background genome recovery analysis*:** A phylogenetic tree was constructed to ascertain the genetic relationship using ape package in R. To estimate recurrent parent genome recovery two separate sets were made. Set-I includes NIBGE-BR1 and NIBGE-BR18 along with Super Basmati and IRBB57 as recurrent and donor parents respectively. Set-II includes NIBGE-DT11 and NIBGE-DT12 along with Super Basmati and IR55419-04 as recurrent and donor parents respectively. The genotyping sets were filtered to remove monomorphic sites and parental

heterozygous sites. Genome recovery was visualized using Graphical Genotype version 2 (GGT2) (van Berloo, 2008). To construct the hClust tree and PCoA plots of both the sets we used flapjack software (Milne *et al*., 2010).

***Annotation and identification of functional nsSNPs*:** SNPs in genic and intergenic regions were annotated by SNPEff V 4.2 and SNiPlay tools (Cingolani *et al.,* 2012, Dereeper *et al.,* 2011). SIFT 4G determined the effect of nsSNPs on protein by comparing it with *Oryza sativa* IRGSP-1.0.23 genome build (http://sift.bii.astar.edu.sg/index.html) (Ng and Henikoff, 2006). The nsSNPs were further filtered to retain only deleterious SNPs based on SIFT score. For a better visualization the identified loci were drawn as circos diagram using J circos (An *et al.,* 2014).

RESULTS

***tGBS and sequencing reads filtration*:** Ion proton sequencing run generated a total of 2050620 reads with an average read length of 113 bp. The average read depth was 15.38, with 84.3% of reads aligning to the reference genome. Reads that aligned at more than one location of reference genome were regarded as non-unique or ambiguous reads. Those reads were filtered out, and finally, 61% of total reads were retained as unique reads for downstream analysis. Individual statistics are given in (Table 1). After the quality trimming and SNP calling a set of 16530 SNP were obtained. Further filtering was performed based on minor allele frequency, no of alleles and heterozygosity range as explained in the material and along with missing call rate. Finally, SNPs were filtered 4206 MRC50 SNPs were obtained.

***SNPs distribution*:** After the quality filtration of individual data of all lines, a combined MCR50 VCF file was generated based on the minimum call rate. The file contained 4206 SNPs that were called in at least 4 genotypes. Chromosome one contains 455 SNPs, highest in the number compared with rest of the chromosomes. While lowest numbers of SNPs were identified on chromosome 10 (Table 2). Out of these 4206 positions 2926 were transition while 1280 were transversion SNPs. Among transition SNPs 1494 were between C and T while 1432 were between A and G base. Out of 1280 transversions 323 were between A and C, 360 between A and T, 265 between C and G and 332 between G and T bases.

Table 2. Distribution of MCR50 SNPs among the 12 chromosomes

Chromosome No.	No. of SNPs	Chromosome No.	No. of SNPs
1	455	7	366
2	389	8	320
3	361	9	246
4	421	10	221
5	310	11	396
6	400	12	321

***Phylogenetic and background genome recovery analysis*:** A phylogenetic tree was constructed based on all 4206 ns SNPs using the ape package in "R" The phylogenetic analysis shows the genetic relationship among all seven lines included in this study. Two distinct clusters can be seen in the tree. Both parental lines from IRRI origin IRBB57 and IR55419 were clustered together while all 4 daughter lines were clustered along with Super Basmati (Recurrent Parent). (Fig. 2). Confirming close relation with recurrent parent. Chromosome wise distribution of SNPs used for background genome recovery analysis of both of the sub sets is given in (Fig. 3 A&B**).** For background genome recovery analysis of both sub sets (SET-I and SET-II) the genotyping files were filtered separately, for the removal of monomorphic and parental heterozygous sites. In sub set-1 that includes NIBGE-BR1, NIBGE-BR18 as progeny and Super Basmati and IRBB57 as parental lines, 2638 polymorphic sites were retained after filtration. SET-I that includes NIBGE-DT11 and NIBGE-DT12 along with Super Basmati and IR55409-4 was also filtered using the same parameters and retained 1727.

Basmati background genome recovery analysis of sub SET-I has revealed that NIBGE-BR1 has recovered 90% of recurrent parent genome, while NIBGE–BR18 has recovered 87% of Basmati genome. NIBGE-BR18 have some extra

Table 1. Individual lines-wise statistics of tGBS run results .

Rice Genotype	Total reads	Total trimmed reads	Average read length	Alignment %	*Unique aligned reads (%)
Super Basmati	445667	386306	115	84	61.43
IRBB57	630630	313293	116	81	60.97
IR55419-04	181154	170031	113	81.3	62.73
NIBGE-DT11	166944	147523	115	83.5	56.89
NIBGE-DT12	174393	164382	112	84	64.72
NIBGE-BR18	185818	174248	111	85.3	60.69
NIBGE-BR1	266014	229899	111	84	61.51

***Unique aligned reads are those mapped only on the single unique location of the genome

chunks of donor (IR55419-04) at chromosome 1, 2, 3 and 7. The graphical representation also showed that NIBGE–BR18 has higher number of heterozygous regions compared with NIBGE-BR1.

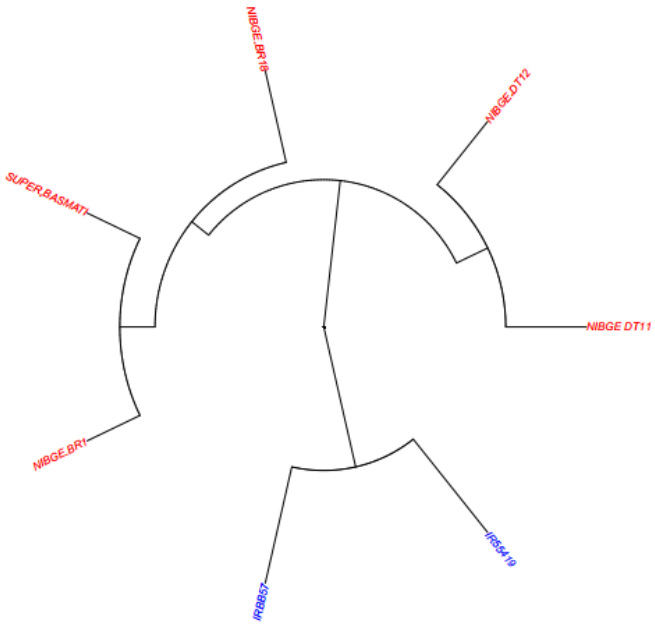

Figure 2. Phylogenetic tree of seven rice lines based on the set of 4206 MRC50 SNPs showing the relationship of progeny lines with Basmati parent clearly showing close relation of all progeny genotypes as compared with the IRRI bred donor parents.

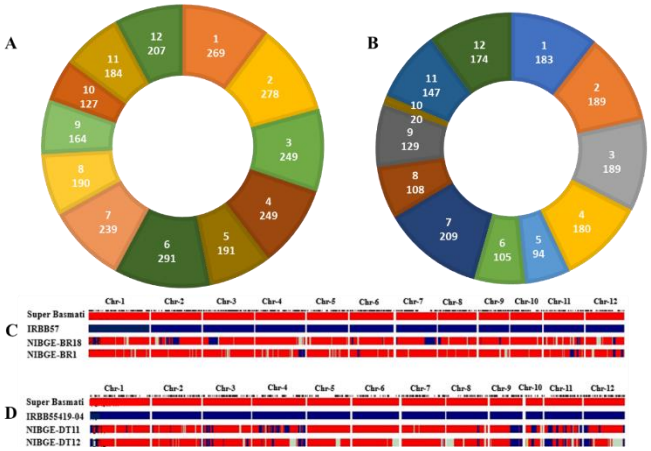

Figure 3. Chromosome wise distribution of polymorphic SNPs & Graphical representation of parental genome recovery. Here figure 3A shows distribution of polymorphic SNPs between Super Basmati and IRBB57, figure 3B shows distribution of polymorphic SNPs between Super Basmati and IR55419. While figure 3C and 3D Graphical representation of parental genome recovery of SET-I and SET-II respectively. Red bar shows the recurrent parent genome while dark blue color represents donor parent genome. Light blue color represent the heterozygous regions.

Similarly, in genotyping SET-II, NIBGE-DT11 and NIBGE-DT12 both showed 90% recovery of Super Basmati genome compared with the donor parent. NIBGE-DT 11 retained a larger portion of donor parent segments at chromosome 4, while more heterozygosity was visualized in NIBGE-DT12 (Fig. 3C&D).

Three-dimensional Principal coordinate analysis (PCoA) and phylogenetic clustering (hClust) using Flapjack software has demonstrated that all 4 NIBGE lines (BR1, BR18, DT11 and DT12) have shown closer similarity with Super Basmati compared with both the donor parents (Fig. 4).

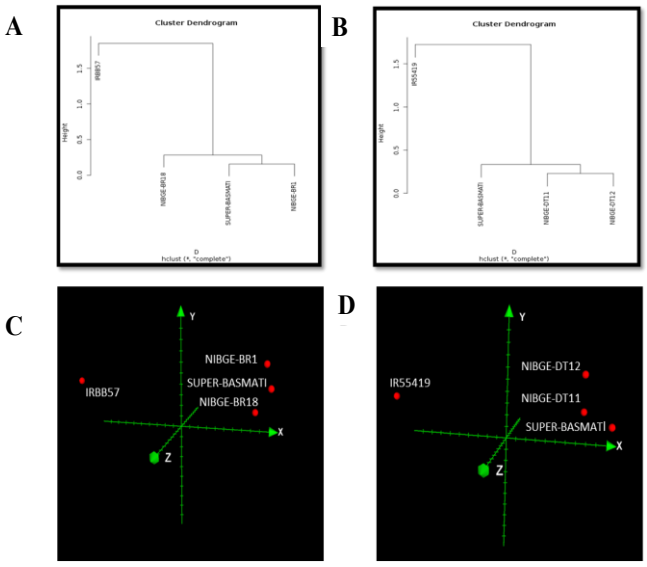

Figure 4. Cluster dendrogram and PCoA analysis of Set - I and Set-II, (A) h*Cluster dendrogram analysis of Set-I (B) h*Cluster dendrogram of Set-II (C) PCoA analysis of Set-I (D) PCoA analysis of Set-II.

Annotation and identification of functional nsSNPs; Functional annotation reveals that 1698 SNPs were present in genic region, while rest of them were in non-genic regions. Among these genic SNPs 882 were exonic. In our study we found 312 synonymous and 409 nsSNPs**.** SIFT analysis divided the nsSNPs into two categories i.e. 'Tolerated' and 'Deleterious'. Out of 409 nsSNPs 20 were found to be deleterious. Among these 20 SNPs 4 were present in already reported genes with known functions, while 16 novel SNPs were found. All 20 deleterious SNPs are drawn in the form of circos diagram (Fig. 5).

Six very important deleterious SNPs has been probed based on their deleterious role (Table 3). One deleterious SNP has been identified on chromosome one at locus

LOC_Os01g41260 with allele variation from G to A. It gives a codon shift from GGA to AGA replacing glycine to alanine.

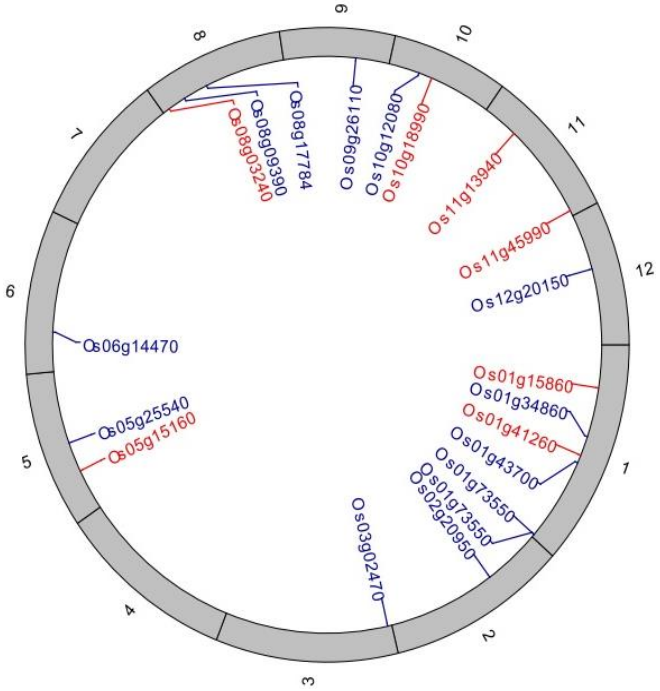

Figure 5. Circos representation of SIFT analysis results of nsSNPs. Red font shows the genes near SNPs causing crucial variations with SIFT score as zero.

This creates a protein change in OsFDB F-box and FDB domain containing protein. In our genotypes sets both IRRI bred line IR55419 and IRBB57 have G at this position while Super Basmati, NIBGE-BR18 and NIBGE-BR1 have A at this position. Another highly deleterious SNP with SIFT score 0 is found at position 8606005 of chromosome 5 (locus LOC_Os05g15160).

The nucleotide variation from C to T causes codon change from GCG to GTG while replacing Alanine with Valine. At chromosome 10 a deleterious nsSNP T/A was present at position 9667381 at the locus LOC_Os10g18990 (Leucine rich region) altering the codon and replacing Leucine with Histidine. Two deleterious SNPs were identified at chromosome 11 C to T variation at the position 7716773 and A to G variation at position 27834859 within LOC_Os11g13940 and LOC_Os11g45990 respectively. On chromosome 8 a G to A SNP is found at 1504912.

DISCUSSION

In this study, tGBS (Schnable *et al.,* 2013) was used for SNP-typing of a set of rice lines comprising of 7 rice lines i.e., Super Basmati, IR55419-04, IRBB57, NIBGE-BR1, NBGE-BR18, NIBGE-DT11 and NIBGE-DT12. tGBS has been employed for different genomics studies of multiple crops and proved to yield more accurate SNPs in term of read depth and SNP sites confidence compared with other GBS techniques (Elshire *et al.,* 2011). A recent study by Ali *et al.,* 2018 included 11 early backcross introgression populations with 564 individuals and 12 parents has demonstrated the ability of this technique in exploring the genetic diversity and functional SNPs in rice. LMD 50 SNPs (Low missing data SNPs) successfully differentiated among all eleven populations along with their parental materials.

For current study the sequence run was performed on an Ion Proton sequencing machine with average read length 113. Approximately, 16% of reads remained unmapped. The reads unable to be mapped can be due to the difference in the *Japonica*, basmati and *Indica* subspecies genome (Tang *et al.*, 2015). The other reason of this phenomena could be the errors generated during the sequencing runs (Mehra *et al.,* 2015). Chromosome wise maximum number SNPs were reported on chromosome one due the reason that chromosome one is the largest rice chromosome (Chen *et al.,* 2002). The high density of SSR markers was also reported on Chromosome one which strengthen our results (McCouch *et al.,* 2002; Ul Haq and Ijaz 2019).

The phylogenetic analysis shows the genetic relationship among all seven lines included in this study. Two distinct clusters can be seen in the tree. Both parental lines from IRRI origin IRBB57 and IR55419 were clustered together while all 4 daughter lines were clustered along with Super basmati. This is due to the fact that all these lines were developed by keeping in view all the desired traits of Super Basmati background along with BLB and drought tolerance from donor parents.

Out of these 4206 positions 2926 were transition while 1280 were transversion SNPs (Fig. 3). Transition SNPs were more than double than the transversion SNPs. In different studies it has been proved that the transition SNPs are greater in number compared to the transversion SNPs (Hu *et al.,* 2014; Hwang *et al.,* 2014). Almost 40% SNPs were found in genic region while 3.8% of identified SNPs were present in the UTRs. Out of the genic SNPs 861 (20% of total) were present in intronic regions, the amount of intronic SNPs is always greater than the SNPs present in exonic region of a gene (Roy and Reddy Lachagari, 2017).

Out of these 4206 SNPs ~40% were found in genic region while 3.8% of identified SNPs were present in the UTRs. The percentage of genic region SNPs was comparatively higher than the output of other GBS techniques but results are consistent with a recent tGBS study where almost 40% of SNPs were found in the genic region and 4% were present in the regulatory regions (Ali *et al.,* 2018). Out of the genic SNPs 861 (20% of total) were present in intronic regions, the amount of intronic SNPs is always greater than the SNPs present in exonic region of a gene (Ali *et al.*, 2018; Roy and Reddy-Lachagari, 2017). As evident from the annotation

results, number of transition SNPs were more than double as the transversion SNPs. In different studies it has been proved that the transition SNPs are greater in number compared to the transversion SNPs (Hwang *et al.,* 2014; Hu *et al.,* 2014; Ali *et al.,* 2018). In the coding regions 312 sSNPs were identified while number of nsSNPs was 409. Number of nsSNPs were greater than the number of sSNPs in the CDS regions of all genotypes as it has been already reported in different rice based GBS studies (Rathinasabapathi *et al.,* 2015; Mehra *et al.,* 2015).

SIFT (Sorting Intolerant from Tolerant) analysis divided the nsSNPs into two categories i.e., Tolerated and Deleterious. These highly impactful nsSNPs are referred as the deleterious nsSNPs. The SNPs present in non-coding regions merely have any direct impact on the amino acid alteration, but they may have a considerable impact on splice sites and the transcription factor binding sites and consequently alter the extent of transcription (Xu *et al.,* 2011). Out of 409 nsSNPs 389 were regarded as Tolerated SNPs while 20 were found to be deleterious. Among these 20 SNPs 4 were present in already reported genes with known functions, while 16 novel SNPs were found. All 20 deleterious SNPs are drawn in the form of circos diagram. Out of these 20 SNPs 6 were highly deleterious having a SIFT score near 0. Deleterious SNP identified on chromosome one at locus LOC_Os01g41260 with allele variation from G to A. This transition gives a codon shift from GGA to AGA replacing glycine to alanine. This creates a protein change in OsFDB F-box and FDB domain containing protein. In our genotypes sets both IRRI bred line IR55419 and IRBB57 have G at this position while Super Basmati, NIBGE-BR18 and NIBGE-BR1 have A at this position. Among multiple roles of F-box it is also involved in response to light and a biotic stress stimulus (Jain *et al.,* 2007). The analysis of SNP-seek data has revealed that most of the lines of Basmati origin have A at this position, although the major allele contains G. Another highly deleterious SNP with SIFT score 0 is found at position 8606005 of chromosome 5 in the locus LOC_Os05g15160 (Phosphate/phosphate translocator, putative protein). The nucleotide variation from C to T causes codon change from GCG to GTG while replacing Alanine with Valine. The GO terms of biological function of this locus explains its activities in membrane transports especially in response to abiotic stress stimulus (Liu *et al.,* 2013). In our set of genotypes both IRRI bred parents, IR55419 and IRBB57 have C at this position while Super Basmati, NIBGE-BR18 and NIBGE-BR1 have T at this position while sequencing results of NIBGE-DT11 and NIBGE-DT12 shows missing data as "N". Our sequencing results and SNP-seek data base analysis has confirmed the presence of C on this position in IR55419 while about 60% of all aromatic varieties available in SNP-seek data possess the allele T at this position, more specifically all Basmati varieties of Pakistani origin available in SNP-seek have the allele T at this position. This position can be uses for the development of Pakistan Basmati specific marker as all of Basmati varieties available in 3K rice genome have T instead of C (otherwise reference allele). At chromosome 10 a deleterious nsSNP T/A is present at position 9667381 at the locus LOC_Os10g18990 (Leucine rich region) altering the codon from CTT to CAT and replacing Leucine with Histidine amino acid. Leucine rich regions are involved in conferring the resistance against many stresses such as Bacterial leaf blight. In current set of genotypes allele T is only present in IRBB57 while rest of the genotypes have A at this position. SNP-seek has designated it as a major allele as it is present in 52% of total genotypes present in 3K project. More studies are needed to find out the function of this SNP near Leucine rich region loci. Two deleterious SNPs were identified at chromosome 11 C to T variation at the position 7716773 and A to G variation at position 27834859 within LOC-_Os11g13940 (NBS-LRR disease resistance protein) and LOC_Os11g45990 (Von Willebrand factor type A domain containing protein) respectively. Chromosome 11 is well known to harbor multiple Bacterial blight resistance genes in rice including *Xa3*, *Xa4*, *Xa6*, *xa9*, *Xa11*, *Xa21*, *Xa22*, *Xa23*, and *Xa26* (*Gu et al., 2008*; Ronald *et al.,* 1992; *Song et al.,* 1995; Sun *et al.,* 2003; Wang *et al.,* 2003; Xiang *et al.,* 2006; Zhang *et al.,* 1998). The SNP at position 7716773 causing variation in NBS-LRR disease resistance protein. This protein has important role in conferring the resistance against BB as well, as many R genes confer resistance through LRR repeats. Both NIBGE-BR1 and NIBGE-BR18 has T allele at this position while rest of the Basmati origin varieties along with IR55419-04 have allele C at this location. SNP-seek data confirm the abundance of C allele in all Basmati varieties other than Basmati-385. C is the major allele with its presence in 84% of total 3k varieties. M At chromosome four *Xa1* has been known to encode Nucleotide binding LRR (Yoshimura *et al.,* 1998). On chromosome 8 a G to A alteration was identified at position 1504912 The SNP-seek data has revealed that at this position G is the major allele as it is present in 92% of all genotypes available in the database, while ~85% of all aromatic genotypes have A at this position. Moreover, all Basmati origin genotypes of database possess A at this position. Betaine aldehyde dehydrogenase genes (BADH1 and BADH2) are responsible to provide aromatic character in rice. BADH2 is present on chromosome 8 residing the locus os08g0424500 from position 20379823 to 20385975 (He *et al.,* 2015; Kovach *et al.,* 2009; Shao *et al.,* 2013). Although, our identified SNP is at considerable distance from BADH2 locus, but its abundance in the aromatic group is indicating its association with the aroma genes. All of the genotypes with Basmati background under this study have A at this position and strengthen the idea that this position can be used for the development of Basmati specific molecular marker.

***Conclusions*:** Current study confirms that all 4 progeny genotypes have recovered more than 80% of Basmati genome. The background genome recovery analysis results are further strengthen by Phylogenetic and PCoA analysis results. The study also shows that tGBS is a powerful and accurate tool for background recovery analysis as well as functional annotation studies. The functional annotation of SNPs identified 20 deleterious SNPs. Out of these, 6 highly important SNPs were studied in detail. This study has helped us to identify a highly deleterious SNP present at chromosome 11 causing variation in NBS-LRR that is linked to resistance mechanism to multiple stresses. The study also identified a Basmati specific SNP at chromosome 8 that can be used for development of Basmati specific Molecular marker as a tool of DNA finger printing.

***Conflict of interest*:** The authors declare that they have no conflict of interest.

***Acknowledgement*:** The authors extend their gratitude to the Higher Education Commission (HEC), Pakistan, for providing research fellowship under International Research Support Initiative Program (IRSIP) during 2017 to support research work at IRRI, Philippines.

***Author's contribution statement*:** UM, MALS performed field work UM and ADB performed bioinformatics analysis. UM and SMS written the manuscript. MA and JA Conceptualized and supervised research and finally reviewed and edited the manuscript.

REFERENCES

Ali, J., U. M. Aslam, R. Tariq, V. Murugaiyan, P. S. Schnable, D. Li, C. M. Marfori-Nazarea, J. E. Hernandez, M. Arif, J. Xu and Z. Li. 2018. Exploiting the genomic diversity of rice (*Oryza sativa* L.): SNP-typing in 11 early-backcross introgression-breeding populations. Frontiers in Plant Sciences. 9:849.

An, J., J. Lai, A. Sajjanhar, J. Batra, C. Wang and C. C. Nelson. 2014. J-Circos: an interactive Circos plotter. Bioinformatics. 31:1463-1465.

Arshad, H. M. I., S. T. Sahi, M. Atiq and W. Wakil. 2016. Appraisal of resistant genes and gene pyramid lines of rice against indigenous pathotypes of *Xanthomonas oryzae pv. oryzae* in Punjab, Pakistan. Pakistan Journal of Agricultural Sciences. 53:365-370

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko and E. A. Johnson. 2008. "Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 3:e3376.

Chung, Y. S., S. C. Choi, T.-H. Jun and C. Kim. 2017. Genotyping-by-sequencing: a promising tool for plant genetics research and breeding. Horticulture Environment and Biotechnology. 58:425-431.

Cingolani, P., A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu and D. M. Ruden. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin). 6:80-92.

Dereeper, A., S. Nicolas, L. Le Cunff, R. Bacilieri, A. Doligez, J. P. Peros, M. Ruiz and P. This. 2011. SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. BMC Bioinformatics. 12:1471-2105.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler and S. E. Mitchell. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS One. 6:e19379.

Gu, K., J. S. Sangha, Y. Li and Z. Yin. 2008. High-resolution genetic mapping of bacterial blight resistance gene *Xa10*. Theoretical and Applied Genetics. 116:155-163.

Harper, A. L., M. Trick, J. Higgins, F. Fraser, L. Clissold, R. Wells, C. Hattori, P. Werner and I. Bancroft. 2012. Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. Nature Biotechnology. 30:798-802.

He, Q., J. Yu, T.-S. Kim, Y.-H. Cho, Y.-S. Lee and Y.-J. Park. 2015. Resequencing reveals different domestication rate for *BADH1* and *BADH2* in rice (*Oryza sativa*). PLoS One. 10:e0134801.

Hu, Y., B. Mao, Y. Peng, Y. Sun, Y. Pan, Y. Xia, X. Sheng, Y. Li, L. Tang, L. Yuan and B. Zhao. 2014. Deep re-sequencing of a widely used maintainer line of hybrid rice for discovery of DNA polymorphisms and evaluation of genetic diversity. Molecular Genetics and Genomics. 289:303-315.

Hwang, S. G., J. G. Hwang, D. S. Kim and C. S. Jang. 2014. Genome-wide DNA polymorphism and transcriptome analysis of an early-maturing rice mutant. Genetica. 142:73-85.

Ijaz, S. and I.A. Khan. 2009. Molecular characterization of wheat germplasm using microsatellite markers. Genetics and Molecular Research. 8:809-815.

Ijaz, S. 2011. Microsatellite markers: An important fingerprinting tool for characterization of crop plants. African Journal of Biotechnology. 10:7723-7726

Ijaz, S., I. Ul. Haq and B. Nasir. 2020. In silico identification of expressed sequence tags based simple sequence repeats (EST-SSRs) markers in Trifolium species. ScienceAsia. 46:6-10.

Islam, M. S., G. N. Thyssen, J. N. Jenkins and D. D. Fang. 2015. Detection, validation, and application of genotyping-by-sequencing based single nucleotide

polymorphisms in upland cotton. Plant Genome. 8:e2014.2007.0034.

Jain, M., A. Nijhawan, R. Arora, P. Agarwal, S. Ray, P. Sharma, S. Kapoor, A. K. Tyagi and J. P. Khurana. 2007. F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. Plant Physiology. 143:1467-1483.

Kovach, M. J., M. N. Calingacion, M. A. Fitzgerald, S. R. McCouch and B. A. Larkins. 2009. The origin and evolution of fragrance in rice (*Oryza Sativa* L.). Proceedings of National Academy of Sciences. 106:14444-14449.

Leiboff, S., X. Li, H.-C. Hu, N. Todt, J. Yang, X. Li, X. Yu, G. J. Muehlbauer, M. C. P. Timmermans, J. Yu, P. S. Schnable and M. J. Scanlon. 2015. Genetic control of morphometric diversity in the maize shoot apical meristem. Nature Communications. 6:8974.

Li, F., G. Fan, K. Wang, F. Sun, Y. Yuan, G. Song, Q. Li, Z. Ma, C. Lu, C. Zou, W. Chen, X. Liang, H. Shang, W. Liu, C. Shi, G. Xiao, C. Gou, W. Ye, X. Xu, X. Zhang, H. Wei, Z. Li, G. Zhang, J. Wang, K. Liu, R. J. Kohel, R. G. Percy, J. Z. Yu, Y.-X. Zhu, J. Wang and S. Yu. 2014. Genome sequence of the cultivated cotton *Gossypium arboreum*. Nature Genetics. 46: 567-572.

Liu, L., Q. Mei, Z. Yu, T. Sun, Z. Zhang and M. Chen. 2013. An integrative bioinformatics framework for genome-scale multiple level network reconstruction of rice. Journal of Integrative Bioinformatics. 10:94-102.

Mahender, A., J. Ali, G. D. Prahalada, M. A. L. Sevilla, C. H. Balachiranjeevi, J. Md, U. Maqsood and Z. Li. 2019. Genetic dissection of developmental responses of agro-morphological traits under different doses of nutrient fertilizers using high-density SNP markers. PLoS One. 14:e0220066.

McCouch, S. R., L. Teytelman, Y. Xu, K. B. Lobos, K. Clare, M. Walton, B. Fu, R. Maghirang, Z. Li, Y. Xing, Q. Zhang, I. Kono, M. Yano, R. Fjellstrom, G. DeClerck, D. Schneider, S. Cartinhour, D. Ware and L. Stein. 2002. Development and mapping of 2240 new SSR markers for rice (*Oryza sativa L.*). DNA Research. 9:257-279.

Mehra, P., B. K. Pandey and J. Giri. 2015. Genome-wide DNA polymorphisms in low phosphate tolerant and sensitive rice genotypes. Scientific Reports. 5:1-14.

Miah, G., M. Y. Rafii, M. R. Ismail, A. B. Puteh, H. A. Rahim and M. A. Latif. 2015. Recurrent parent genome recovery analysis in a marker-assisted backcrossing program of rice (*Oryza sativa*.). Comptes Rendus Biologies. 338:83-94.

Mondini, L., A. Noorani, and M. A. Pagnotta. 2009. Assessing plant genetic diversity by molecular tools. Diversity. 1:19–35

Nadeem, M. A., M.A. Nawaz, M. Q. Shahid, Y. Doğan, G. Comertpay, M. Yıldız, R. Hatipoğlu, F. Ahmad, A. Alsaleh, N. Labhane, H. Özkan, G. Chung and F.S. Baloch. 2018. DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. Biotechnology & Biotechnological Equipment. 32: 261-285.

Nasir, M., B. Iqbal, M. Hussain, A. Mustafa and M. Ayub. 2019. Chemical management of bacterial leaf blight disease in rice. Journal of Agriculture Research. 57:99-103.

Ott, A., S. Liu, J. C. Schnable, C.-T. E. Yeh, K.-S. Wang and P. S. Schnable. 2017. tGBS® genotyping-by-sequencing enables reliable genotyping of heterozygous loci. Nucleic acids Research. 45:e178-e179.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher and H. E. Hoekstra. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One. 7:e37135.

Rathinasabapathi, P., N. Purushothaman, R. Vl and M. Parani. 2015. Whole genome sequencing and analysis of Swarna, a widely cultivated *indica* rice variety with low glycemic index. Scientific Reports. 5:e11303.

Ronald, P. C., B. Albano, R. Tabien, L. Abenes, K.-s. Wu, S. McCouch and S. D. Tanksley. 1992. Genetic and physical analysis of the rice bacterial blight disease resistance locus, *Xa21*. Molecular and General Genetics. 236:113-120.

Roy, S. C. and V. B. Reddy Lachagari. 2017. Assessment of SNP and InDel variations among rice lines of Tulaipanji x Ranjit. Rice Science. 24:336-348.

Sabar, M., G. Shabir, S. M. Shah, K. Aslam, S. A. Naveed and M. Arif. 2019. Identification and mapping of QTLs associated with drought tolerance traits in rice by a cross between Super Basmati and IR55419-04. Breeding Science. 69:169-178.

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

Sheikh, A., M. A. Mahmood, A. Bashir and M. Kashif. 2006. Adoption of rice technological package by the farmers of irrigated Punjab Pakistan. Journal of Agriculture Research. 44:341-352.

Song, W., G. Wang, L. Chen, H. Kim, L. Pi, T. Holsten, J. Gardner, B. Wang, W. Zhai and L. Zhu. 1995. The rice disease resistance gene, *Xa21*, encodes a receptor-like protein kinase. Science. 270:804-801.

Sun, X., Z. Yang, S. Wang and Q. Zhang. 2003. Identification of a 47-kb DNA fragment containing Xa4, a locus for bacterial blight resistance in rice. Theoretical and Applied Genetics. 106:683-687.

Ul Haq, I. and S. Ijaz. 2019. Assessment of genetic diversity based on ISSR markers in neopestalotiopsis species collected from guava (Psidium guajava l.) Plants affected with canker disease in Pakistan. Applied Ecology Environmental Research. 17:11803-11811.

Van Tassell, C. P., T. P. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel, C. T. Lawley, C. D. Haudenschild, S. S. Moore, W. C. Warren and T. S. Sonstegard. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods. 5:247-252.

Van-Berloo, R. 2008. GGT 2.0: versatile software for visualization and analysis of genetic data. Journal of Heredity. 99:232-236.

Wang, C., M. Tan, X. Xu, G. Wen, D. Zhang and X. Lin. 2003. Localizing the bacterial blight resistance gene, xa 22(t), to a 100-kilobase bacterial artificial chromosome. Phytopathology. 93:1258-1262.

Xiang, Y., Y. Cao, C. Xu, X. Li and S. Wang. 2006. *Xa3*, conferring resistance for rice bacterial blight and encoding a receptor kinase-like protein, is the same as *Xa26*. Theoretical and Applied Genetics. 113:1347-1355.

Xu, X., S. Pan, S. Cheng, B. Zhang, D. Mu, P. Ni, G. Zhang, S. Yang, R. Li, J. Wang, G. Orjeda, F. Guzman, M. Torres, R. Lozano, O. Ponce, D. Martinez, G. D. Cruz, S. K. Chakrabarti, V. U. Patil, K. G. Skryabin, B. B. Kuznetsov, N. V. Ravin, T. V. Kolganova, A. V. Beletsky, A. V. Mardanov, A. Di Genova, D. M. Bolser, D. M. A. Martin, G. Li, Y. Yang, H. Kuang, Q. Hu, X. Xiong, G. J. Bishop, B. Sagredo, N. Mejía, W. Zagorski, R. Gromadka, J. Gawor, P. Szczesny, S. Huang, Z. Zhang, C. Liang, J. He, Y. Li, Y. He, J. Xu, Y. Zhang, B. Xie, Y. Du, D. Qu, M. Bonierbale, M. Ghislain, M. D. Herrera, G. Giuliano, M. Pietrella, G. Perrotta, P. Facella, K. O'Brien, S. E. Feingold, L. E. Barreiro, G. A. Massa, L. Diambra, B. R. Whitty, B. Vaillancourt, H. Lin, A. N. Massa, M. Geoffroy, S. Lundback, D. Dellapenna, C. R. Buell, S. K. Sharma, D. F. Marshall, R. Waugh, G. J. Bryan, M. Destefanis, I. Nagy, D. Milbourne, S. J. Thomson, M. Fiers, J. M. E. Jacobs, K. L. Nielsen, M. Sønderkær, M. Iovene, G. A. Torres, J. Jiang, R. E. Veilleux, C. W. B. Bachem, J. D. Boer, T. Borm, B. Kloosterman, H. V. Eck, E. Datema, B. T. L. Hekkert, A. Goverse, R. C. H. J. V. Ham, R. G. F. Visser, C. 2011. Genome sequence and analysis of the tuber crop potato. Nature. 475:189-195.

Yoshimura, S., U. Yamanouchi, Y. Katayose, S. Toki, Z. X. Wang, I. Kono, N. Kurata, M. Yano, N. Iwata and T. Sasaki. 1998. Expression of Xa1, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. Proceedings of National Academy of Science. 95:1663-1668.

Zafar, K., M. Z. Khan, I. Amin, Z. Mukhtar, S. Yasmin, M. Arif, K. Ejaz and S. Mansoor. 2020. Precise CRISPR-Cas9 mediated genome editing in super basmati rice for resistance against bacterial blight by targeting the major susceptibility gene. Frontiers in Plant Sciences. 11:575.

Zhang, Q., S. Lin, B. Zhao, C. Wang, W. Yang, Y. Zhou, D. Li, C. Chen and L. Zhu. 1998. Identification and tagging a new gene for resistance to bacterial blight (*Xanthomonas oryzae pv. oryzae*) from *O. rufipogon*. Rice Genetic Newsletters. 15:138-142.