CAHIERS MAGELLANES-NS
Volume 06 Issue 2
2024

ISSN:1624-1940
DOI 10.6084/m9.figshare.261421
http://magellanes.com/

EXPLORATION OF GENETIC VARIATIONS IN KETEKI JOHA RICE (*ORYZA SATIVA* L.) THROUGH RNA-SEQUENCING

Kanishka Purkait[1], Nickolsova Handique[1], Uddipta Borthakur[1,2]*, Nibedita Sarma[2], Manish Raj Mishra[3], Manashi Kalita[1]

[1]Department of Botany, Handique Girls' College, Guwahati, 781001, Assam, India
[2]Department of Botany, Gauhati University, Guwahati, 781014, Assam, India
[3]Department of Molecular Biology and Biotechnology, Tezpur University, Sonitpur, 784028, Assam, India.

***Corresponding Author:** Uddipta Borthakur
Address: Department of Botany, Gauhati University, Guwahati, 781014, Assam, India

Abstract

The study investigates genetic variations, particularly single nucleotide polymorphisms (SNPs) and insertions/deletions (InDels), in the Keteki Joha rice variety using RNA-sequencing data. By pre-processing raw reads, mapping them to the Japonica genome, and applying variant calling tools, a comprehensive analysis of the genetic diversity within Keteki Joha rice was performed. The study identified a total of 342,641.3 variants, with SNPs constituting 68% and InDels 32%. Chromosome 1 exhibited the highest number of variants, while chromosome 9 had the least. The research also examined the functional impact of these variants, revealing significant effects on gene structure and protein function. This exploration of genetic diversity in Keteki Joha provides insights into its adaptability and potential for breeding programs, emphasizing the importance of conserving this unique rice variety.

Keywords: *Keteki Joha, genetic variation, SNPs, InDels, RNA-sequencing*

Introduction

Rice (*Oryza sativa* L.) is a staple food that is eaten by nearly half of the world's population. Asian nations cultivate and consume 90% of the world's rice (Tyagi et al., 2004). A wide variety of rice cultivars naturally occur in India (Samal et al., 2014).

Assamese Joha rice is a special variety of fragrant rice that is grown as winter rice and is highly prized for its exceptional quality. Joha rice has superior cooking qualities, a distinct scent, a superfine kernel, and outstanding palatability (Das et al., 2010). Joha rice varieties have a non-functional betaine aldehyde dehydrogenase 2 (BADH2), which also reduces grain production and gives them their scent (Kovach et al., 2009).

Joha rice has a limited production potential and is susceptible to photoperiod. Its low productivity and yield make it unfavourable for growing, even if it tastes and smells great. As a result, the researchers are least interested in it. The average Joha rice yields between one and 1.5 t/ha (Das et al., 2010). It is estimated that Assam is home to between 50 and 80 aromatic varieties of Joha rice, many of which have disappeared from the face of the earth (Das et al., 2010). An enhanced variety of Joha rice called Keteki Joha (Savitri X Badshabhog) was created by the Regional Agricultural Research Station (RARS) in Titabar, Assam, India. Consequently, Keteki Joha is the original hybrid Joha variety and is said to

ISSN:1624-1940
DOI 10.6084/m9.figshare.261421
http://magellanes.com/

have yield potential that is 3.5 times higher than standard types (Das et al., 2010). The average yield potential of Ketaki Joha was found to be 4.5 t/ha (Rice Knowledge Bank, Assam)

Genetic variation plays a crucial role in shaping the diversity of living organisms, and single nucleotide polymorphisms (SNPs) and insertions/deletions (InDels) are the most common types of genetic variation found in genomes (Bhangale et al., 2005). The advent of high-throughput sequencing technologies has made it possible to identify and characterize genetic variation at a genome-wide scale (Bernstein et al., 2012) and RNA sequencing (RNA-seq) has emerged as a powerful tool for this purpose (Wang et al., 2019).

RNA-seq is a technique that allows researchers to capture a snapshot of the transcriptome, the complete set of RNA molecules that are transcribed from a genome, in a particular cell type or tissue at a given time. By comparing RNA-seq data from different individuals or populations, it is possible to identify genetic variants that affect gene expression or splicing, as well as to quantify gene expression levels and detect alternative splicing events.

SNPs and InDels are of particular interest because they are highly abundant in genomes and can have significant functional consequences. SNPs can alter the amino acid sequence of a protein, affect protein stability or activity, or influence RNA processing (Pai et al., 2012) whereas InDels can cause frameshifts (Kunkel and Bebenek, 2000; Huang et al., 2002; Kunkel,1985) that lead to truncated or altered protein products, or affect splicing by disrupting splice sites or creating new ones.

In recent years, several studies have used RNA-seq to identify and characterize SNPs and InDels in a variety of species, including humans (Lezmi and Benvenisty, 2021), model organisms (Jeena, 2021) and non-model organisms (Espinosa et al., 2020). These studies have revealed a wealth of new genetic variation and have provided insights into the functional consequences of this variation. For example, RNA-seq data has been used to identify SNPs that affect gene expression in cancer cells, to detect InDels that cause genetic disorders in humans, and to discover novel splice sites that affect gene function in plants (Wang et al., 2024).

Therefore, this study aims to explore the genetic variants, specifically single nucleotide polymorphisms (SNPs) and insertions/deletions (InDels), in Keteki Joha rice. Understanding these genetic variations is crucial for assessing the genetic diversity within this unique rice variety, which is known for its distinctive aroma and flavour. By analyzing these genetic markers, we aim to uncover insights into the genetic architecture of Keteki Joha rice, contributing to the broader understanding of its adaptability, resilience, and potential for breeding programs. This research not only highlights the significance of genetic diversity in rice cultivation but also aims to support efforts in conservation and sustainable agricultural practices.

Materials and methods

Pre-processing of raw reads and mapping

Pre-sequenced RNA seq data performed by Illumina HiSeq 2000 platform (Illumina, USA) were downloaded from NCBI (National Center for Biotechnology Information) with 3 biological replicates with the accession numbers SRR13123865, SRR13123866, SRR13123867 .

ISSN:1624-1940
DOI 10.6084/m9.figshare.261421
http://magellanes.com/

The quality of sequencing raw reads was evaluated using FastQC (Andrews, 2010). Based on the quality control report, the sequencing data were filtered and trimmed to remove low-quality reads (Phread 33 score <25) sequences that might negatively impact downstream analysis. After quality control, adapter sequences were removed using the software Trimmomatic (Bolger et al., 2014).
The raw reads were mapped to the Joponica genome retrieved from Rice Genome Annotation Project database (http://rice.plantbiology.msu.edu/)). Briefly, index was generated for the genome and reference-based mapping was performed using STAR v2.14 (Dobin et al., 2013). Samtools was used to manipulate the mapped reads where, Integrative Genome Browser was used to visualize the mapped reads graphically (Robinson et al., 2011).

Variant Calling and Filtering

For variant calling the raw reads were mapping using the 2-PASS mode of STAR. After alignment, the resulting alignment files were often processed further to remove duplicates, filter low-quality reads, and sorted the resulting files for downstream analysis using Picard tools. This step is important for removing PCR artifacts and ensuring accurate variant calling. Used a variant calling tool GATK to identify SNPs and InDels from the aligned reads (McKenna et al., 2010; DePristo et al., 2011). This tool compared the aligned reads to the reference genome or transcriptome and identified sites where there were differences. Filters were applied to the identified variants to remove any false positives and retain only high-confidence variants. Common filtering criteria include depth of coverage, allele frequency, and quality score.

Annotation of variant and their effects

A database was built for the reference genome using SNPeff's built-in database builder. This step is important to allow SNPeff to efficiently and accurately annotate the variants. Annotate variants: SNPeff (Cingolani et al., 2012) was used to annotate the variants using the built-in database. The output from this step includes information on the location of the variant, its effect on the gene structure, and its functional impact on the protein.

Predict the effect of variants: Predicted the functional impact of variants using SNPeff's built-in algorithms. These algorithms predicted the effect of variants on protein function based on various criteria, such as the type of amino acid change, conservation, and structural features.

Results

Mapping statistics of RNA-sequencing data of Keteki Joha rice

The RNA-sequencing data analysis for Keteki Joha rice revealed the following mapping statistics across three replicates. Each replicate had around 18 million input reads with an average read length of 300 bases. Uniquely mapped reads constituted approximately 84.22% across all replicates. The average mapped read length was consistent at 296.5 bases. The total number of splices averaged around 4.9

ISSN:1624-1940
DOI 10.6084/m9.figshare.261421
http://magellanes.com/

million, with annotated splices accounting for the majority (around 4.6 million). The mismatch rate per base was low, averaging 0.45%, while the deletion and insertion rates per base were 0.05% and 0.04%, respectively. Multi-mapping reads accounted for about 4.4%, and around 10% of reads were too short to be mapped. No chimeric reads were detected, indicating the high quality of the sequencing data

Table 1. Mapping statistics across three replicates of RNA-sequencing data of Keteki Joha rice

Metric	Replicate 1	Replicate 2	Replicate 3	SD	SE
Number of input reads	18,514,276	18,914,076	18,714,176	200,400	115,800
Average input read length	300	300	300	0	0
Uniquely mapped reads number	15,279,341	15,559,239	15,419,290	139,949	80,947
Uniquely mapped reads ratio	84.22%	84.22%	84.22%	0%	0%
Average mapped length	296.44	296.55	296.50	0.055	0.032
Number of splices: Total	4,827,652	4,984,399	4,905,025	78,747	45,548
Number of splices: Annotated (sjdb)	4,276,144	4,725,215	4,500,680	224,535	129,732
Number of splices: GT/AG	4,751,345	4,902,822	4,827,083	75,739	43,748
Number of splices: GC/AG	44,056	40,091	42,073	1,982	1,145
Number of splices: AT/AC	1,727	4,661	3,194	1,467	847
Number of splices: Non-canonical	30,524	30,225	30,375	149	86
Mismatch rate per base	0.44%	0.46%	0.45%	0.01%	0.01%
Deletion average length	1.28	1.28	1.28	0	0
Insertion rate per base	0.04%	0.04%	0.04%	0%	0%
Insertion average length	1.18	1.13	1.16	0.03	0.02
Number of reads mapped to multiple loci	864,440	762,219	813,330	50,890	29,405
% of reads mapped to multiple loci	4.76%	4.20%	4.34%	0.28%	0.16%
Number of reads mapped to too many loci	58,131	36,735	47,433	10,699	6,186
% of reads mapped to too many loci	0.32%	0.20%	0.26%	0.06%	0.03%
Number of reads unmapped: too short	1,939,482	1,673,043	1,806,263	133,220	76,961
% of reads unmapped: too short	10.69%	9.22%	9.66%	0.73%	0.42%
Number of reads unmapped: other	82	469	275	194	112
% of reads unmapped: other	0.00%	0.00%	0.00%	0.00%	0.00%
Number of chimeric read pairs	0	1	0	0	0
% of chimeric reads	0.00%	0.00%	0.00%	0.00%	0.00%

Analysis of variant count and variant rate of RNA-sequencing data of Keteki Joha rice

The total variants so obtained from the average of three samples of both InDels and SNPs together came out to be 342641.3 out of which 236332 (68%) were SNPs and 106309.3 (32%) were InDels at an average. *Oryza sativa* has a total of 12 chromosomes out of which chromosome number 1 had the highest number of SNPs with a count of 28,975 average. On the other hand, chromosome number 5 with an average of 11,993. The highest number of InDels was observed on chromosome number 1 as well with an average of 14,913 InDels and the least was observed on chromosome 9 with an average of 5,663. An average of 116 SNPs and 58 InDels were observed on synonymous chromosome (Sy) and 188 SNPs 27 InDels on Unmapped chromosome(Un) (Table.4.2.1) As a result, the rate of variants was obtained from the variant number and highest rates were observed on Un chromosome and the least at chromosome number 3 of InDels. While the highest rates were obtained on Sy chromosome and the least at chromosome 6 of SNPs. The rates are shown graphically on the Table 1.

CAHIERS MAGELLANES-NS
Volume 06 Issue 2
2024

ISSN:1624-1940
DOI 10.6084/m9.figshare.261421
http://magellanes.com/

Table 2. Showing total variants (i.e. SNPs and InDels) of three replicates of RNA Seq data of Keteki Joha .

Chromosome	Replica 1		Replica 2		Replica 3	
	InDels	SNPs	InDels	SNPs	InDels	SNPs
1	15,704	29,182	14,518	28,624	14,518	29,118
2	13,446	27,415	12,502	26,943	12,502	28,203
3	13,771	23,898	13,116	23,228	13,116	23,828
4	8,993	17,243	8,511	17,014	8,511	17,448
5	8,215	12,350	7,690	11,640	7,690	11,988
6	10,362	27,377	9,804	27,308	9,804	28,103
7	8,306	18,723	8,028	18,948	8,028	19,050
8	7,520	18,094	7,227	17,938	7,227	18,321
9	5,804	13,070	5,593	13,099	5,593	13,603
10	6,209	17,448	6,052	17,569	6,052	18,096
11	6,092	15,843	5,544	16,664	5,544	17,444
12	5,932	14,107	5,575	14,238	5,575	14,917
Sy	66	134	54	98	54	117
Un	24	179	28	208	28	178
Total	**1,10,444**	**2,35,063**	**1,04,242**	**2,33,519**	**1,04,242**	**2,40,414**

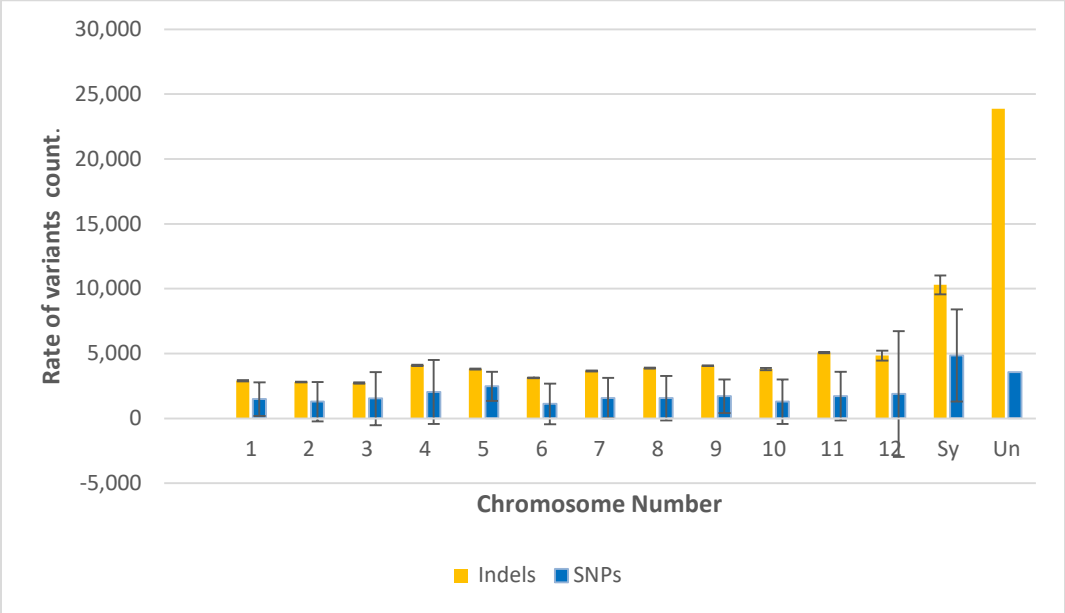

Fig 1. Variant rate of SNPs and InDels, Sy denotes synonymous and Un denotes unmapped chromosomes. (Error bars represent mean ± Standard Deviation, n = 3)

Number of effects caused by variants of RNA-sequencing data of Keteki Joha rice

The highest number of effects by impact was observed on 'Modifier' (363144.7) impact on SNPs of

CAHIERS MAGELLANES-NS
Volume 06 Issue 2
2024

ISSN:1624-1940
DOI 10.6084/m9.figshare.261421
http://magellanes.com/

Keteki Joha while the lowest effect was observed on 'high impact' (1931.67) of SNPs. On the other hand, highest number of effects by impact was observed on 'Modifier' (95287.67) impact on SNPs of Keteki Joha while the lowest effect was observed on 'high impact' (84578.67) at an average.

Fig 2. Number of effects by impact of SNPs and InDels (Error bars represent mean ± Standard Deviation, n = 3)

Number of effects by function and transition-transversion variants in SNP of RNA-sequencing data of Keteki Joha rice

The total ratio of 1.91 (Transition: Transversion ratio) at an average was observed on the Keteki Joha SNPs. Out of which 281421 were transitions and 147507 were transversion variants. The highest number of effects by function variants was as a result of 'Missense' mutation at an average of 39675.33 on SNPs followed by 'Silent' (36501.33) and at last by 'Nonsense' (610.67).

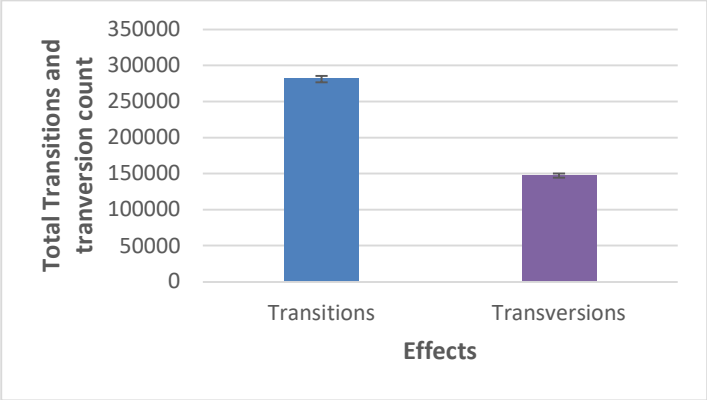

Fig 3. Total transitions and transversions in SNP variant (Error bars represent mean ± Standard Deviation, n = 3)

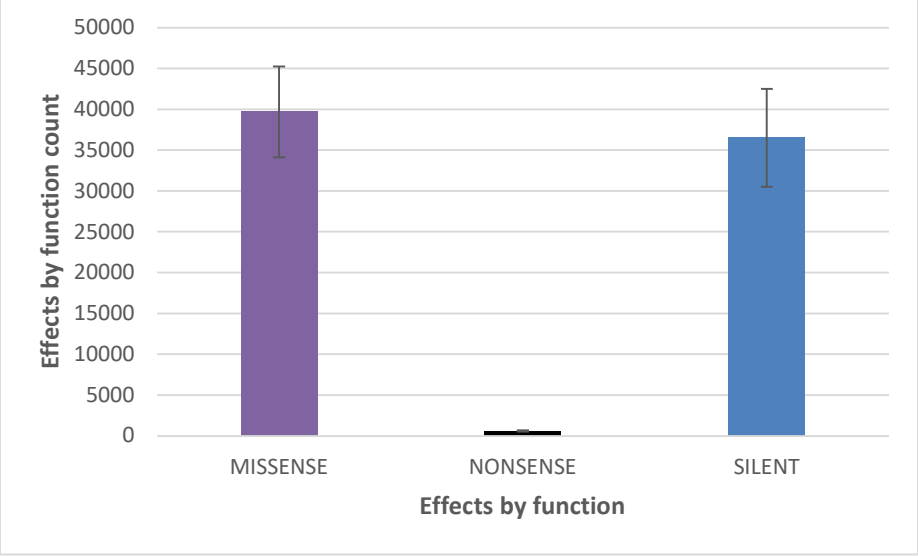

Fig 4. Number of effects by function in SNP Variants (Error bars represent mean ± Standard Deviation, n = 3)

CAHIERS MAGELLANES-NS
Volume 06 Issue 2
2024

ISSN:1624-1940
DOI 10.6084/m9.figshare.261421
http://magellanes.com/

Number effects by type of RNA-sequencing data of Keteki Joha rice

Different effects by type were distinctively found for both InDels as well as SNPs. The highest significant effect was shown by 'Introns' on both SNPs (141433.3) and InDel (121579.3) at an average followed by UTR 5' (55311.33) on SNPs and 'Splice-site region' (88165) in InDels. The least was observed on 'Transcript' with an average of 47.66 on SNPs and Intragenic region (0.33) in InDels.

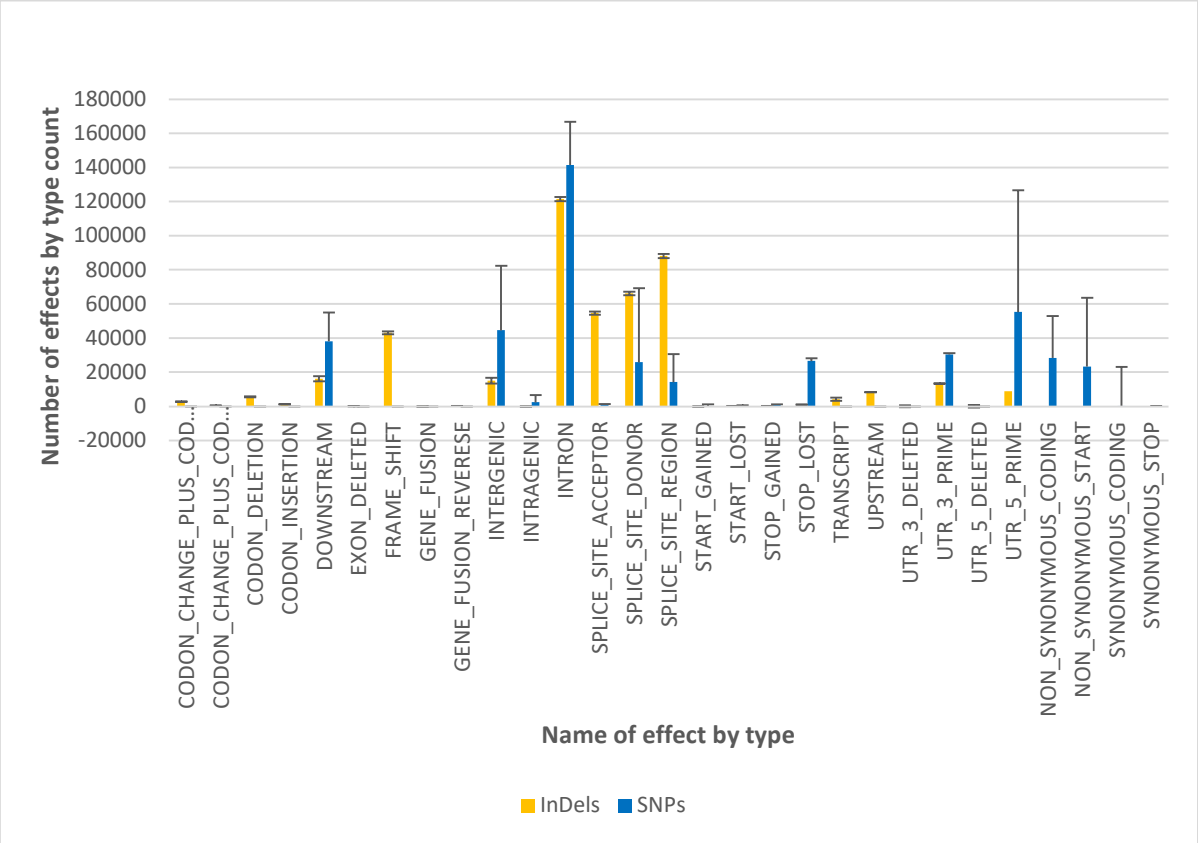

Fig 5. Number of effects by types in Keteki Joha InDels (Error bars represent mean ± Standard Deviation, n = 3)

Number effects by region of RNA-sequencing data of Keteki Joha rice

The highest amount of effects was observed on 'Intron' regions of snp variants with an average of 105402.7 followed by 'Downstream region' (88872.67) in SNPs. The least was observed on 'Splice-site acceptor' region with an average of 406.67. On the other hand, the highest amount of effects was observed on Exon region (47439) followed by Intron region (33094.66) and the least on 'Gene' at an average of 307.66 of InDels.

CAHIERS MAGELLANES-NS
Volume 06 Issue 2
2024

ISSN:1624-1940
DOI 10.6084/m9.figshare.261421
http://magellanes.com/

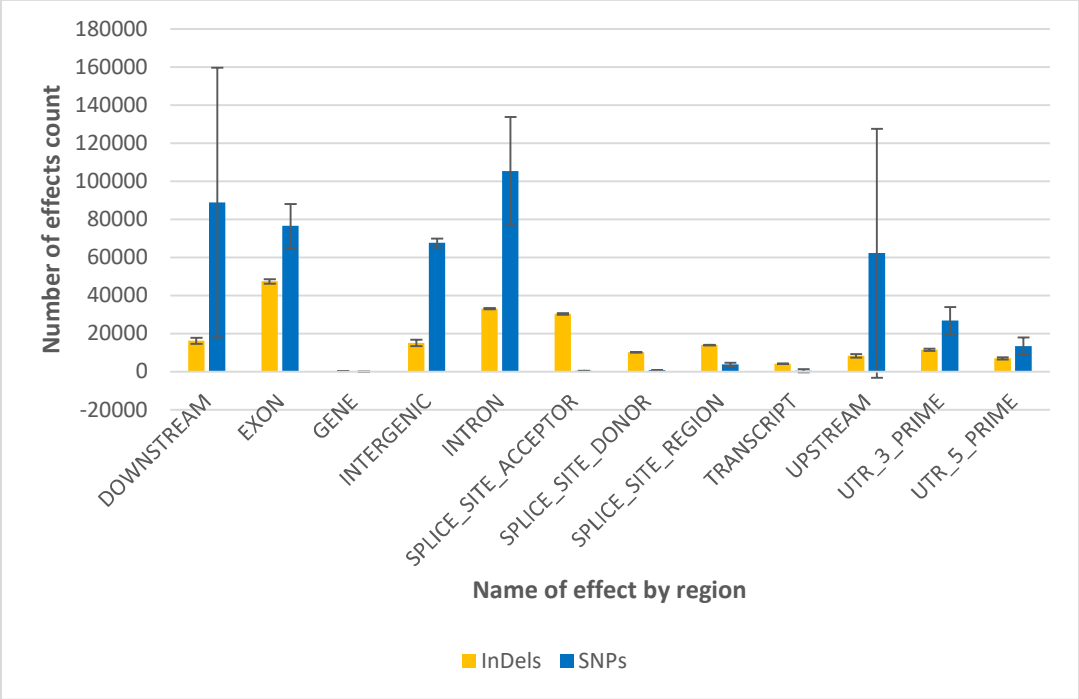

Fig 6. Number of effects by region (Error bars represent mean ± Standard Deviation, n = 3)

Identification of variants in randomly selected 10 genes of RNA-sequencing data of Keteki Joha rice

10 genes were randomly selected, and the variants occurred on it were calculated out as an average for three replications InDels and SNPs each, for Keteki Joha cultivar. The gene Mitogen activated protein (id. LOC_Os10g38950) showed maximum numbers of effects by 'intron' at an average of 6.66, followed by 'splice-site-region' at an average of 5.33 in InDels. Where the highest impact was shown by 'modifier' (1.33) of SNPs at an average for three replicates. The gene for WAX2 (id. LOC_Os02g40784) showed highest variants by both 'intron' and splice-site-region' at an average of 5.33 for InDels. Whereas highest impact on SNPs was impacted by 'intron' alone at 7.66 average. Similarly, LOC_Os01g59350 (transcription factor) showed the highest effect by 'intron' (2.33) followed by 'splice-site-acceptor' region and high impact variants (1.66) by InDels and 'variant impact modifier' showed the highest number of effects in SNPs with a n average of 6.33. OsPIP2 (2-3) gene for Aquaporin with id. LOC_Os02g41860 showed highest effect by 'intron' (3) in InDels and by modifier impact (7.33) on SNPs, id. LOC_Os01g47370, MYB family transcription factor showed the highest on 'modifier' (5.33) followed by 'UTR 5``'(5) on InDels and highest on 'modifier'(9) followed by 'upstream gene variant'(7) in SNPs, LOC_Os04g46400 (AP2 domain containing protein) showed 1 variants by both 'UTR 5`` and 'modifier' for InDel and 1 variant each for 'modifier' and 'upstream' in SNPs. LOC_Os03g57240 (ZOS3-19 - C2H2 zinc finger protein) showed highest variants by 'modifier' in SNPs with a count of 1.33, whereas no InDels variants were observed on it. Further LOC_Os03g18600 (cyclase/dehydrase family protein) showed highest variants by 'UTR 3`' (2.66) and 'modifier' (2.66) on InDels and by 'modifier' (4.33) in SNPs, LOC_Os02g14730 (ubiquitin carboxyl-

ISSN:1624-1940
DOI 10.6084/m9.figshare.261421
http://magellanes.com/

termi0l hydrolase family protein) showed the highest on 'intron' by 4.33 in InDels and 16.66 in SNPs. And finally, LOC_Os11g05290 (stress responsive A/B Barrel domain containing protein) showed the highest number of variants by 'modifier' (2) in InDels and a value of 5.66 in SNPs at an average.

Table 3. Average InDels and SNPs for 10 randomly selected genes in Keteki Joha rice across three replications

Serial no.	Gene Id	Name	Function
1.	LOC_Os10g38950	Mitogen activated protein	Signal transduction, regulation of gene expression
2.	LOC_Os02g40784	WAX2 (Protein GLOSSY 1-4)	WAXY coating formation
3.	LOC_Os01g59350	bZIP Transcription factor 8	Sequence-specific D0 binding (nucleus)
4.	LOC_Os02g41860	Aquaporin	Integral component of membrane, waster channel activity
5.	LOC_Os04g46400	AP2 domain containing protein	D0-binding transcription factor activity (nucleus)
6.	LOC_Os03g57240	ZOS3-19 - C2H2 zinc finger protein	Nucleic acid binding protein
7.	LOC_Os03g18600	Cyclase/dehydrase family protein	Abscisic acid-activated signaling pathway, protein phosphatase inhibitor activity
8.	LOC_Os02g14730	Ubiquitin carboxyl-termi0l hydrolase family protein	Metal ion binding, protease activity, root development, cell division
9.	LOC_Os11g05290	Stress responsive A/B Barrel domain containing protein	Stress responsive alpha-beta barrel domain
10.	LOC_Os01g47370	MYB family transcription factor	D0 binding

Discussion

The comprehensive RNA-sequencing data analysis for Keteki Joha rice yielded high-quality mapping statistics, highlighting the robustness of the sequencing process. The consistent average of 18 million input reads per replicate and the high percentage of uniquely mapped reads (84.22%) indicate reliable data acquisition. The average mapped read length of 296.5 bases aligns well with expected RNA-seq results, providing a solid basis for subsequent analyses. Notably, the high number of annotated splices (around 4.6 million) and the low mismatch rate per base (0.45%) underscore the precision of the mapping process. The low rates of deletions (0.05%) and insertions (0.04%) further emphasize the accuracy of the sequencing.

Compared to previous studies, the mapping quality and efficiency in this analysis are commendable. For instance, in a study on rice RNA-seq data by Wang et al., (2009), the unique mapping rate was reported to be around 75%, significantly lower than the 84.22% observed here. Similarly, another study by Lu et al., (2010) on *Oryza sativa* reported a higher mismatch rate of around 0.7%, compared to our 0.45%, indicating improved sequencing accuracy in our study. Additionally, the absence of chimeric reads in our data further highlights the enhanced quality of our RNA-sequencing process, compared to

ISSN:1624-1940
DOI 10.6084/m9.figshare.261421
http://magellanes.com/

the presence of a small percentage of chimeric reads (0.01%) reported in the study by Zhang et al., (2010).

The exploration of genetic variation in Keteki Joha rice using RNA sequencing data has provided significant insights into the genetic architecture and potential breeding improvements for this unique cultivar. The identification of a substantial number of single nucleotide polymorphisms (SNPs) and insertions/deletions (InDels) emphasizes the genetic diversity within this variety, aligning with previous studies on genetic variation in rice and other crops.

The total variants identified in Keteki Joha were 342,641.3 on average, with SNPs comprising 68% and InDels 32% of this total. This high level of genetic variation is consistent with the findings in other rice varieties and non-model organisms, where SNPs and InDels are abundant and can significantly impact gene function and adaptation (Kawahara et al., 2013; Sakai et al., 2014). The distribution of these variants across the 12 chromosomes, with chromosome 1 having the highest number of SNPs and InDels, indicates regions of potential interest for further functional and evolutionary studies. The functional annotation of variants revealed that most SNPs had a 'modifier' impact, while the least number of effects were 'high impact' variants. This distribution suggests that while many variants may have subtle effects on gene function, a smaller number of SNPs and InDels could lead to significant phenotypic changes. These findings are in line with studies on the genetic basis of traits in rice, where both high and low-impact variants contribute to phenotypic diversity and adaptation (Huang et al., 2010).

The transition-transversion ratio of 1.91 observed in Keteki Joha is comparable to ratios reported in other plants and animals, reflecting the evolutionary pressures shaping these genomes (Moragues et al., 2007). The high number of missense mutations among the SNPs indicates potential changes in protein function, which could influence traits such as aroma, yield, and stress resistance. These results underscore the importance of specific genetic changes in the adaptation and improvement of rice varieties (Yamamoto et al., 2010).

The analysis of variant effects by genomic region highlighted the prominence of intronic and upstream/downstream regions, which are critical for gene regulation and expression. The high number of intronic SNPs and InDels suggests possible regulatory roles, while the presence of variants in untranslated regions (UTRs) may impact mRNA stability and translation efficiency (Andorf et al., 2010). These patterns are consistent with the regulatory landscape observed in other rice studies, where non-coding regions play vital roles in trait variation (Zhang et al., 2016).

The examination of ten randomly selected genes revealed significant variant impacts, particularly in genes involved in stress responses, transcriptional regulation, and metabolic pathways. For example, the high number of intronic variants in the Mitogen-activated protein (MAP) gene and the splice-site variants in the WAX2 gene suggest functional modifications that could affect plant development and environmental interactions. These findings align with previous research on the genetic basis of stress tolerance and metabolic efficiency in rice (Liu et al., 2018).

Conclusion

This study provides a comprehensive overview of the genetic variation in Keteki Joha rice, highlighting the substantial SNP and InDel diversity and their potential impacts on gene function and regulation. The findings contribute to a broader understanding of rice genetics and offer valuable insights for breeding

ISSN:1624-1940
DOI 10.6084/m9.figshare.261421
http://magellanes.com/

programs aimed at improving yield, aroma, and resilience. Future research should focus on the functional validation of key variants and their roles in trait expression, as well as exploring the genetic diversity in other aromatic rice varieties to further enhance conservation and sustainable agricultural practices.

References

1. Andorf, C. M., Lawrence, C. J., Harper, L. C., Schaeffer, M. L., Campbell, D. A., and Sen, T. Z. (2010). The Locus Lookup tool at MaizeGDB: identification of genomic regions in maize by integrating sequence information with physical and genetic maps. *Bioinformatics*, 26(3), 434-436.
2. Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.
3. Bhangale, T. R., Rieder, M. J., Livingston, R. J., and Nickerson, D. A. (2005). Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes. *Human molecular genetics*, 14(1), 59-69.
4. Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120.
5. Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... and Ruden, D. M. (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*, 6(2), 80-92.
6. Das, A., Kesari, V., and Rangan, L. (2010). Aromatic joha rice of Assam-A review. *Agricultural Reviews*, 31(1), 1-10.
7. DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5), 491-498.
8. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15-21.
9. ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57.
10. Espinosa, E., Arroyo, M., Larrosa, R., Manchado, M., Claros, M. G., and Bautista, R. (2020, April). Micro-variations from RNA-seq experiments for non-model organisms. In *International Work-Conference on Bioinformatics and Biomedical Engineering* (pp. 542-549). Cham: Springer International Publishing.
11. Huang, Q. Y., Xu, F. H., Shen, H., Deng, H. Y., Liu, Y. J., Liu, Y. Z., ... and Deng, H. W. (2002). Mutation patterns at dinucleotide microsatellite loci in humans. *The American Journal of Human Genetics*, 70(3), 625-634.
12. Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q. I., Zhao, Y., ... and Han, B. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature genetics*, 42(11), 961-967.
13. Jeena, G. (2021). *A bioinformatics approach to quantify the effects of the underlying regulatory mechanisms on natural variation in gene expression by allele-specific expression analysis in Arabidopsis thaliana accessions using RNA-Seq Data* (Doctoral dissertation, Universität zu Köln).

ISSN:1624-1940
DOI 10.6084/m9.figshare.261421
http://magellanes.com/

14. Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., ... and Matsumoto, T. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*, 6, 1-10.
15. Kovach, M. J., Calingacion, M. N., Fitzgerald, M. A., and McCouch, S. R. (2009). The origin and evolution of fragrance in rice (*Oryza sativa* L.). *Proceedings of the National Academy of Sciences*, 106(34), 14444-14449.
16. Kunkel, T. A. (1985). The mutational specificity of DNA polymerase-beta during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations. *Journal of Biological Chemistry*, 260(9), 5787-5796.
17. Kunkel, T. A., and Bebenek, K. (2000). DNA replication fidelity. *Annual review of biochemistry*, 69(1), 497-529.
18. Lezmi, E., and Benvenisty, N. (2021). Identification of cancer-related mutations in human pluripotent stem cells using RNA-seq analysis. *Nature Protocols*, 16(9), 4522-4537.
19. Liu, C. T., Wang, W., Mao, B. G., and Chu, C. (2018). Cold stress tolerance in rice: physiological changes, molecular mechanism, and future prospects. *Yi chuan= Hereditas*, 40(3), 171-185.
20. Lu, T., Lu, G., Fan, D., Zhu, C., Li, W., Zhao, Q., ... and Han, B. (2010). Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome research*, 20(9), 1238-1249.
21. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9), 1297-1303.
22. Moragues, M., Moralejo, M., Sorrells, M. E., and Royo, C. (2007). Dispersal of durum wheat [Triticum turgidum L. ssp. turgidum convar. durum (Desf.) MacKey] landraces across the Mediterranean basin assessed by AFLPs and microsatellites. *Genetic resources and crop evolution*, 54, 1133-1144.
23. Pai, A. A., Cain, C. E., Mizrahi-Man, O., De Leon, S., Lewellen, N., Veyrieras, J. B., ... and Gilad, Y. (2012). The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels.
24. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology*, 29(1), 24-26.
25. Sakai, H., Tanaka, T., Antonio, B. A., Itoh, T., and Sasaki, T. (2014). The first monocot genome sequence: *Oryza sativa* (rice). In *Advances in botanical research* (Vol. 69, pp. 119-135). Academic Press.
26. Samal, R., Reddy, J. N., Rao, G. J. N., Roy, P. S., Subudhi, H. N., and Pani, D. R. (2014). Haplotype diversity for Sub1QTL associated with submergence tolerance in rice landraces of Sundarban region (West Bengal) of India.
27. Tyagi, A. K., Khurana, J. P., Khurana, P., Raghuvanshi, S., Gaur, A., Kapur, A., ... and Sharma, S. (2004). Structural and functional analysis of rice genome. *Journal of genetics*, 83, 79-99.
28. Wang, J., Dean, D. C., Hornicek, F. J., Shi, H., and Duan, Z. (2019). RNA sequencing (RNA-Seq) and its application in ovarian cancer. *Gynecologic oncology*, 152(1), 194-201.

ISSN:1624-1940
DOI 10.6084/m9.figshare.261421
http://magellanes.com/

29. Wang, W., Zhang, N., Chen, L., Zhao, X., Shan, Y., Yang, F., ... and Gu, S. (2024). Whole-genome sequencing and RNA sequencing analysis reveals novel risk genes and differential expression patterns in hepatoblastoma. *Gene*, 897, 147991.
30. Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1), 57-63.
31. 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, 1-14.
32. Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., ... and Wang, J. (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome research*, 20(5), 646-654.
33. Zhang, J., Chen, L. L., Xing, F., Kudrna, D. A., Yao, W., Copetti, D., ... and Zhang, Q. (2016). Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. *Proceedings of the National Academy of Sciences*, 113(35), E5163-E5171.

Author's Contributions

In this study, the author UB provided guidance and mentorship in designing the experiment and manuscript, incorporating feedback from all the co-authors. Authors KP and NH contributed equally to the experimental work. Authors NS, MRM and MK drafted the manuscript while the final manuscript was edited by authors UB and NS.

Conflict of Interest

The authors do not have any conflict of interests.