Can Tho University Journal of Science
website: ctujs.ctu.edu.vn

DOI: 10.22144/ctu.jen.2022.041

Genetic variation of Nang Thom Cho Dao rice variety based on whole genome sequencing

Huynh Ky[1*], Van Quoc Giang[1], Nguyen Loc Hien[1], Nguyen Chau Thanh Tung[1], Huynh Nhu Dien[1], Nguyen Nhut Thanh[2], Vo Cong Thanh[1], and Yeap Swee Keong[3]

[1]*College of Agriculture, Can Tho University, Viet Nam*

[2]*Department of Sciences and Technology of Long An, Long An City, Viet Nam*

[3]*China-ASEAN College of Marine Sciences Xiamen University Malaysia, Sepang, Selangor Malaysia*

**Correspondence: Huynh Ky (email: hky@ctu.edu.vn)*

Article info.

Received 12 Jan 2022
Revised 11 Feb 2022
Accepted 19 Feb 2022

Keywords

Bioinformatic, InDels, Nang Thom Cho Dao, SNPs, Variants

ABSTRACT

High-performance sequences are generating increasingly comprehensive catalogs of crop genetic variation. To make optimal use of this vast collection of data for research purposes, a robust and reproducible analytical pipeline discipline is required that is capable of accurately detecting and favoring variants. The entire genome sequencing data from the rice variety Nang Thom Cho Dao was analyzed using the appropriate bioinformatic pipeline. A total of 21 million reads with 6,6 GB of data were analyzed. SNPs and indels from the Nang Thom Cho Dao genome were found to be variable when compared to the Nipponbare reference rice genome. The result showed that the novel Indel of BADH2 gene in Nang Thom Cho Dao genome. The study will contribute valuable information to the development of genetic markers for rice breeding strategies using Nang Thom Cho Dao rice varieties.

1. INTRODUCTION

In the plant world, the defining phenotype is frequently used to identify organisms based on their appearance. However, the appearance of the same species or sub-species cannot be distinguished, despite the fact that their genetic material is distinct and results in distinctive characteristics. The unique characteristics contribute to the diversity of the plant kingdom and contribute to the large gene pool available for crop improvement selection and breeding schemes. To differentiate the distinct genetic materials, genome sequencing has been effectively applied to numerous plant species such as rice (Li et al., 2014), soybean (Shimomura et al., 2015), maize (Chandler & Brendel, 2002), cotton (Zhu et al., 2013), etc. useful for identifying the valuable traits of the crop, since the phenotypic observation could not be detected.

The advantage of whole genome sequencing in crops by NGS (next-generation sequencing) technology facilitate the creation of millions of new markers, especially for agronomically important genes (Thottathil et al., 2016). Rapid sequencing methods will almost certainly result in the faster identification of single nucleotide polymorphism (SNP) markers that make it easier to distinguish allelic variants of a given trait, making them more useful in crop breeding (Salgotra et al., 2014). In order to extract a few nucleotide variations for some traits from GB data from the genome, a bioinformatic facility will be required.

In this study, the complete genome of the Nang Thom Cho Dao (NTCD) variety was analysed using bioinformatics methods in order to identify a specific allele exhibiting the NTCD variety.

2. MATERIALS AND METHOD

2.1. Whole-genome sequencing of *Oryza sativa* L. cv. NTCD

NTCD seeds were obtained from Long An and kept at Can Tho University's gene bank. Undergreenhouse conditions, seedlings were grown in pots. Using the CTAB (cetyltrimethyl ammonium bromide) technique, genomic DNA was extracted from immature leaves (Doyle, 1991). The quantity and quality of genomic DNA were determined using a spectrophotometer and agarose gel electrophoresis. The full genome re-sequencing of the sample was performed on the Illumina HiSeq 2000™ by Novogen (Novogen, Malaysia). The genomic DNA was randomly sheared into short fragments of approximately 350 bp. The obtained fragments were used to construct a library using the NEBNext® DNA Library Prep Kit according to the manufacturer's instructions. To summarize, the required fragments (300–500 bp in size) were PCR enriched using P5 and indexed P7 oligos, followed by dA-tailing and further ligation with the NEBNext adapter. Following purification and quality control, the resulting library is ready for sequencing.

2.2. Genome mapping and variant calling in NTCD

In general, the application of computational genetic variation genetic pipeline is as in Figure 1. Using fastp tool V0.20.0, an ultrafast FASTQ preprocessor, quality checking and preprocessing of raw paired-end reads were conducted (Cock et al., 2010). The quality-filtered reads were mapped to the latest version of Os-Nipponbare-Reference-IRGSP-1.0 (Kawahara et al., 2013) available on Ensembl Plants website (Bolser et al., 2016) using HISAT2 software V2.1.0 (Keel & Snelling, 2018), SAMtools toolkit V1.9 totally eliminated low-mapping quality (MAPQ30 (Li et al., 2009). For variant detection, both SNPs and InDels were separately called via SAMtools toolkit V1.9 (Li et al., 2009) and VarScan. According to Li (2014), Filtering variations that overlap with low-complexity regions (LCRs) is the most successful method for identifying spurious heterozygotes.. The step was primarily masked features fall, as DUST in low-complexity regions by using minimap toolkit V0.2 (https://github.com/lh3/minimap) (Li, 2014).

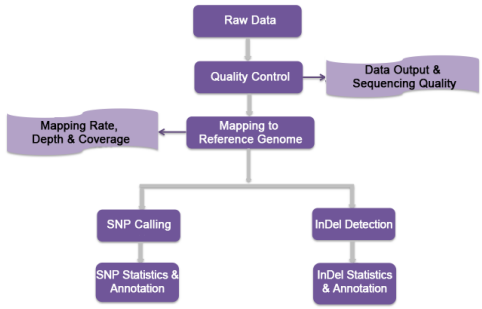

Figure 1. Bioinformatics analysis pipeline

2.3. Annotation of variants

In this analysis, SnpEff build V4.3+T.galaxy3 (Cingolani et al., 2012) on Galaxy UI (https://usegalaxy.org) was used to Build a reference database for functional annotation, using Os-Nipponbare-Reference-IRGSP-1.0 as the reference genome (ftp:/ftp.ensemblgenomes.org/pub/plants/release-45/fasta/oryza sativa/dna/) and a GFF3 as the annotation file (ftp:/ftp.ensemblgenomes.org/pub/plants/release45/gff3/ Eventually, the genomic distribution of SNPs and InDels was estimated and identified primarily using the awk and sort/uniq command lines.. The Awk program was used to obtain the VCF file's chromosomal locations, then to arrange those positions into windows of 10 kb and finally to sort/uniq the positions to obtain the count of variations in every window of 10 kb. The result was then presented with the Circa software.

3. RESULTS AND DISCUSSION

3.1. Whole-Genome Sequencing and Mapping

Raw reads were stored in a FASTQ file (Cock et al., 2010). More than 21 million clean readings corresponding to 6.6 Gb of sequencing data have been produced. For the present 373,245,194 bp reference genome of the Nipponbare genome, 95.81% of the NTCD sequence was mapped. The average depths inside the reference genome (without Ns) were 16.43X, and 1X coverages were 90.25%. This result is within the defined normal range and may be useful in detecting and analyzing later variations (Petrackova et al., 2019).

The effective sequencing data were aligned to the reference sequence using the HISAT2 program with default parameters, and mapping rate and coverage were calculated based on the alignment findings.

SAMTOOLS eliminated the duplicates (Li et al., 2009).

3.2. Annotation of SNPs in NTCD

Single nucleotide polymorphism (SNP) refers to a variation in a single nucleotide that may occur at some specific position in the genome, including the transition and the transversion of a single nucleotide. The results of SNPs annotation are summarized in Table 1 using SnpEff (Galaxy Version 4.3+T.galaxy1). Based on the analysis for SNPs variation, point mutation with transition (Ts) type was 1,422,100, and transversion (Tv) type was 580,530. The ratio Ts/Tv was 2,45, indicating that the quality of sequencing was sufficient (Wang et al., 2015). The variation of SNPs region showed highly in the intergenic region but less in UTR and exon (Figure 2).

Table 1. Annotation of SNPs in NTCD

Types	Count
3'-UTR_variant	24,075
5'UTR premature start_codon_gain_variant	2,354
5' prime_UTR_variant	14,847
downstream_gene_variant	399,523
initiator_codon_variant	5
intergenic_region	822,479
intron_variant	115,505
missense_variant	31,754
non_coding_transcript_exon_variant	5
non_coding_transcript_variant	5
splice_acceptor_variant	97
splice_donor_variant	94
splice_region_variant	3,643
start_lost	44
stop_gained	576
stop_lost	270
stop_retained_variant	73
synonymous_variant	26,718
upstream_gene_variant	413,965
Total	1,856,032

The Circos research analyzed the genomic organization of DNA polymorphism on all 12 NTCD chromosomes. For each NTCD genome chromosome, the number of DNA polymorphisms (SNPs) was proportional to the chromosome's length (Figure 2). The amount of high-impact SNPs on each NTCD chromosome, however, varied.

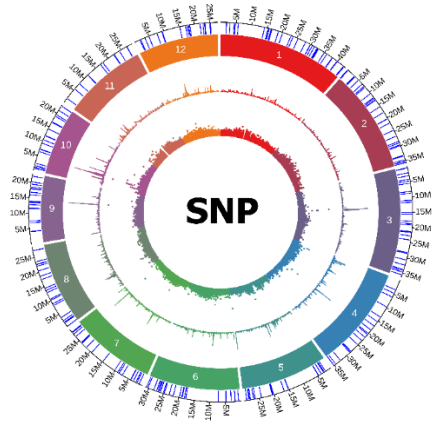

Figure 2. Distribution of SNPs in NTCD-T on each rice chromosome (25 M window size)

The outermost circles indicate 12 different coloured rice chromosomes. The middle indicated SNPs polymorphism and the innermost represents SNP distribution in NTCD. The blue bar indicated the high-impact SNPs.

3.3. Annotation of InDels in NTCD

Table 2. Annotation of InDels in NTCD

Types	Count
3_prime_UTR_variant	3,315
5_prime_UTR_variant	1,747
conservative_inframe_deletion/insertion	317
disruptive_inframe_deletion/insertion	402
downstream_gene_variant	42,884
frameshift_variant	973
intergenic_region	64,265
intron_variant	16,036
non_coding_transcript_variant	5
splice_acceptor_variant	12
splice_donor_variant	26
splice_region_variant	438
start_lost	19
stop_gained	13
stop_lost	18
upstream_gene_variant	42,576
Total	173,056

Insertion and deletion (INDEL) mutations are a significant source of genomic diversity. InDel refers to the insertion or deletion of ≤ 50 bp sequences in the DNA (Table 2). The genome-wide diversity of INDELs (with <50bp) was almost tenfold lower than that of SNPs. The number of InDels in the exonic region was 50-fold lower than the number of SNPs. The result was also observed in the Great Tit genomic variation (Barton & Zeng, 2019). In total,

173,056 InDels were detected in NTCD when compared with the reference genome. The NTCD genome contained a deletion of the *BADH2* gene, a previously undetected Indel (supplement table).

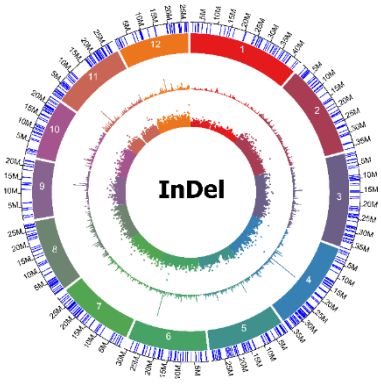

Figure 3. NTDC whole genome InDels variations

From outer to inner, rice chromosome, middle polymorphic, innermost InDel distribution. The blue bar indicated high impact InDels.

Genome-wide mapping of the circus diagram presented InDels density in 12 chromosomes of the NTCD genome (Figure 3). The number of InDels was distributed unevenly across the 12 chromosomes, with the highest (20 X 1000) on chromosome 3 and the fewest on chromosome 2. (12 X 1000).

4. CONCLUSIONS

In this study, bioinformatics as tools have been attempted to examine the variability in the NTCD genome when compared with the Nipponbarre reference genome. The novel Indel was detected in the NTCD genome. Our results will help develop valuable DNA markers for rice breeding programs, not only for Mekong delta' rice but also for the world's aromatic rice.

ACKNOWLEDGMENTS

This study was funded by Vietnamese Science foundation. The sequencing service was funded in part by the Can Tho University Improvement Project VN14-P6 supported by a Japanese ODA loan. The seeds were provided from CTU genebank.

REFERENCES

Barton, H. J., & Zeng, K. (2019). The impact of natural selection on short insertion and deletion variation in the great tit genome. *Genome Biology and Evolution*, *11*(6), 1514-1524. https://doi.org/10.1093/gbe/evz068

Bolser, D., Staines, D. M., Pritchard, E., & Kersey, P. (2016). Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. In D. Edwards (Ed.), *Plant Bioinformatics: Methods and Protocols* (pp. 115-140). Springer New York. https://doi.org/10.1007/978-1-4939-3167-5_6

Chandler, V. L., & Brendel, V. (2002). The maize genome sequencing project. *Plant Physiology*, *130*(4), 1594. https://doi.org/10.1104/pp.015594

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & 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. https://doi.org/10.4161/fly.19695

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, *38*(6), 1767-1771. https://doi.org/10.1093/nar/gkp1137

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., Schwartz, D. C., Tanaka, T., Wu, J., Zhou, S., Childs, K. L., Davidson, R. M., Lin, H., Quesada-Ocampo, L., Vaillancourt, B., Sakai, H., Lee, S. S., Kim, J., Numa, H., Itoh, T., Buell, C. R., & Matsumoto, T. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice (New York, N.Y.)*, *6*(1), 4-4. https://doi.org/10.1186/1939-8433-6-4

Keel, B. N., & Snelling, W. M. (2018). Comparison of Burrows-wheeler transform-based mapping algorithms used in high-throughput whole-genome sequencing: application to illumina data for livestock genomes. *Frontiers in genetics*, *9*, 35-35. https://doi.org/10.3389/fgene.2018.00035

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics (Oxford, England)*, *30*(20), 2843-2851. https://doi.org/10.1093/bioinformatics/btu356

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Genome Project Data Processing, S. (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J.-Y., Wang, J., & Zeigler, R. S. (2014). The 3,000 rice genomes project: new opportunities and challenges for future rice research. *GigaScience*, *3*(1). https://doi.org/10.1186/2047-217x-3-8

Petrackova, A., Vasinek, M., Sedlarikova, L., Dyskova, T., Schneiderova, P., Novosad, T., Papajik, T., & Kriegova, E. (2019). Standardization of sequencing coverage depth in ngs: recommendation for detection of clonal and subclonal mutations in cancer diagnostics. *Frontiers in oncology*, *9*, 851-851. https://doi.org/10.3389/fonc.2019.00851

Salgotra, R. K., Gupta, B. B., & Stewart, C. N. (2014). From genomics to functional markers in the era of next-generation sequencing. *Biotechnology Letters*, *36*(3), 417-426. https://doi.org/10.1007/s10529-013-1377-1

Shimomura, M., Kanamori, H., Komatsu, S., Namiki, N., Mukai, Y., Kurita, K., Kamatsuki, K., Ikawa, H., Yano, R., Ishimoto, M., Kaga, A., & Katayose, Y. (2015). The *Glycine max* cv. Enrei genome for improvement of japanese soybean cultivars. *International Journal of Genomics*, *2015*, 358127. https://doi.org/10.1155/2015/358127

Thottathil, G. P., Jayasekaran, K., & Othman, A. S. (2016). Sequencing crop genomes: a gateway to improve tropical agriculture. *Tropical life sciences research*, *27*(1), 93-114. https://pubmed.ncbi.nlm.nih.gov/27019684

Wang, J., Raskin, L., Samuels, D. C., Shyr, Y., & Guo, Y. (2015). Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics (Oxford, England)*, *31*(3), 318-323. https://doi.org/10.1093/bioinformatics/btu668

Zhu, Y.-N., Shi, D.-Q., Ruan, M.-B., Zhang, L.-L., Meng, Z.-H., Liu, J., & Yang, W.-C. (2013). Transcriptome analysis reveals crosstalk of responsive genes to multiple abiotic stresses in cotton (*Gossypium hirsutum* L.). *PloS one*, *8*(11), e80218-e80218. https://doi.org/10.1371/journal.pone.0080218