Revealing Genetic Diversity in the Rice Breeding Lines (Oryza sativa x O. rufipogon) Based on Whole Genome Sequencing Variants (SNPs/InDels)

Shukdeb Sarkar
University of North Bengal
Jayita Sarkar
University of North Bengal
Bhumika T. Yonzon
University of North Bengal
Sona Limboo
University of North Bengal
Subhas Chandra Roy
subhascr2011@gmail.com

University of North Bengal

Research Article

Keywords: Rice, Whole-genome sequencing, SNP/InDel, Genome wide variation, Interspecific hybridization

Posted Date: February 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8630035/v1

Additional Declarations: No competing interests reported.

Abstract

Whole genome sequencing makes it possible to find a vast number of DNA polymorphisms, including SNPs and InDels, between cultivars. Genetic variation between cultivars and the genomic sites of heritable characteristics may be quickly identified due to these DNA polymorphisms. The whole genomes of four rice lines, comprising two offspring lines(YR13 and YR32) and their parents (*Oryza sativa* cv. Yamuna and *Oryza rufipogon*), were resequenced in order to determine the sequence diversity resulting from interspecific hybridization using Illumina next-generation sequencing. The range of sequencing coverage was 18.46× to 27.51×, and the mapping efficiency against the reference genome of *O. sativa* Nipponbare was high (98.21–98.73%). Variant discovery indicated 452,810–1,393,901 polymorphisms per line, with transition/transversion ratios of 2.43–2.50 and SNPs accounting for the majority (88–91%). Functional annotation showed that 7.8–9.6% of variants occurred in exonic regions, with missense mutations being the most frequent (52–54%), while high-impact coding mutations accounted for less than 1%.Unique variants ranged 973–4,041, whereas 15,137 were common to all lines. The progeny lines primarily carried SNPs from a number of agronomically significant genes and QTLs, such as NOG1 (grain number), qSH1 and SH4 (seed shattering), GS2 (grain size), Hd1/Hd6 (heading date), Badh2 (aroma), Saltol (salt tolerance), and Sub1A (submergence tolerance), indicating successful introgression of trait-associated alleles. The transcription, transport, signal transduction, and carbohydrate/lipid metabolism pathways were shown to be enriched by Gene ontology and KEGG analysis. Low to moderate differentiation across lines was revealed by genome-wide Fst analysis (mean SNP Fst = 0.012; InDel Fst = 0.017), which reflected both parental contributions and new recombination patterns in progeny. These findings provide a thorough genetic resource for functional genomics and marker-assisted breeding in rice improvement.

Highlights

• Whole genome resequencing of interspecific rice parents and progeny lines

• SNPs accounting for the majority of the variation (88–91%)

• Key traits associated allels introgressed from wild rice

• Enrichment in transcription, transport and metabolic pathways

• Low to moderate genomic differentiation between lines

1. Introduction

The genomes of *Oryza sativa*, the domesticated rice, contain information to explain the wide range of physiological, morphological and ecological variations seen in the several varieties grown for food. Genetic polymorphisms are primarily responsible for phenotypic variances, and the way in which these polymorphisms interact with an individual's environment greatly influences how that individual exhibits a trait. A wide range of molecular markers, including simple sequence repeats (SSR) and restriction fragment length polymorphism, have been developed for use in genetic study based on these DNA variants [1]. Through large-scale resequencing of whole genomes, the development and affordable use of next-generation sequencing technologies have greatly enhanced our capacity to investigate genetic diversity across the genome [2].

The identification of millions of single nucleotide polymorphisms (SNPs), insertion-deletions (InDels), and genome-wide DNA polymorphisms has been made possible by advancements in genome sequencing technologies. These are extremely useful tools for determining the association between genomes and heritable characteristics as well as for assessing genetic diversity within a population [3, 4]. With the availability of reference genome sequences for a number of crop species, it is now possible to quickly identify candidate genes using bioinformatics analysis and find SNPs by comparing the reference sequence with those of different cultivars [5, 6].

It is crucial to use trait-specific SNPs to characterize genotypes in order to examine the frequency of advantageous alleles for different variables of interest associated with biotic and abiotic stressors as well as grain quality characteristics [7]. This chance to choose genotypes utilizing modern breeding techniques like genomic selection and marker-assisted forward breeding to choose elite parents with exceptional traits has been made possible by the high throughput SNP platform. A breeding program may be able to satisfy the diverse needs of stakeholders from various societal segments by increasing the genetic gain if it has many effective QTLs or genes.

In genetic research, SNP and InDel are the most common polymorphisms in the genomes of the majority of organisms. Large numbers of SNP and InDel may now be found by comparing the rice genome to the Nipponbare reference sequence, which was recently used to sequence the rice genome with excellent precision using the Japonica rice cultivar Nipponbare [8]. Recently, high-throughput sequencers have been used to resequence the whole genomes of rice cultivars, using Nipponbare as a reference. The japonica rice cultivar Koshihikari, which is closely related to Nipponbare, has had its whole genome resequenced [9]. By comparing these two genomes, 67,051 SNPs in total have been found. The dynamics of genomic compositions were also examined using typing arrays based on SNPs for historical representative rice cultivars. Whole genome resequencing identified 168,228 DNA polymorphisms in a landrace cultivar of Japonica rice, and direct use of InDels as DNA markers confirmed their validity [10]. In order to discover genes of agronomic importance, 6.5 million high-quality SNPs were found by resequencing 50 accessions of wild and farmed rice. Based on the SNPs collected, thousands of genes with much reduced diversity were also found. Domestication was thought to be the time when these candidate genes were selected [11].

In this present investigation, we used whole genome sequencing to analyze the SNP and InDel variations between two parents, Yamuna and wild rice, and their progeny lines (YR 13 and YR32) and evaluate parental genome introgression into the progeny lines to obtain genetic architectural knowledge of SNP variations linked to traits for crop improvement.

2. Materials and methods

2.1 Plant materials

An experiment was planned with four rice lines comprised of cultivated rice Yamuna, a wild rice (*Oryza rufipogon* Griff.) and their two breeding lines YR13 and YR32. Rice cultivar Yamuna (*Oryza sativa* L.) used as a maternal parent, collected from Alipurduar district (26°31N, 89°32 E) West Bengal, India; wild rice (*Oryza rufipogon* Griff. [Herbarium sheet submitted to NBU herbarium with Accession no. 13048]) used as a paternal parent collected from the marshy areas of University of North Bengal campus (Darjeeling district 26.70ºN,

88.35ºE) and maintained in Plant Genetics & Molecular Breeding Laboratory, Department of Botany, University of North Bengal, India; and their two breeding lines (*Oryza sativa* x *O. rufipogon*) based on their contrasting phenotypic character (Table 1) were selected from the segregation population. Genomic DNA was extracted from four rice lines using Qiagen DNeasy Plant mini kit (Cat No. 69104). The concentration and purity of genomic DNA was quantified using Nanodrop Spectrophotometer (Thermo Scientific; 2000). The integrity of DNA was observed by agarose gel electrophoresis. DNA concentration quantified using Qubit dsDNA HS assay kit (Cat No.Q32854).

Table 1
Phenotypic characteristics of 4 rice lines

Traits	Wild	Yamuna	YR13	YR32
Culm: attitude	Spreading	Erect	Erect	Erect
Panicle: Attitude of branches	Spreading	Erect	Spreading	Spreading
Seed shattering	Shattered	Non shattered	Shattered	Non shattered
Awn	Awn	Awnless	Awn	Awnless
Hull color	Black	Straw	Straw	Straw
Pericarp color	Red	White	White	Red
Aroma	Very low	Absent	Absent	Very low
Heading time	Very late	Late	Late	Very late
Panicle length	Short	Long	Very long	Long
Grain shape	Long slender	Long bold	Basmati type	Long slender
Grain number	Low	High	High	Medium
Plant height	Very long	Medium	Very long	Very long
Salt tolerance: Germination stage	Tolerant	Moderate	Tolerant	Moderate
Submergence	Tolerant	Moderate	Tolerant	Moderate

2.2 Library construction and genome sequencing

Library preparation was performed using QIASeq FX DNA Library Preparation protocol (Cat#180475) by following manufacturer's instructions. About 40 ng of Qubit quantified DNA was enzymatically fragmented, end-repaired and A-tailed in a one-tube reaction using the FX Enzyme Mix provided in the QIASeq FX DNA kit. The end-repaired and adenylated fragments were subjected to adapter ligation, whereby index-incorporated Illumina adapter was ligated, to generate sequencing library (Supplementary Table 1). Then Illumina sequencing was carried out for the four rice samples Wild Rice, Yamuna, YR13 and YR32 (Genotypic Technology Pvt. Ltd. Bangalore, India).

2.3 Reads mapping, variant calling and annotation

Raw data was trimmed for adapter removal and low quality read filtering using Trimgalore-v0.4.01. The high quality processed reads were mapped against reference rice genomes; *Oryza sativa ssp. japonica* cultivar Nipponbare (RGAP v7) using BWA-v0.7.182 aligner. The post-processing of alignment data was performed using Samtool-v0.1.203 and Picard-v3.1.14 tool. Next, the alignment data was used for variant prediction through gatk-4.5.0.05 HaplotypeCaller. Annotation of the variants was done using snpEff-v3.3h6 program.

2.4 SNP detection, circus plot, venn diagram,gene ontology and KEGG

The SNPs predicted per sample were used to generate SNP Density plot of 1Mb window size using the R package Cmplot-v4.5.17. All the variants (SNPs, and INDELs) predicted per sample aligned were used to generate a Circos plots using the R package Circlize-v0.4.168. The reference rice proteins were further annotated using DIAMOND BlastP9 program against Uniprot Viridiplantae database with minimum 30% identity cutoff for functional assignment and gene ontology information. The reference rice genes were also annotated against KAAS10 database for retrieval of pathway information. Venn diagrams of genes having variants were generated between samples using Venny-v2.111.

2.5 SNP variation analysis of some important genes

Amino acid sequence of some important genes having locus id were downloaded from UniProt database. All these sequences were used to perform a blast homology search against protein sequences from *Oryza sativa* RGAP v7. Upon filtering the BLAST result with minimum percent identity and query coverage of 75%, all the genes showed homology against *Oryza sativa* reference proteins. Variants were filtered from the annotated files based on the matched Transcript IDs of references for each sample.

2.6 Genetic diversity analysis

Genetic diversity (F_{st}; fixation index) analysis was performed for all the variants – SNPs and InDels from samples aligned with the Nipponbare reference genome. F_{st} was calculated using Vcftools between Yamuna, YR 13 and YR 32 with wild rice. SNP-F_{st} and InDel-F_{st} was calculated in a 100 Kb window size with a step size of 10 Kb. SNP-F_{st} and InDel-F_{st} suggests a moderate differentiation. A scatter plot for Fst values across the chromosomes were generated using Rpackage ggplot2.

3. Result

3.1 Sequencing and mapping of the reads

Whole genome sequencing was performed using NGS technology of four rice lines including two progeny lines (YR13(SRX28947800) and YR32 (SRX28947801)), their parental lines Yamuna(SRX28947803) and wild rice (*Oryza rufipogon* (SRX28947802)).The sequencing coverage was ~ 18X to 27X against *Oryza sativa* reference (Table 2). The Illumina processed reads were mapped against the reference genome Nipponbare and the percentage of mapped reads for samples was in the range ~ 98.21% to 98.73%. The mean genome coverage was 98.56% with highest in YR32 (98.73%) and lowest in wild rice (98.21%) (Table 3).The predicted variants were filtered with minimum read depth 20 and average quality of 30 prior to annotation regarding protein level

change related information. A total of 57,904 to 203,568 non-synonymous variants were predicted across Wild Rice, Yamuna, YR13 and YR32 with respect to the reference rice genome (Table 4).

Table 2
Read statistics of Wild rice, Yamuna, YR13 and YR32

Sample Name	Raw Read Count	Sequencing Coverage Nipponbare RGAP(X)	Processed Read Count	%Reads Retained
Wild Rice	22946804	18.46	22717271	99.00
Yamuna	26422325	21.25	26173105	99.06
YR 13	34206155	27.51	33801374	98.82
YR 32	27702052	22.28	27440838	99.06

Table 3
Mapping statistics against Nipponbare (RGAP v7) reference genome

Sample	Number of Reads	Number of Mapped reads	%Mapped reads
Wild Rice	22717271	22311251	98.21
Yamuna	26173105	25827614	98.68
YR 13	33801374	33342796	98.64
YR 32	27440838	27092636	98.73

Table 4
Variants statistics for reference genomeNipponbare (RGAP v7)

Sample	Total variants count	Filtered variants count	Non-synonymous variants count
Wild Rice	4990155	451043	57904
Yamuna	4069471	608313	74840
YR 13	4885044	1389180	203568
YR 32	4722602	891682	123887

3.2 Variant (SNP/InDels) discovery

After removing all the false positive variants, a total of452,810 variants were obtained among which 66,513and 386,297 were homozygous and heterozygous respectively and categorically 399,570were SNPs 24,576 insertions and 28,441were deletions in case of wild rice (Fig. 1) (Supplementary table 2). In Yamuna out of 609,696 variants 216,797 were homozygous and 392899were heterozygous and categorically 540,839 were SNPs, 32704 insertions and 35864 deletions. In case of two progeny lines YR13 and YR32, the total variation was 1,393,901 and 894,666 out of which 189,256 and 113,675 were homozygous and 1,204,645 and 780,991 were heterozygous respectively. YR13 contains 1,209,121 SNPs, 87,218 insertions and 97,076 deletions and YR32 contains 781,711 SNPs, 52,324 insertions and 60,318 deletions. The distribution pattern of SNPs in the chromosome is uneven, some region plotted lightly and some region plotted densely (Fig. 2) (Supplementary Fig. 1)

3.3 Analysis and annotation of SNPs and InDels.

Among the 399,570 SNPs, 283,790 were transitions (T/C and G/A) and 115,780 were transversion (T/G,C/G, T/A, and C/A) and the transition to transversion ratio was 2.45 in wild rice. In case of Yamuna out of540,839 SNPs 386,213 were transition and 154,626 were transversion and the transition to transversion ratio was 2.49. In case of YR13 and YR32, out of 1,209,121 and 781,711 SNPs 857,297 and 556,506 were transition and 351,824 and 225,205 were transversion. The transition to transversion ratio was 2.43 and 2.47 in YR13 and YR32 respectively (Supplementary Table 3).

SNPs were annotated and categorized in intergenic regions, introns, 3'-UTR, 5'-UTR and exon based on their location. We evaluated at the SNPs that might have a significant impact on changes in protein and gene expression. In our study all rice lines showed a majority of variants in intergenic region 19.725% in wild, 19.96% in Yamuna, 20.57, 20.577% in YR13 and 20.039% in YR32. Within the coding region of the genome 9.60%, 9.63%, 7.83% and 8.60% of variants were found in wild, Yamuna, YR13 and YR32, which may be functional SNPs. Coding SNPs, exonic SNPs, and cDNA SNPs are the terms used to describe SNPs that are found in exons and the accompanying cDNAs, respectively. Synonymous SNPs (synSNPs) are exonic SNPs, which do not alter the amino acid composition of the encoded domain or protein, while non-synonymous SNPs (nsSNPs) alter the encoded amino acid (Supplementary Table 4).

Based on their functional annotation, the SNPs and InDels were primarily divided into four types: modifier (upstream, downstream, intergenic, and UTRs), low impact (synonymous coding, alternate start/stop, and start gained), moderate impact (nonsynonymous), and high impact (affecting splice site, stop, and start codons). Modifier SNPs accounted for the highest percentage of all SNPs (wild 90.27%, Yamuna 90.24%, YR13 92.04% and YR32 91.27%) and high effect SNPs showed lowest percentage of all SNPS (wild 0.61%, Yamuna 0.66%, YR13 0.52% and YR32 0.57%)(Supplementary Table 5). There were three categories for the high and moderate effect SNPs affecting the amino acids: missense, nonsense, and silent. Missense made up (wild 52.00%, Yamuna52.9%, YR13 54.10% and YR32 53.42%), nonsense (Wild 2.18%, Yamuna 2.32%, YR13 2.12% and YR32 2.22%), and silent (wild 45.80%, Yamuna 44.77%, YR13 43.77% and YR32 44.34%), of the total. Furthermore, only a few high impact SNPs and InDels were found to have an influence on the transcript's splice site (splice acceptor and splice donor), stop codon (stop site gain and stop site loss), and start codon loss (Supplementary Table 6).

3.4 Distribution of SNPs and InDels between Yamuna, Wild, YR13 and YR32

SNPs and InDels in Yamuna, wild, YR13, and YR32, in comparison to Nipponbare, were examined and presented as Venn diagrams in order to find genome-wide polymorphisms between these species. The SNPs and InDel unique to Yamuna were 540,839 and 68,568 and to Wild rice were 408,063 and 53,050 and to YR13 were 1,209,121 and 184,294 and to YR32 were781,711 and 112,642. There were 1173 unique variation in Yamuna, 973 in wild, 4041 in YR13 and 1487 in YR32 with 15137variations common to all(Fig. 3).

In this study 4 rice lines including 2 parent lines and 2 progeny lines shared 32.2% common variants of all the SNPs and InDels. Gene Ontology analysis of the protein-coding genes with variants in Yamuna, wild rice, YR13 and YR32 identified many significant GO terms (Supplementary Figs. 2, 3, 4 and 5). GO and KEGG enrichment

studies were conducted in order to better examine the probable functions that are impacted in Yamuna, wild rice and their progeny lines. Genes involved in biological processes such as transcription, transport and catabolism, lipid metabolism, signal transduction, translation, carbohydrate and amino acid metabolism were significantly enriched.

3.5 SNP in Genes/QTLs associated with agronomically important traits

The genes and it associated with QTLs for agronomically important traits were compared between 4 rice lines including two parents Yamuna and wild rice and their two progeny lines YR13 and YR32 and the single nucleotide polymorphisms corresponding to each of the genes

Table 5
SNP variations in some agronomically important gene/QTL

Traits	Name of Gene/QTL	Gene id RAPDB	Gene id RGAP	Wild rice	Yamuna	YR13	YR32
Grain number	NOG1	Os01g0752200	LOC_Os01g54860.1	0	2	7	3
Seed shattering	qSH1	Os01g0848400	LOC_Os01g62920.1	0	0	6	1
Plant height	sd1	Os01g0883800	LOC_Os01g66100.1	1	1	1	1
Grain size	GS2	Os02g0701300	LOC_Os02g47280.2	0	0	4	0
		Os02g0701300	LOC_Os02g47280.1	0	0	4	0
Hull color	Lsi1	Os02g0745100	LOC_Os02g51110.1	1	0	1	2
		Os02g0745100	LOC_Os06g12310.1	1	0	3	3
Awn	Awn3-1	Os03g0418600	LOC_Os03g30519.1	0	0	0	0
		Os03g0418600	LOC_Os03g30519.2	0	0	0	0
Heading date	Hd6	Os03g0762000	LOC_Os07g02350.1	0	2	4	5
		Os03g0762000	LOC_Os03g10940.1	1	0	4	1
		Os03g0762000	LOC_Os03g55490.2	0	0	1	2
		Os03g0762000	LOC_Os03g55490.1	0	0	1	5
Seed shattering	AP2	Os04g0649100	LOC_Os04g55560.2	0	1	4	0
Seed shattering	SHAT1	Os04g0649100	LOC_Os04g55560.3	0	1	4	0
Seed shattering	SH4	Os04g0670900	LOC_Os04g57530.1	0	0	3	0
Heading date	Hd1	Os06g0275000	LOC_Os06g16370.1	0	1	5	1
Prostrate growth	PROG1	Os07g0153600	LOC_Os07g05900.1	5	0	0	0
Aroma	Badh2	Os08g0424500	LOC_Os08g32870.1	0	0	2	1
Aroma	BAD1	Os08g0424500	LOC_Os04g39020.1	0	1	3	7
Panicle length	SP1	Os11g0235200	LOC_Os11g12740.1	0	2	0	6
Salt tolerance	Saltol	Os01g0343300		1	5	2	2
Submergence	Sub1A	Os09g0454900		0	1	3	2
	SNORKEL1	Os12g0602450		18	0	0	0

were analyzed based on the nucleotide sequence (Table 5). Os01g0752200 (grain number) gene not found in wild rice, 2 SNPs in Yamuna, 7 SNPs in YR13 and 3 SNPs in YR32. Os07g0153600 (prostrate growth) gene showed 5 SNPs in wild rice (Non synonymous) and absent in other lines. Os01g0883800 (plant height) gene showed 1 SNPs variation in each lines(Non synonymous in wild rice). Os02g0701300 (grain size) gene were absent in wild rice, Yamuna and YR32 and 4 SNPs in. Os02g0745100 (hull color) gene showed 1 SNPs in wild, absent in Yamuna, 1 SNPs in YR13 and 2 SNPs in YR32. Os03g0762000 (heading date) gene absent in wild rice and showed 2 SNPs in Yamuna (1 Non synonymous), 4 SNPs in YR13 and 5 SNPs in YR32. Os04g0649100 (seed shattering) gene showed absence in wild rice and YR32, 1 SNPs in Yamuna, 4 SNPs in YR13. Os08g0424500 (aroma) gene showed absence SNPs in wild rice and Yamuna, 2 SNPs in YR13 and 1 SNPs in YR32(Non synonymous). Os11g0235200 (panicle length) gene showed absence in wild rice and YR13, 2 SNPs in Yamuna and 6 SNPs in YR32(1 Non synonymous).Os01g0343300 (Saltol) gene showed 1 SNPs in wild rice, 5 SNPs in Yamuna, 2 SNPs in YR13 and 2 SNPs in YR32. Os09g0454900 (Sub1A) gene showed absence in wild rice, 1 SNPs in Yamuna, 3 SNPs in YR13 and 2 SNPs in YR32. Os12g0602450 (SNORKEL1) gene showed 18 SNPs in wild rice, and absence in other rice. In most of the cases the SNPs are more frequent in both breeding lines than their parents in response to a particular gene.

3.6 Genetic diversity analysis

Considering that the level of distinction between subpopulations is raised by natural selection, the genomic areas that are substantially distinct may be potential candidates for selection. The Fst value of the whole genome, which is determined by setting the sliding step length should be 10 kb and window size should be 100 kb(Fig. 4). On examining genetic differentiation of 4 rice lines the average Fst value is 0.012 for SNPs and 0.016 for InDels and weighted Fst value is 0.022 for SNPs and 0.029 for InDels (Table 6).

Table 6
F_{st}statistics of Yamuna, YR13, YR32 and Wild rice (*Oryza rufipogon*)

F_{st} estimates	**SNP-F_{st}**	**INDEL-F_{st}**
Weir and Cockerham mean Fst estimate	0.012466	0.016804
Weir and Cockerham weighted Fst estimate	0.022798	0.029989

4. Discussion

The rice genome contains millions of SNPs, according to a genome-wide study based on sequencing of two cultivars, Nipponbare and 93 – 11 [12, 13]. Only small number of rice cultivars like Omachi [10], Eiko and Rikuu 132 [14] Bengal and Nona Bokra [15] Koshihikari [16] have undergone whole-genome resequencing amid the quick advancements in next-generation sequencing technology. In order to increase the pool of DNA polymorphisms (SNPs and InDels) present throughout the rice genome, which could be used for genetic analysis as well as in the improvement of rice, emphasis has been placed on and efforts are being made to sequence a diverse set of more rice genotypes through whole-genome sequencing [17].

In the current study, the whole genome of four ricelines including two parents and two progeny lines were resequenced and mapped to Nipponbare as areference genome in order to find genome-wide DNA variations.

The uniquely mapped reads from these lines covered > 98% of the reference genome. The highest mapped reads on to the reference genome was shown in YR 32 and lowest in wild rice.

The distribution pattern and variation analysis of SNPs across the 12 chromosomes were not uniform. These irregular SNP and InDels distributions indicate to the existence of low SNP areas, which have been referred to as "SNP deserts" in other reports [18, 19, 20].It is thought that SNP deserts are important locations for the domestication process and have well preserved functionality. According to Tenaillon et al. [21] greater SNP rates may be associated with a higher level of diversity. In all four datasets, there was more transition SNPs than transversion SNPs. It was previously observed [22, 19, 23] that transitional SNPs were more common than transversion SNPs, and this finding has been termed "transition bias." Protein architectures are conserved to provide conformational benefits in natural selection, as evidenced by the larger ratio of transitional SNPs over transversion SNPs [24, 25, 26]. Previous research on the mutant rice (var. Jugal) revealed that, in contrast to regular rice, it had more carpels and larger floral meristems [27]. Ts / Tv ratio ranging from 2.42 to 2.49 which was higher than earlier reports [22], he inclusion of SNPs from the entire rice genome may be the cause of the higher ratio. The research results indicated that both progeny lines had much more SNP variants than their parent lines, which might be the result of genome mixing.

SNPs are most often found in intra-genic areas if they have functional impacts on phenotypic features. As a result, we categorized SNPs according to their genomic positions. Over 19% of the putative SNPs were found in intergenic regions, whereas roughly 5% were found in coding areas.Analysis of SNPs or InDels within individual linesindicated that YR13 had the highest number ofpolymorphic SNPs and InDels, 1,209,121 and 184,294 whereas wild ricehadthe lowest number of polymorphic SNPs and InDels, 408,063 and 53,050.SNPs or InDels unique to each line were identified for Yamuna (1173), wild (973), YR13 (4041) and YR32 (1487). The variations common to all is15137 accounting 32.2% all the SNPs and InDels. Indicate that the 4 lines i.e. Yamuna, Wild,YR13 and YR32 were diverged from each other. Because the ability to execute heterosis depends on the genetic variations between parents, genetic variety is essential to hybrid rice projects that aim to develop heterotic rice hybrids [28].GO is an essential bioinformatics technique that uses controlled term organization to try and understand the function of genes or proteins.

Sequence comparison of the genes related with agronomically important characters revealed polymorphism in terms of SNPs. Environmental factor like soil pH, temperature, and natural variation in agronomic traits could influence traits expression and SNP associations. Thus further study across different environment would be beneficial to validate the applicability of these SNPs. The SNPs which shows non synonymous effect, alter the codon and codes different amino acid which ultimately change the protein sequence and its function. As wild rice showed prostrate growth, it's PROG1 (Os07g0153600) gene showed SNPs variations with Nipponbare reference genome, may support significant SNPs are present in wild rice gene. Grain size (GS2) gene in YR13 showed 4 SNPs variation may be the reason for higher grain size i.e. basmati type. Heading time (Hd1, Hd6) of wild rice is much higher than other lines; this may the reason for SNPs variation. Panicle length of 4 lines was different from each other, number of SNP variation also differs. Wild rice has higher salt tolerance trait, so it may be the reason of lesser SNPs variation in comparison to other rice lines. Previous studies have found SNPs associated to salt tolerance traits like grain weight, spikelet fertility, often implicating candidate genes associated to stress response pathways [29].SNPs variation does not means the alteration of gene expression. The number of SNPs in a particular gene/QTL may be synonymous or non synonymous. If it is synonymous

than it does not alter amino acid, so there is no change in phenotypic expression. If the SNPs is non synonymous, it alters the amino acid sequence during translation and may alter the phenotypic expression [30].On the other hands transcription factor often identify a significant number of binding sites, an SNP typically has little effect on expression or the interaction between the TF and binding site. Allele-specific gene expression can occasionally result from an SNP that increases or decreases the binding. Rarely, an SNP may create a new binding site or remove the native binding site. As a result, the original TF can no longer control the gene. Functional SNPs in TF binding sites may therefore predictably result in variations in phenotypes and gene expression [31].

Wright [32] proposed that pairwise F_{st} values greater than 0.15 showed significant genetic difference. The average observed heterozygosity was 0.012 for SNPs and 0.016 for InDels, which was less than the results of Tarang et al. [33] and Suvi et al. [34]. This low estimate can be attributed to rice's autogamous reproduction. However, Nachimuthu et al. [35] found that the average anticipated heterozygosity, which was calculated to be 0.524, was consistent with their findings. This is explained by the fact that a wide range of genetic variety is produced by the exchange of genes between genotypes. The calculated F_{st} value was also less than that of wheat (*Triticum aestivum*; F_{st} = 0.15), or maize (F_{st} = 0.14; [36]). These results indicate that the rice lines chosen for this study have a moderate level of genetic diversity.

5. Conclusion

The whole-genome sequencing (WGS)-based identification of SNPs and InDels in the *Oryza sativa* × *O. rufipogon* breeding lines provides a comprehensive insight into the underlying genetic diversity and allelic variations introgressed from the wild progenitor. These variants are crucial for identifying novel alleles, mapping agronomic traits, and enhancing rice breeding through MAS and introgression. The results demonstrate that wild rice (*O. rufipogon*) contributes a rich reservoir of novel alleles that enhance genomic variability, offering unique opportunities for broadening the genetic base of cultivated rice. The detected SNPs and InDels reveal the extent of divergence between the parental genomes along with breeding lines. The rich genetic reservoir of *O. rufipogon* offers significant potential for improving cultivated rice, especially for traits like disease resistance and yield. Continued advancements in sequencing technologies and pangenome analyses will further unlock the potential of wild rice for global food security. Whole-genome sequencing of *Oryza sativa* x *O. rufipogon* breeding lines conclusively demonstrates that *O. rufipogon* is a wealthy reservoir of genetic diversity, harboring significantly higher SNP and InDel variation compared to cultivated rice. These variants, particularly in genes like abiotic stress Saltol, SNORKEL1, for aroma-BAD1, and yield-related loci (GS2, Hd1, NOG1) enable precise marker-assisted selection and introgression of novel alleles to enhance agronomic traits. The extensive polymorphisms and MNP variants identified underscore the critical role of wild rice in overcoming domestication bottlenecks, offering transformative potential for improving rice yield, resilience, and adaptability to meet global food security challenges.

Declarations

CRediT authorship contribution statement

Shukdeb Sarkar: Writing –original draft, review & editing, Validation, Methodology, Resources, Data curation, Software.**Jayita Sarkar**: Writing – review &editing. **Bhumika T Yonzon**: Writing – review &editing.**Sona Limboo**- Writing – review &editing**; Subhas Chandra Roy**: Conceptualization, designing of research methodology, overall supervision, manuscript review and editing, validation, germplasm resources selection, and project administration.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Acknowledgements

We greatly acknowledged the financial support received from DST, Govt. of India for providing Advanced Instruments Facility (DST-FIST Instrument facility; Sanction No. SR/FST/LS-I/2021/900 dt.25.03.2022), in the Department of Botany, University of North Bengal.

Data Availability

The whole genome sequence data of four rice samples was submitted to NCBI archives and are available with accession number - Yamuna (SRX28947803), wild rice (*Oryza rufipogon* [SRX28947802]), YR 13 (SRX28947800) and YR 32 (SRX28947801). The data generated from whole genome sequence are included in this article and its supplementary information files.

Ethics approval and Consent to participate

The cultivated rice Yamuna was collected from farmers and is available locally. No special permissions were required under local regulations or institutional guidelines. The experimental research on wild rice (*Oryza rufipogon*) complied with national and institutional guidelines for research involving plants. The wild rice was taxonomically identified by Prof. Manoranjan Chowdhury, Department of Botany, University of North Bengal, Darjeeling, India. A voucher specimen of wild rice (*Oryza rufipogon*) (Accession no. 13048) has been deposited in the NBU Herbarium, Department of Botany, University of North Bengal, Darjeeling, India. We developed YR13 and YR32 breeding lines through interspecific hybridization (*Oryza sativa* x *O. rufipogon*) and maintained in Plant Genetics & Molecular Breeding Laboratory, Department of Botany, University of North Bengal. The study needs no further affirmations.

Clinical trial

Not applicable.

Consent to Publication

Not applicable.

Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Jones, N., Ougham, H., Thomas, H. and Pašakinskienė, I., 2009. Markers and mapping revisited: finding your gene. New phytologist, 183(4).
2. Bentley, D.R., 2006. Whole-genome re-sequencing. Current opinion in genetics & development, 16(6), pp.545-552.
3. Chen, H., He, H., Zou, Y., Chen, W., Yu, R., Liu, X., Yang, Y., Gao, Y.M., Xu, J.L., Fan, L.M. and Li, Y., 2011. Development and application of a set of breeder-friendly SNP markers for genetic analyses and molecular breeding of rice (Oryza sativa L.). Theoretical and applied genetics, 123(6), pp.869-879.
4. Osman, A., Jordan, B., Lessard, P.A., Muhammad, N., Haron, M.R., Riffin, N.M., Sinskey, A.J., Rha, C. and Housman, D.E., 2003. Genetic diversity of Eurycomalongifolia inferred from single nucleotide polymorphisms. Plant Physiology, 131(3), pp.1294-1301.
5. Edwards, D. and Batley, J., 2010. Plant genome sequencing: applications for crop improvement. Plant biotechnology journal, 8(1), pp.2-9.
6. Kim, M.Y., Lee, S., Van, K., Kim, T.H., Jeong, S.C., Choi, I.Y., Kim, D.S., Lee, Y.S., Park, D., Ma, J. and Kim, W.Y., 2010. Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine sojaSieb. and Zucc.) genome. Proceedings of the National Academy of Sciences, 107(51), pp.22032-22037.
7. Debsharma, S.K., Rahman, M.A., Quddus, M.R., Khatun, H., Disha, R.F., Roy, P.R., Ahmed, S., El-Sharnouby, M., Iftekharuddaula, K.M., Aloufi, S. and Alzuaibr, F.M., 2022. SNP based trait characterization detects genetically important and stable multiple stress tolerance rice genotypes in salt-stress environments. Plants, 11(9), p.1150
8. Project, I.R.G.S., 2005. The map-based sequence of the rice genome. Nature, 436(7052), pp.793-800.
9. Yamamoto, T., Nagasaki, H., Yonemaru, J.I., Ebana, K., Nakajima, M., Shibaya, T. and Yano, M., 2010. Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. BMC genomics, 11(1), p.267.
10. Arai-Kichise, Y., Shiwa, Y., Nagasaki, H., Ebana, K., Yoshikawa, H., Yano, M. and Wakasa, K., 2011. Discovery of genome-wide DNA polymorphisms in a landrace cultivar of japonica rice by whole-genome sequencing. Plant and cell physiology, 52(2), pp.274-282.
11. Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F., Li, X., Dong, Y., Gutenkunst, R.N., Fang, L., Huang, L. and Li, J., 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature biotechnology, 30(1), pp.105-111.
12. Feltus, F.A., Wan, J., Schulze, S.R., Estill, J.C., Jiang, N. and Paterson, A.H., 2004. An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. Genome research, 14(9), pp.1812-1819.
13. Shen, Y.J., Jiang, H., Jin, J.P., Zhang, Z.B., Xi, B., He, Y.Y., Wang, G., Wang, C., Qian, L., Li, X. and Yu, Q.B., 2004. Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiology, 135(3), pp.1198-1205.

14. Nagasaki, H., Ebana, K., Shibaya, T., Yonemaru, J.I. and Yano, M., 2010. Core single-nucleotide polymorphisms—a tool for genetic analysis of the Japanese rice population. Breeding science, 60(5), pp.648-655.
15. Chai, C., Shankar, R., Jain, M. and Subudhi, P.K., 2018. Genome-wide discovery of DNA polymorphisms by whole genome sequencing differentiates weedy and cultivated rice. Scientific reports, 8(1), p.14218.
16. Tomita, M., 2024. Whole genome sequencing and trait investigation revealed a semidwarf isogenic rice variety with a genome and grain quality similar to the high-demand cultivar Koshihikari. Genet Mol Res, 23(1).
17. McCouch, S.R., Zhao, K., Wright, M., Tung, C.W., Ebana, K., Thomson, M., Reynolds, A., Wang, D., DeClerck, G., Ali, M.L. and McClung, A., 2010. Development of genome-wide SNP assays for rice. Breeding Science, 60(5), pp.524-535.
18. Wang, L., Hao, L., Li, X., Hu, S., Ge, S. and Yu, J., 2009. SNP deserts of Asian cultivated rice: genomic regions under domestication. Journal of evolutionary biology, 22(4), pp.751-761.
19. Hu, Y., Mao, B., Peng, Y., Sun, Y., Pan, Y., Xia, Y., Sheng, X., Li, Y., Tang, L., Yuan, L. and Zhao, B., 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(3), pp.303-315.
20. Singhabahu, S., Wijesinghe, C., Gunawardana, D., Senerath-Yapa, M.D., Kannangara, M., Edirisinghe, R. and Dissanayake, V.H.W., 2017. Whole genome sequencing and analysis of Godawee, a salt tolerant indica rice variety. J Rice Res, 5(177), p.2.
21. Tenaillon, M.I., Sawkins, M.C., Long, A.D., Gaut, R.L., Doebley, J.F. and Gaut, B.S., 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). Proceedings of the National Academy of Sciences, 98(16), pp.9161-9166.
22. Subbaiyan, G.K., Waters, D.L., Katiyar, S.K., Sadananda, A.R., Vaddadi, S. and Henry, R.J., 2012. Genome-wide DNA polymorphisms in elite indica rice inbreds discovered by whole-genome sequencing. Plant Biotechnology Journal, 10(6), pp.623-634.
23. Hwang, S.G., Hwang, J.G., Kim, D.S. and Jang, C.S., 2014. Genome-wide DNA polymorphism and transcriptome analysis of an early-maturing rice mutant. Genetica, 142(1), pp.73-85.
24. Wakeley, J., 1996. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Trends in ecology & evolution, 11(4), pp.158-162.
25. Rathinasabapathi, P., Purushothaman, N., Vl, R. and Parani, M., 2015. Whole genome sequencing and analysis of Swarna, a widely cultivated indica rice variety with low glycemic index. Scientific Reports, 5(1), p.11303.
26. Zhang, F., Xu, T., Mao, L., Yan, S., Chen, X., Wu, Z., Chen, R., Luo, X., Xie, J. and Gao, S., 2016. Genome-wide analysis of Dongxiang wild rice (OryzarufipogonGriff.) to investigate lost/acquired genes during rice domestication. BMC plant biology, 16(1), p.103.
27. Das, S.P., Deb, D. and Dey, N., 2018. Micromorphic and molecular studies of floral organs of a multiple seeded rice (Oryza sativa L.). Plant Molecular Biology Reporter, 36(5), pp.764-775.
28. Zheng, W., Ma, Z., Zhao, M., Xiao, M., Zhao, J., Wang, C., Gao, H., Bai, Y., Wang, H. and Sui, G., 2020. Research and development strategies for hybrid japonica rice. Rice, 13(1), p.36.

29. Era, F. M., Raihan, M. S., Jahan, N., Pandey, S., Alalawy, A. I., Al-Duais, M. A., ... & Islam, A. A. (2024). Identification of significant SNPs for yield-related salt tolerant traits in rice through genome-wide association analysis. *Cellular and Molecular Biology*, *70*(12), 18-25.
30. Hunt, R., Sauna, Z.E., Ambudkar, S.V., Gottesman, M.M. and Kimchi-Sarfaty, C., 2009. Silent (synonymous) SNPs: should we care about them?. Single nucleotide polymorphisms: methods and protocols, pp.23-39.
31. Chorley, B. N., Wang, X., Campbell, M. R., Pittman, G. S., Noureddine, M. A., & Bell, D. A. (2008). Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. *Mutation Research/Reviews in Mutation Research*, *659*(1-2), 147-157.
32. Wright, S. (1984). Variability within and among natural populations. *Evol Genet Popul*, *4*.
33. Tarang, A. and Gashti, A.B., 2016. The power of microsatellite markers and AFLPs in revealing the genetic diversity of Hashemi aromatic rice from Iran. Journal of Integrative Agriculture, 15(6), pp.1186-1197.
34. Suvi, W.T., Shimelis, H., Laing, M., Mathew, I. and Shayanowako, A.I.T., 2020. Assessment of the genetic diversity and population structure of rice genotypes using SSR markers. ActaAgriculturaeScandinavica, Section B—Soil & Plant Science, 70(1), pp.76-86.
35. Nachimuthu, V.V., Muthurajan, R., Duraialaguraja, S., Sivakami, R., Pandian, B.A., Ponniah, G., Gunasekaran, K., Swaminathan, M., KK, S. and Sabariappan, R., 2015. Analysis of population structure and genetic diversity in rice germplasm using SSR markers: an initiative towards association mapping of agronomic traits in Oryza sativa. Rice, 8(1), p.30.
36. Jiao, Y., Zhao, H., Ren, L., Song, W., Zeng, B., Guo, J., Wang, B., Liu, Z., Chen, J., Li, W. and Zhang, M., 2012. Genome-wide genetic changes during modern breeding of maize. Nature genetics, 44(7), pp.812-815.

Figures

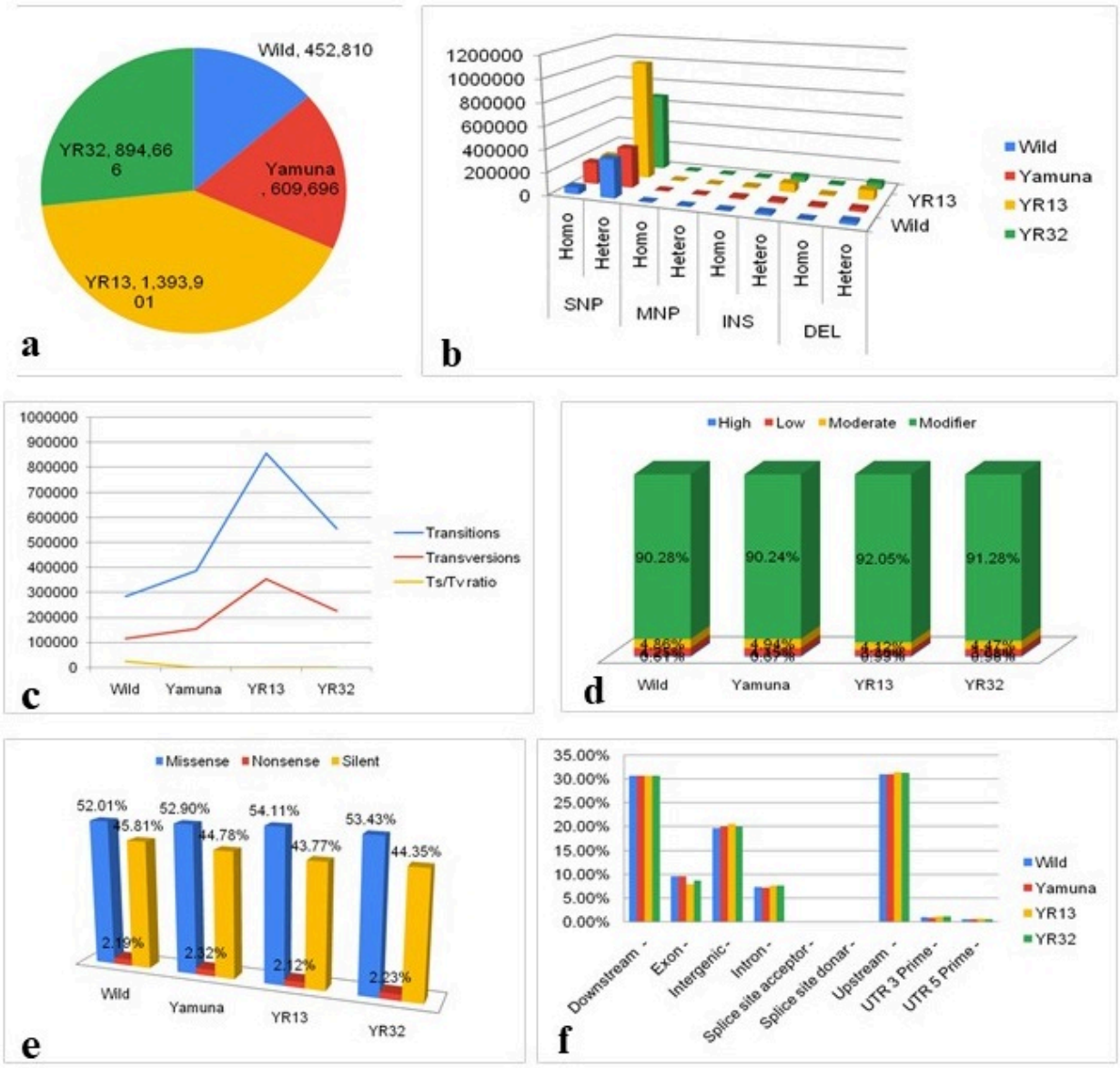

Figure 1

Types of variants **(a)** Total number of changes by type **(b)** Number of changes by type for Wild rice, Yamuna, YR13 and YR32 **(c)**Ts/Tv (Transitions/Transversions) for Wild rice, Yamuna, YR13 and YR32 **(d)**Number of effects by impact for Wild rice, Yamuna, YR13 and YR32 **(e)**Number of effects by functional class for Wild rice, Yamuna, YR13 and YR32 **(f)**Number of effects by type and region for Wild rice, Yamuna, YR13 and YR32

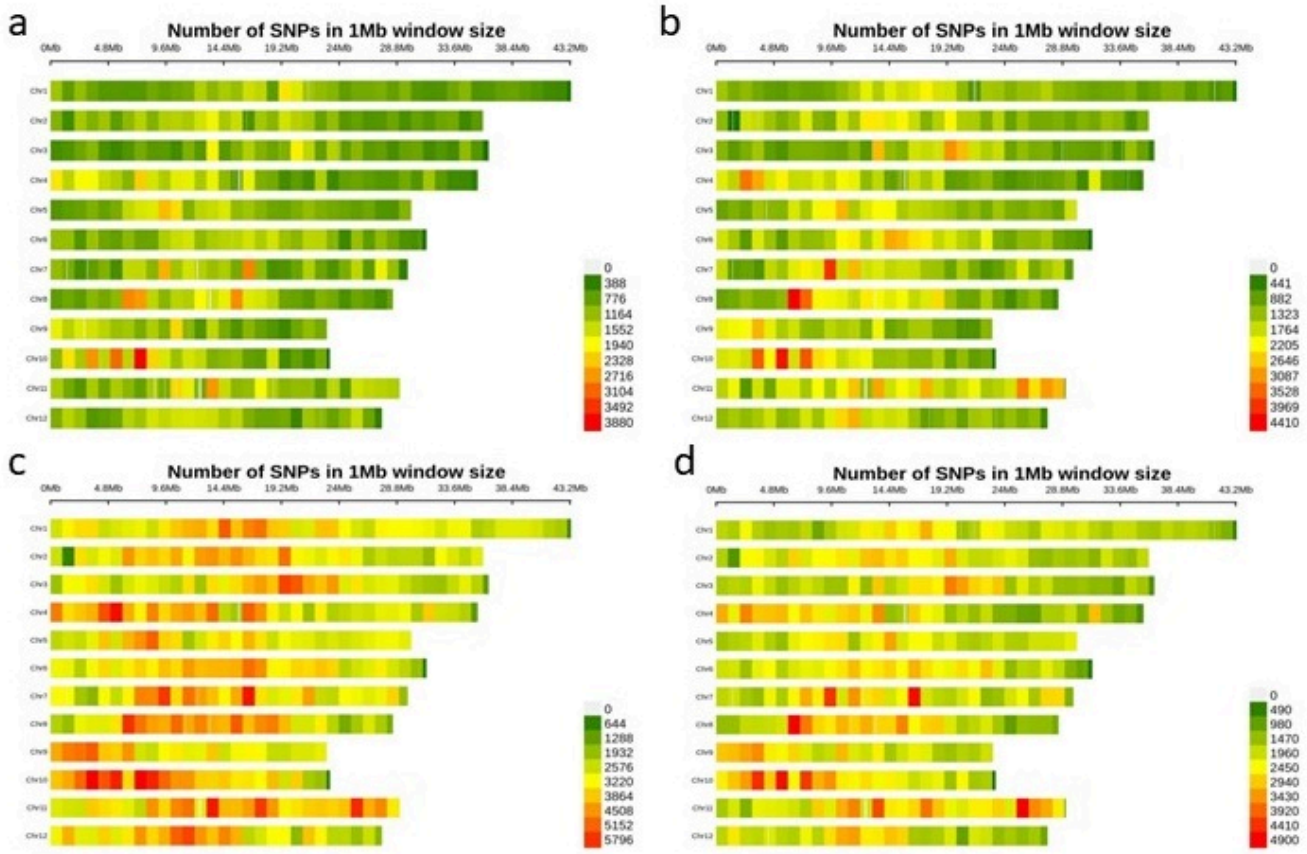

Figure 2

SNP density plot across 12 chromosomes representing number of SNPs within 1 Mb window size. The horizontal axis shows the chromosome length (Mb); the different color depicts SNP density in rice parental lines: **a)** Yamuna (**b)** Wild rice (**c)** YR13 (**d)** YR32

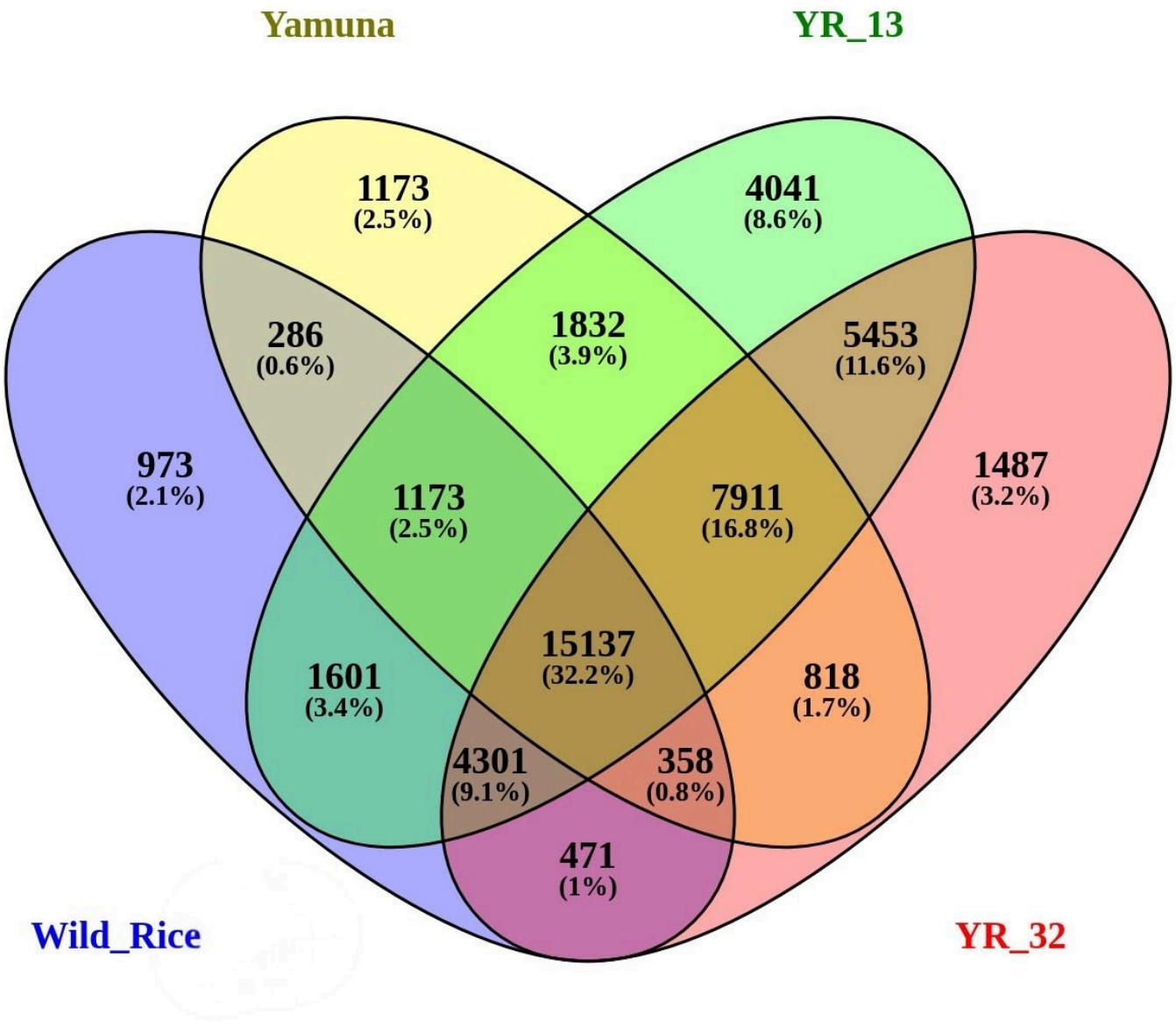

Figure 3

Unique and common genes having SNPs/InDELs variation in four rice lines Yamuna, Wild rice, YR13 and YR32 using whole genome sequencing data depicted in the venn diagram.

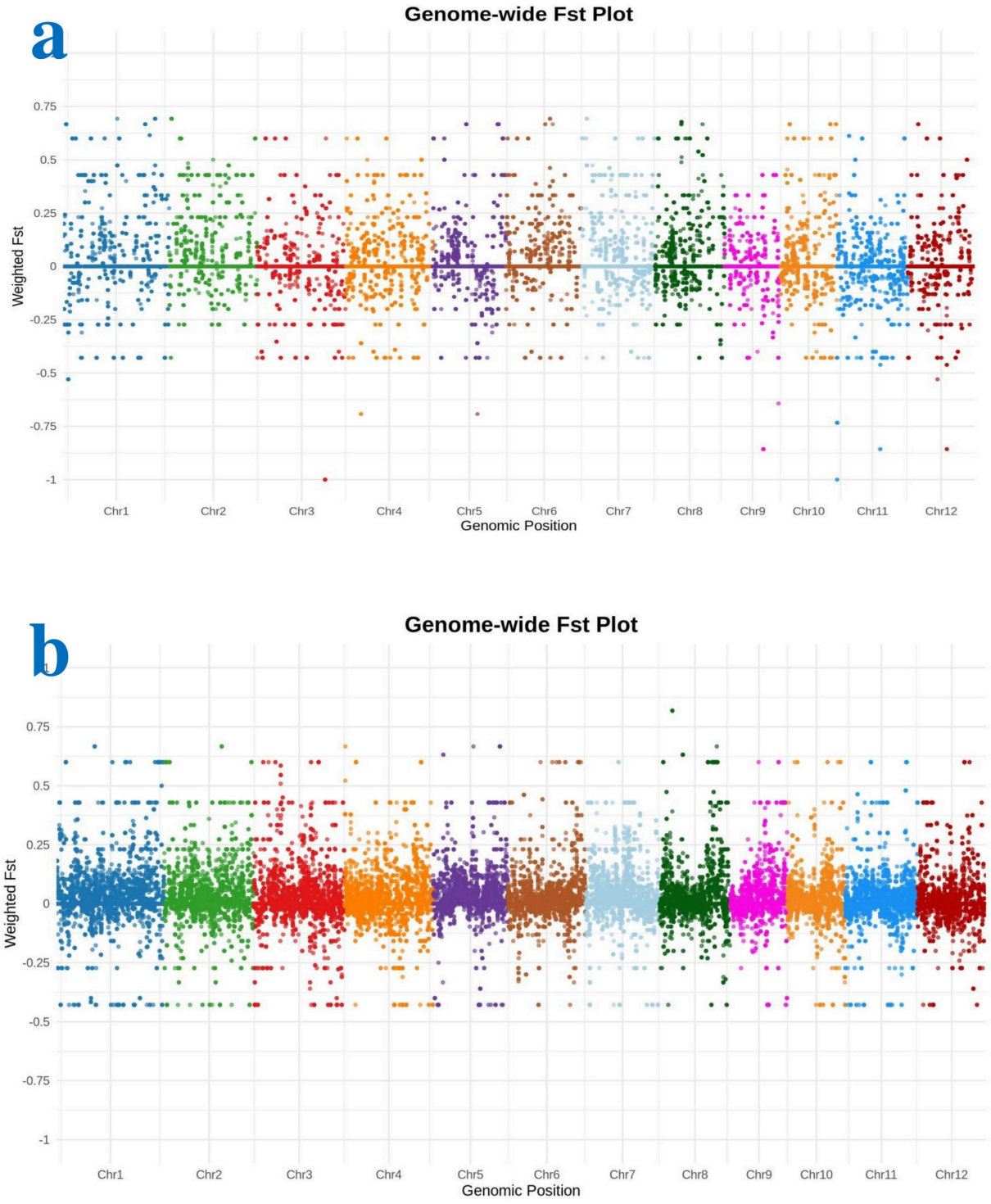

Figure 4

Genetic diversity analysis on the basis of Fst scatter plot in four rice lines(Yamuna,wild rice,YR 13 and YR 32) using whole genome sequencing data (**a**) INDEL-F_{st}scatter plot and (**b**) SNP-F_{st}Scatter plot.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfile.docx
- GraphicalAbstract.jpg
- SupplementaryFigure1.jpg