PLANT SCIENCE TODAY, 2021
Vol 8(3): 437–444
https://doi.org/10.14719/pst.2021.8.3.1047
ISSN 2348-1900 (online)

HORIZON
e-Publishing Group
HePG

RESEARCH ARTICLE

Whole-genome sequencing of three local rice varieties (*Oryza sativa* L.) in Vietnam

Ky Huynh*, Giang Van Quoc, Tung Nguyen Chau Thanh, Hien Nguyen Loc & Vo Cong Thanh
Department of Genetic and Plant Breeding, College of Agriculture, Can Tho University, Can Tho 94000, Vietnam
Email: hky@ctu.edu.vn

ARTICLE HISTORY
Received: 06 December 2020
Accepted: 10 April 2021
Available online: 01 May 2021

KEYWORDS
DNA markers
InDel
Next-generation sequencing
Local rice varieties
Vietnam

ABSTRACT

Recently, a new technology, Next-generation sequencing (NGS) has been launched and providing whole-genome sequences that helps identify molecular markers across the genome. DNA markers such as single nucleotides and insertion – deletion (InDel) polymorphisms were widely used for plant breeding particularly to distinguish important traits in rice. These PCR-based markers can be used for the precision detection of polymorphisms. Moreover, PCR-based approaches are simple and effective methods for dealing with the issue of fraudulent labeling and adulteration in the global rice industry. In this study, three local varieties of *Oryza sativa* L. in Vietnam were sequenced with up to ten times genome depth and at least four times coverage (~83%) using the Illumina HiSeq2000™ system, with an average of 6.5 GB clean data per sample, generated after filtering low-quality data. The data was approximately mapped up to 95% to the reference genome IRGSP 1.0. The results obtained from this study will contribute to a wide range of valuable information for further investigation into this germplasm.

Introduction

Rice is the primary food of more than half of the world's population. To date, thousands of rice varieties have been recorded, with different commercial names available in the market, while some unique varieties are particularly popular as they suit the tastes of consumers in a particular region (1). Attempts to create superior varieties/products via conventional breeding have constantly resulted in inconsistent or variations of product quality, which have contributed to the discrepancy of market prices that impact the income of rice farmers (2). Moreover, this variation may subsequently encourage some dishonest stakeholders and traders to mix or mislabel rice varieties to achieve excess profit (3). Traditionally, rice variety identification relies mainly on the expert opinions of breeders, extension services and other expert farmers and morphological descriptions. These methods, however, have several inherent levels of uncertainty. For example, in the absence of formal seed systems, the naming system can vary from time to time and from place to place, leading to inconsistencies in the name of the rice variety. In addition, environmental and plant development conditions influence morphological descriptors at different stages (4). These difficulties can be overcome by molecular markers, not only without environmental effects and stages of development but also with the available reference genomes of many species. In Vietnam, rice is an important food product and represents the first export commodity of the country. Commercially, Nang Thom Cho Dao Thom (NTCD-T) is the most expensive rice variety because of its finer flavor and superior quality. Consequently, local use is not enough, and the cost of other rice varieties is two to three times as high. Thus, developing a suitable method to authenticate this economically important trait of NTCD-T is required.

To solve this authenticity problem for NTCD-T, the molecular marker approach is appropriate given the same genetic background of these rice varieties in the market. DNA markers such as single-nucleotide polymorphism (SNP) and insertion/deletion polymorphism (InDel) markers are useful to distinguish among closely related varieties. Recently, whole rice genome data has become available and helpful for developing SNP and InDel markers using next-generation sequencing (NGS) technology (5). Compared to the need for unique facilities for SNP detection (6), the codominant technology for InDels is user-friendly and beneficial in some genetic analyses, especially in marker-assisted selection (MAS) (7). With the advancement of NGS technology and cost reduction, InDels were widely detected and created

via resequencing and became a precious resource for the research of many organisms, especially rice (8, 9). This study aimed to detect and create stable and practical InDel markers based on information resequencing from three closely related rice lines, NTCD-T (aroma), NTCD-TN (mild aroma), and Tai Nguyen, compared to a reference genome sequence, Nipponbare, so that the NTCD-T national rice cultivar can be authenticable and distinguished from other Vietnamese rice cultivars. Our findings thus set up the foundation for a long-term assessment of the purity of other rice varieties not only in Vietnam but also in authentic rice markets around the world.

Materials and Methods

Plant Material

NTCD-T belongs to *Oryza sativa* L. subsp. *indica*, is an aromatic local rice and is geographically restricted to the Cho Dao district in the Long An province of Vietnam (10°33'29"N, 106°36'23"E), and the Nang Thom Cho Dao mild aromatic line (NTCD-TN) and Tai Nguyen are local varieties sharing similar phenotypic characteristics with NTCD-T. These varieties have been difficult to distinguish in the market and therefore were chosen for comparison purposes regarding authentic NTCD-T.

DNA Isolation and Genome Sequencing

Genomic DNA was extracted from the leaf tissue of rice seedling (two weeks after germination) using the NEXprep™ Plant RNA Mini Kit (Genes Laboratories, Korea). The DNA was purified according to the manufacturer's protocol. DNA purity was checked on 1% agarose gel (w/v), and concentration and quality were determined using Nanodrop (Thermofisher, USA) and Bioanalyzer (Agilent, USA). The whole-genome resequencing of the three samples was performed on an Illumina HiSeq 2000™ from Novogen (Novogen, Malaysia). Library preparation and whole-genome sequencing were performed according to the standard Illumina protocol using Illumina's paired-end sequencing technology according to the Illumina pipeline. The 150-bp paired-end readings generated from the three genotypes were deposited in the NCBI sequence read archive (SRA) under study accession number PRJNA576771.

Genome Mapping and Variant Calling

The quality control and preprocessing of the raw paired-end readings were performed using fastp tool V0.20.0, an ultra-fast FASTQ preprocessor (10). The quality-filtered readings were mapped to the latest Os-Nipponbare-Reference-IRGSP-1.0 build (11), available on the Ensembl Plants website (12), using HISAT2 software V2.1.0 (13, 14), and low mapping quality (MAPQ < 30) was completely removed using SAMtools toolkit V1.9 (15). For variant identification, duplicates were removed from the alignment files using Picard tool V2.18.7 (http://broadinstitute.github.io/picard/). After that, we performed realignment on the InDels that had been identified in the realignment target step. Both the SNPs and the InDels were separately called via SAMtools toolkit V1.9 (15) and BCFtools V1.9 (16); the latest development version is available at GitHub (http://samtools.github.io/bcftools/bcftools.html#call). High-quality variants were retained for further data analyses by filtering out the low mapping quality of raw calls (%QUAL < 20) using BCFtools V1.9 (16). Filtering variants overlapping with low-complexity regions (LCRs) is the most effective method against false heterozygotes (17). This step primarily masked features fall, as DUST, in LCRs using minimap toolkit V0.2 (https://github.com/lh3/minimap) (18, 19).

Variant Analysis

We used SnpEff build V4.3+T.galaxy3 on the Galaxy UI (https://usegalaxy.org) to build the reference database for functional annotation, including the Os-Nipponbare-Reference-IRGSP-1.0 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/release-45/gff3/oryza_sativa). The genomic distribution of SNPs and InDels was eventually calculated and identified using mostly the awk and sort/uniq command lines, the former to determine the chromosome positions of the VCF file before grouping those positions by windows of 10 kb and the latter to determine the counts of the variants in each window of 10 kb. The data was then displayed using the Circa software (http://omgenomics.com/circa/). In addition, both the SNPs' and InDels' variants among the three samples were presented in Venn diagrams using VennyV2.1.0 (https://bioinfogp.cnb.csic.es/tools/venny/index.html).

PCR validation for InDel markers

A set of 2 InDel markers with a 3–10 bp major allele difference in genomes of three samples were randomly selected from the developed InDel markers for accuracy and polymorphism validation by the PCR technique. 100 out of 120 InDel markers were randomly selected to amplify the genomic DNA of three tested samples for accuracy validation. Then, we further detected their polymorphisms by PCR amplification of genomic DNA in a panel of 5 rice cultivars in which three tested sample and including 1 aroma commercial rice Jasmine 85 and local salt tolerant variety Doc Phung with 2 selected pairs of InDel primers (Table 1).

PCR was performed in a 15-µl reaction volume containing 50 ng of template DNA, 7.5 µl of 2X PCR master mix (NEXpro HS-Taq DNA polymerase, Korea), 10 nM of each primer and ddH2O. The DNA amplification protocol included an initial denaturation for 3 min at 95 °C, followed by 35 cycles of denaturation for 30 s at 95 °C, annealing for 90 s at 60 °C and an extension for 30 s at 72 °C, with a final extension for 10 min at 72 °C. The reactions were performed in a C1000 thermal cycler (Bio-rad, Inc., Hercules, CA). The PCR products were

Table 1. List of InDels markers used for detection of NTCD-T lines

Primer	Sequences (5'-3')
NTD-Del-08-5 F	ACAAGTAGGTTTGAAGGACTGC
NTD-Del-08-5 R	CTCCTACCCCATCTTAACTGCC
NTD-Del-09-4 F	GCCGACCGATGCAAAGTAAA
NTD-Del-09-4 R	TGGTTGGTTGTATGCAGTGC

subsequently detected in Qsep 100 fragment analyzer (Bioptic, Taiwan).

Results

Genomic Library Sequencing and Mapping

In this study, high-throughput Illumina sequencing technology was used to sequence commercial Vietnamese rice variety NTCD and two other accessions sharing the similar phenotypic characteristics. The genomic DNA was isolated from rice leaves in each variety to enable a downstream assessment. Illumina HiSeq 2000 was used for high-performance sequencing, and the resulting sequence readings were mapped with BWA to IRGSP-1.0. For the current 373,245,519 bp reference genome, the mapping rate of each sample ranged from 95.76% to 95.87%. Referring to the reference genome (without Ns), the coverage and the average read depths obtained across all twelve chromosomes were 85.23% and 16.43 times for NTCD-T, 83.46% and 12.89 times for Tai Nguyen, and 86.07% and 19.39 times for NTCD-TN respectively (Table 2). This result is in the qualified normal range and may serve in the subsequent variation detection and related analyses.

The SNPs and InDels were called based on unique alignments of the cleaned readings to Os-Nipponbare-Reference-IRGSP-1.0. The total numbers of SNPs (Supplement S1) for NTCD-T and NTCD-TN were 1,959,489 and 2,040,301 respectively, while that of SNPs for Tai Nguyen was 1,859,736. For each comparison, however, the numbers of InDels were considerably less than those of SNPs. The total numbers of InDels were 174,506 and 184,111 for NTCD-T and NTCD-TN respectively and 157,693 for Tai Nguyen. Furthermore, the NTCD-T and NTCD-TN SNPs and InDel densities were closer and higher than those of Tai Nguyen. The result showed that both NTCD-T and NTCD-TN shared a similar genetic background and therefore no significant variation was observed.

Table 2. Sequence information of the 3 genomes obtained by whole-genome re-sequencing analysis.

Sample	Mapped reads	Total reads	Mapping rate (%)	Average depth (X)	1X coverage (%)	4X coverage (%)
NTCD-T	41,996,952	43,833,336	95.81	16.43	90.25	85.23
Tai Nguyen	33,045,998	34,470,558	95.87	12.89	89.74	83.46
NTCD-TN	49,470,956	51,660,194	95.76	19.36	90.59	86.07

Nonrandom genomic organization of DNA Polymorphisms

In this study, all twelve chromosomes of the three genomes have been investigated for the genomic organization of the DNA polymorphisms. Considerable variations between the numbers of identified SNPs and InDels (Fig. 1A, 1B) were observed. The number of DNA polymorphisms (SNPs and InDels) at each chromosome for all three genomes was proportionate to the chromosome length (Fig. 2A, 2B). Overall, the numbers of SNPs and InDels in chromosome 3 were the most abundant, while those of chromosome 9 were the least abundant. In particular, with regard to the DNA polymorphisms in the three genomes, SNP variation was the highest in chromosome 1 and the lowest in chromosome 9 (Fig. 1A). Meanwhile, InDel variation was distributed most abundantly in chromosome 9 (Fig. 1B). The distributions of SNPs and InDels within the chromosomes in the three genomes were not uniform (Fig. 2A, 2B). More DNA polymorphisms were distributed in NTCD-TN compared with those in NTCD-T and Tai Nguyen. The highest-density (>870) SNP regions of 100 kb and the lowest-density (<2) SNP regions of 100 kb were detected in all three genomes. In contrast, the highest-density (>105) InDel regions of 100 kb and lowest-frequency InDel (<2) regions of 100 kb were also detected in all three genomes. However, the substantially differentiated spreads of DNA polymorphisms have also been recorded in many plants including rice.

Analysis of SNPs and InDels

Upon further investigation of the numbers of transitions (Ts) and transversions (Tv), transition-to-transversion (Ts/Tv) ratios as measures for possible random sequence errors were determined. In this study, the Ts/Tv ratios (Fig. 3A) (NTCD-T: 2.619; Tai Nguyen: 2.605; NTCD-TN: 2.628) were approximately higher than the statistical human Ts/Tv ratio (>2.1), indicating the high quality of the SNPs found in an oblique manner. The higher Ts/Tv ratio (termed as

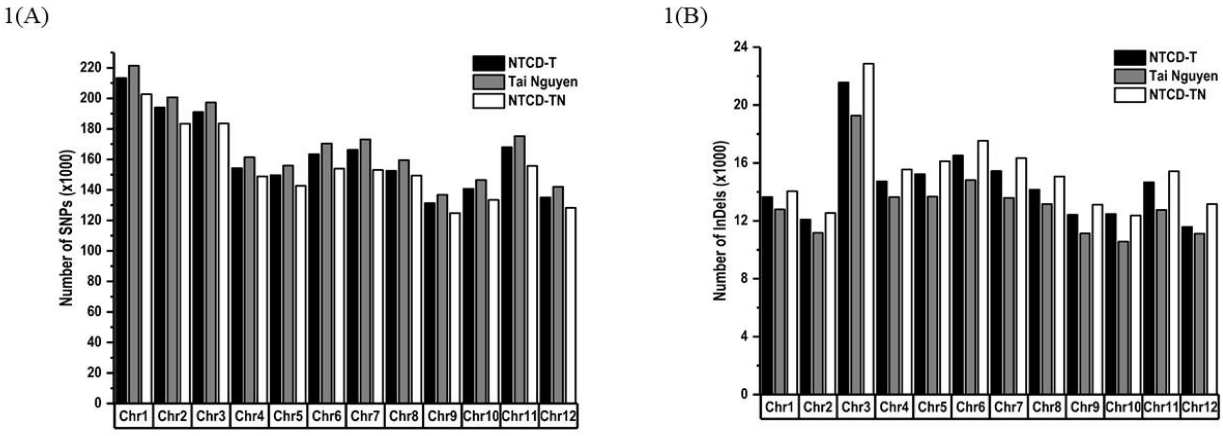

Fig. 1. Total numbers of SNPs (A) and InDels (B) in NTCD-T, Tai Nguyen and NTCD-TN detected on 12 chromosomes. NTCD-T, Tai Nguyen, NTCD-TN refer to SNPs/InDels identified in comparison with reference genomes Nipponbare.

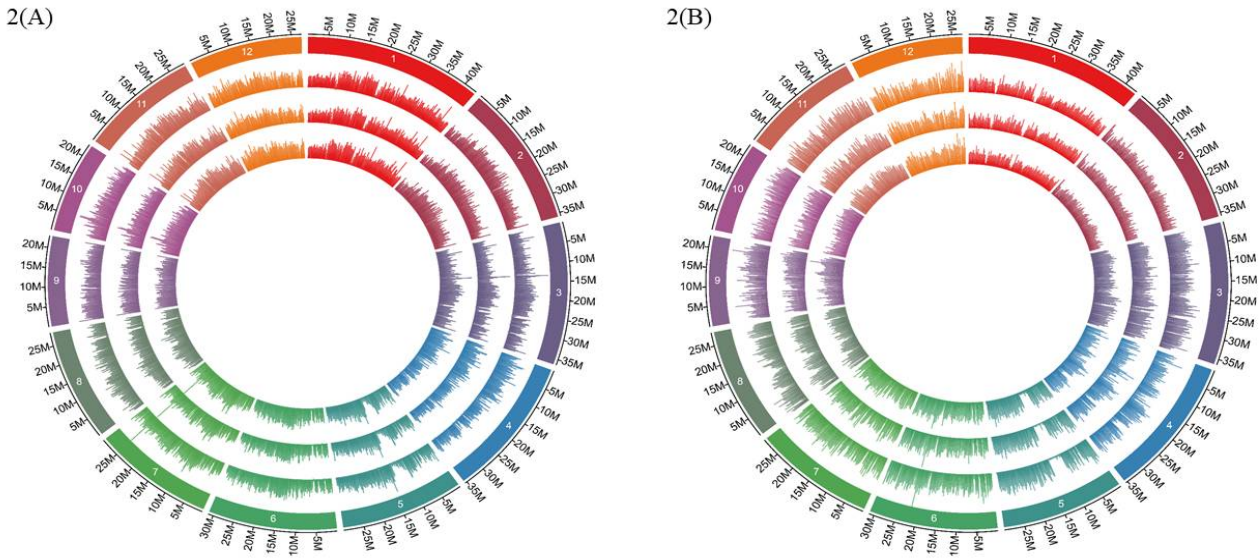

Fig. 2. Distribution of SNPs (A) and InDels (B) in NTCD-T, Tai Nguyen, NTCD-TN on each rice chromosomes (25 M window size). The outermost circles represent 12 rice chromosomes in different colors. The middle and the innermost represent SNP/InDel distribution in NTCD-T, Tai Nguyen and NTCD-TN, respectively. The blue triangle bar indicates the variation of SNP/InDel among these three genomes.

transition bias) had been reported in rice and maize. Thus, the higher frequency of Ts mutations over Tv mutations is due to mispairing and improved Ts tolerance because of fewer chances to change protein structure/functions in Ts compared with Tv. The total number of Ts (A/G and C/T) was significantly higher than that of Tv (A/C, A/T, C/G and G/T) for all three genomes. In all cases, the frequencies of A/G were at a similar level as those of C/T. The frequencies of Tv were not, however, at a similar level; the G/C frequency was lower than those of the three other Tv types (Fig. 3B). Our findings have been consistent with those of previous reports on rice and other plants.

Variant impacts were predicted in protein coding genes by SnpEff V4.3t. The variant impact statistics showed the highest numbers in the modifier group (Supplement S2) (NTCD-T: 7,363,697; Tai Nguyen: 7,021,002; and NTCD-TN: 7,607,886), while the variants predicted with high impact were 1,829; 1,758 and 1,869 in NTCD-T, Tai Nguyen and NTCD-TN respectively. The variants that affected noncoding genes were recognized, and when the information was available, the corresponding biotypes were identified. A biotype is a group of organisms that share the same genotype. Therefore, these variants with high impact will act as markers for identifying biotypes. Furthermore, missense SNPs in the three genomes ranged between 58,598 and 61,635 compared with those in the reference genome (Supplement S3). These SNPs have also been used to identify biotypes, resulting in their own characteristics.

Genomic annotation of DNA polymorphisms

The variants were annotated using Nipponbare as the RefSeq and Ensembl gen sets. In different genomic regions, we performed genome-wide annotation of the SNPs and InDels (Fig. 4A, 4B). In general, the patterns of the SNPs and the InDels were quite similar in different genomic regions for all the comparisons; while the numbers of variants in NTCD-TN and NTCD-T were similar; both variant numbers were higher than that of Tai Nguyen. Most of the detected SNP and InDel variants were detected most abundantly in nonfunctional genome regions such as upstream, downstream and intergenic (Fig. 4C, 4D).

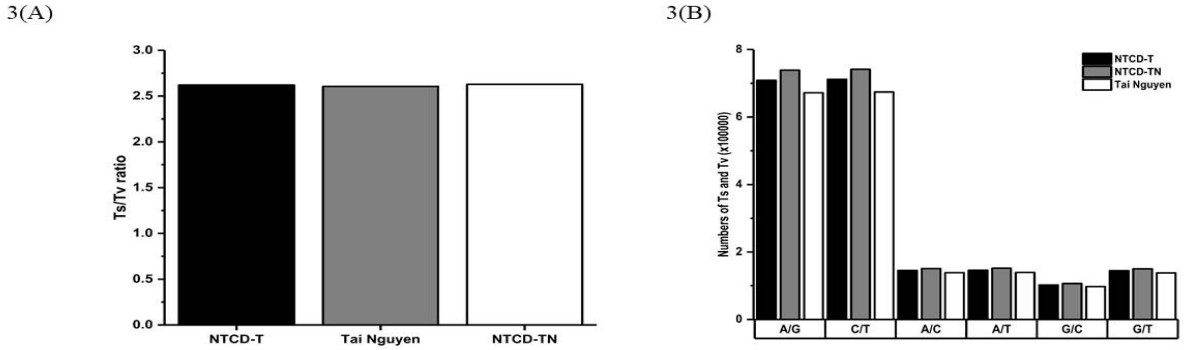

Fig. 3. Number of substitutions type (A) and Ts (transition)/Tv (transversion) ratio (B) in the identified SNPs in NTCD-T, Tai Nguyen and NTCD-TN.

The high frequency of genetic variants in the noncoding region could result from less pressure from natural selection. In the coding regions, however, the DNA polymorphisms were much lower than those in the noncoding regions, although these regions have been shown to play important roles in evolution. This can lead to phenotypical variation in these varieties. Since high-effect variants generate nonfunctional proteins that induce various phenotypic differences in evolution, high-effect SNPs and InDels among the three genomes were investigated in this study. These high-effect variants include the disruption of slicing sites, loss of translation in the start codon and introduction to the premature stop codon. Nonsynonymous SNPs and high-effect InDels were responsible only for 0.79%–0.81% and 0.24%–0.25% in the overall polymorphisms respectively (Supplement S4).

Variant comparisons among the three genomes

The following variants were detected—a total of 1,959,489 SNPs and 174,506 InDels from NTCD-T, 1,859,736 SNPs and 157,693 InDels from Tai Nguyen, and 2,040,301 SNPs and 184,111 InDels compared to the Nipponbare genome. Unique SNPs were recorded for 139,087 of the total NTCD-T SNPs, 337,387 of the total Tai Nguyen SNPs and 189,426 of the total NTCD-TN SNPs, representing 5.4%, 13.1%, and 7.4% of all the SNPs respectively. Similarly, specific InDels were detected for 17,013 of the total NTCD-T InDels, 33,459 of the total Tai Nguyen InDels and 22,878 of the total NTCD-TN InDels, constituting 7.1%, 13.9%, and 9.5% of the overall InDels respectively (Fig. 5A, 5B). The following DNA mutations in the genomes were widely recognized—52,981 SNPs and 5,795 InDels between NTCD-T and Tai Nguyen, 381,507 SNPs and 42,795 InDels between NTCD-T and NTCD-TN, and 83,454 SNPs and 9,543 InDels between Tai Nguyen and NTCD-TN. Approximately, the three genomes commonly shared about 5,091 SNPs and 287 InDels. In agreement with previous reports, genomic variations induced by natural means among rice lines and among cultivars were observed to have similar patterns of chromosome distribution and nucleotide substitution.

Distinguish NTCD-T Varieties from other Rice Varieties based on InDels Marker

To validate two selected InDel markers for distinguishing the NTCD-T rice cultivar from other rice varieties, we further analyzed them via PCR and fragment analysis. The unique size of the amplified fragments between NTCD-T and four other tested cultivars could be discriminated (Fig. 6A, 6B). The validation results show that the marker NTCD-8 could be detected in the NTCD-T cultivar with a deletion of 4–6 bp at the exon of chromosome 8. This suggests that the primer NTCD-8 can be used for an authentic assay of NTCD-T.

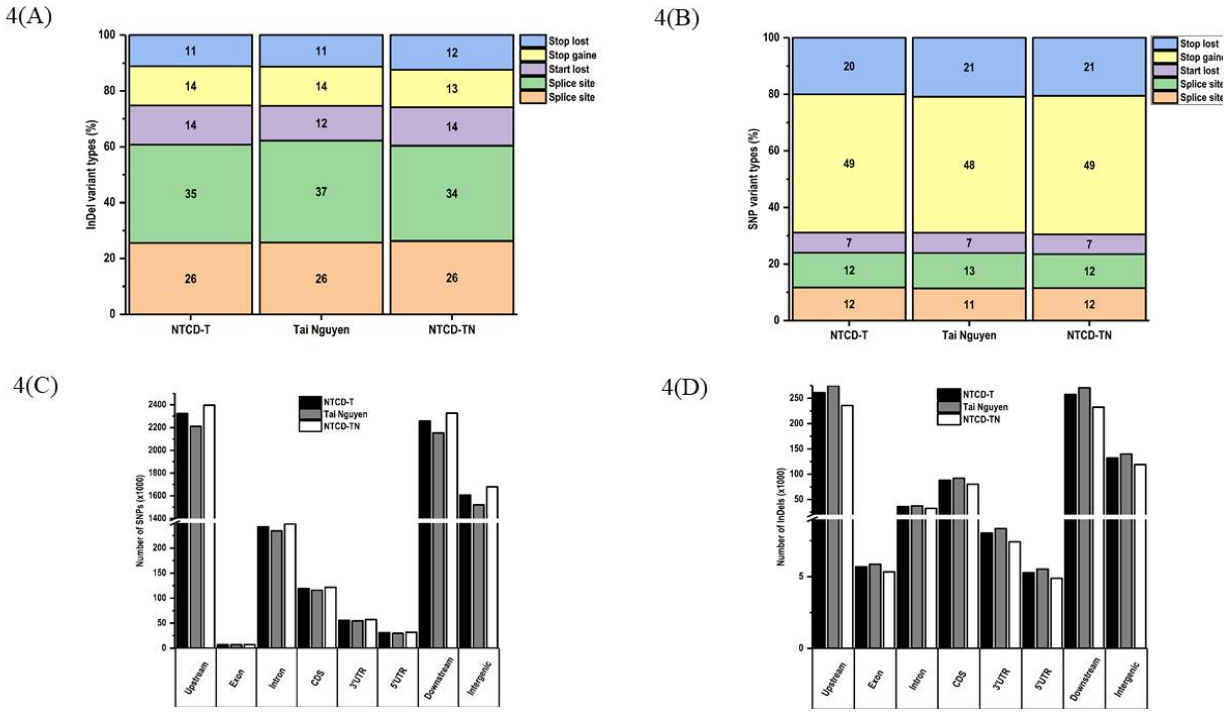

Fig. 4. Annotation of SNPs and InDels. (A) Distribution of SNP of high-effect SNPs in different genic regions; (B) Distribution of SNP of high-effect InDels in different genic regions; (C) distribution of SNPs in different genomic regions; (D) Distribution of InDels in different genomic regions.

Discussion

Adulteration is popular in most agricultural food products, where it is difficult to distinguish adulterants from foodstuffs using the naked eye/visual observation. Adulterated food products are higher in value among similar types. Among all adulteration detection techniques, DNA-based approaches are the most appropriate because they are easier, more affordable and more replicable. The development and cheaper sequence methods of

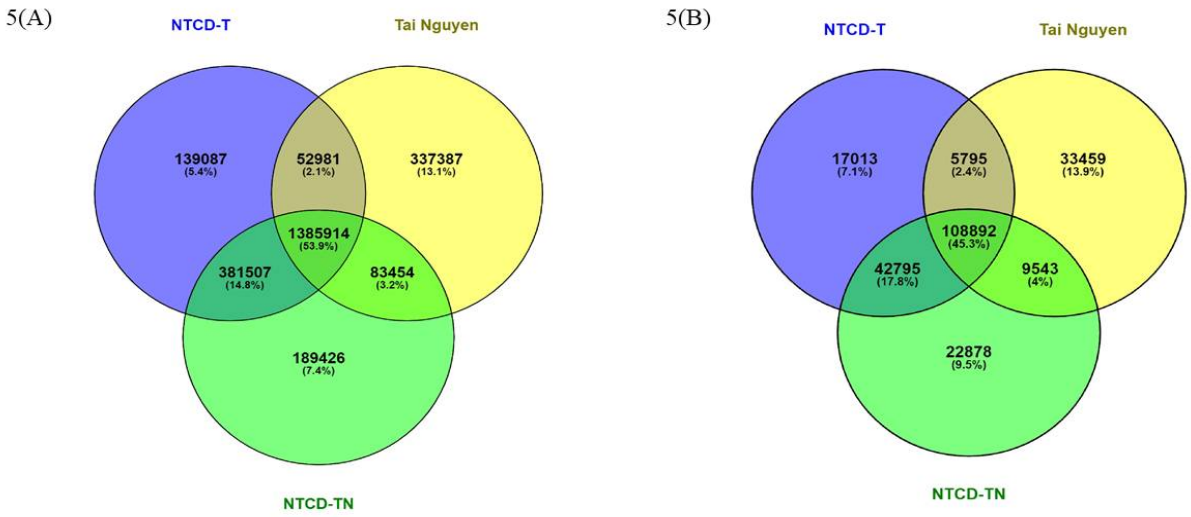

Fig. 5. Ven diagram showing the number of SNPs and InDels. (A) Venn diagram showing overlapping SNPs variants among NTCD-T, Tai Nguyen and NTCD-TN; (B) Venn diagram showing overlapping InDels variants among NTCD-T, Tai Nguyen and NTCD-TN. The overlapped parts indicate common detections in NTCD-T and Tai Nguyen, NTCD-T and NTCD-TN, Tai Nguyen and NTCD-TN, and NTCD-T, Tai Nguyen and NTCD-TN.

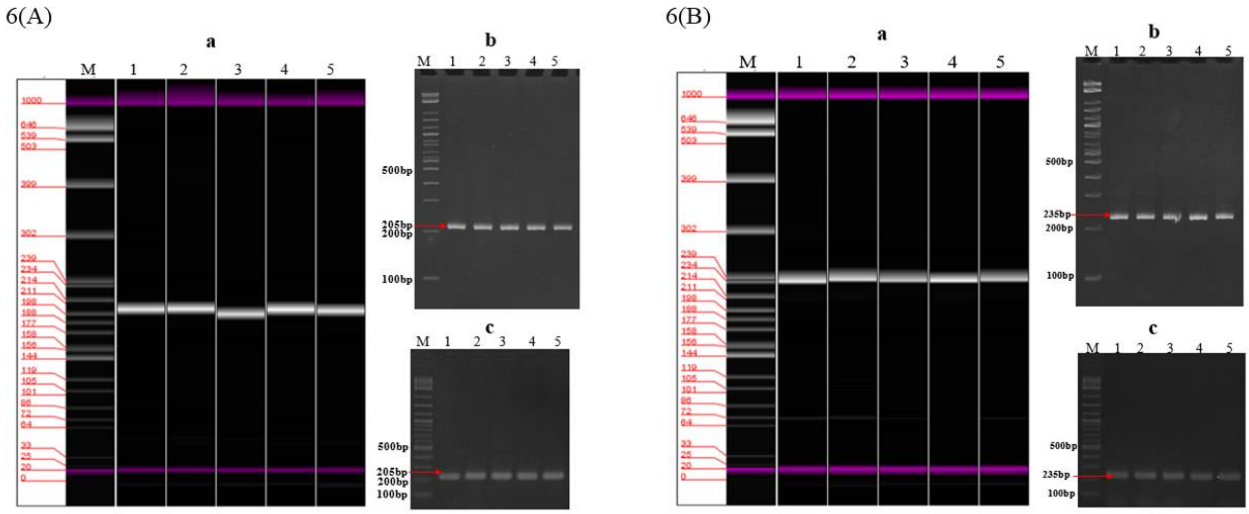

Fig. 6. The PCR profiles of NCD-Del-08-5 and NCD-Del-09-5 5 rice tested samples. (A) NCD-Del-08-5; (B) NCD-Del-09-5. a. fragment analysis using Qsep 100 (Bioptic, Taiwan); b. gel polyacrylaminde 8%; c. gel agarose 2%. M: marker, 1: NTCD-TN, 2: Tai Nguyen, 3: NTCD-T, 4: Jasmin 85, 5: Doc Phung

thousands of DNA markers give unprecedented applications of such DNA-based approaches to address the adulteration crisis in the foodstuffs and rice industry in particular (20, 21). A wide array of DNA-based markers are available for cultivar identification, including RAPD (22), restriction fragment length polymorphism (RFLP) (23), amplified fragment length polymorphisms (AFLP), SNPs (24), microsatellites (25) and InDels (26). Each DNA-based marker has its own advantages and disadvantages in their applications. Among all the DNA markers used to identify varieties in terms of reproductiveness, biological informativeness, and applicability, InDel markers are the most instructive. Taking advantage of a wide database of resequencing from NGS technology, InDel markers have been selected to develop an instructive DNA marker to detect rice lines that share genetic backgrounds. However, the high precision of InDel markers is calculated by the accuracy of the reference genome using an NGS strategy for InDel identification. Moreover, the accuracy of InDel polymorphism identification is affected when resequencing with a short read (27). Nevertheless, the accuracy of the polymorphism was performed based on the sequencing depth, with an average depth of sixteen times (Table 1), similar to that of the most current sequencing project, the 3,000 Rice Genomes Project (28). The present study shows that this profile of the sequencing depth is sufficient to accurately detect the polymorphisms in nucleotides. Many systems have been developed to study InDel markers in rice with 4–40 bp InDels (29), 30–100 bp InDels (30), or even larger InDels (31), i.e., up to 2,000 bp. These InDel PCR amplicons could be separated using polyacrylamide or agarose gels. Furthermore, these studies could only be performed via electrophoresis since the InDel polymorphisms differ by more than 4 bp. Here, fragment analysis

using bioptic technology with three sets of InDel markers was used to distinguish allelic variations of less than 4 bp (Fig. 6).

Conclusion

In this study, the InDels markers selected from rice genome resequencing were proven to be selective and sensitive method to distinguish rice line particularly for the aroma traits based on the simple PCR and polyacrylamide techniques. This result has shown that the potential of using these markers for rice variety identification particularly for the high value NTCD-T rice in Vietnam.

Acknowledgements

This study was funded in part by the Can Tho University Improvement Project VN14-P6 supported by a Japanese ODA loan. Special thanks to Department of Science and Technology, Long An Province for providing NTCD rice variety.

Authors' contributions

Conceptualization, writing—original draft and final version preparation, HK; methodology, formal analysis, VQG; NCTT; NLH; Funding acquisition, project administration, HK; VCT; All authors have read and agreed to the published version of the manuscript.

Conflict of interests

The authors declare no conflict of interest.

Supplementary files

Table S1. Total numbers of SNP and InDel in this study

Table S2. Number of SNPs and InDel effects by impact

Table S3. Number of SNPs effects by functional class

Table S4. Non-synonymous SNPs and high-effect InDels

References

1. Vemireddy LR, Satyavathi VV, Siddiq EA, Nagaraju J. Review of methods for the detection and quantification of adulteration of rice: Basmati as a case study. Journal of Food Science and Technology. 2015;52(6):3187-202. https://doi.org/10.1007/s13197-014-1579-0

2. Ghoshray A. Asymmetric Adjustment of Rice Export Prices: The Case of Thailand and Vietnam. Int J Appl Econ. 2008;5:80-91.

3. Colyer A, Macarthur R, Lloyd J, Hird H. Comparison of calibration methods for the quantification of Basmati and non-Basmati rice using microsatellite analysis. Food Additives and Contaminants: Part A. 2008;25(10):1189-94. https://doi.org/10.1080/02652030802040141

4. Duminil J, Hardy OJ, Petit RJ. Plant traits correlated with generation time directly affect inbreeding depression and mating system and indirectly genetic structure. BMC Evolutionary Biology. 2009;9(1):177. https://doi.org/10.1186/1471-2148-9-177

5. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one. 2011;6(5):e19379. https://doi.org/10.1371/journal.pone.0019379

6. Agarwal G, Jhanwar S, Priya P, Singh VK, Saxena MS, Parida SK *et al.* Comparative analysis of kabuli chickpea transcriptome with desi and wild chickpea provides a rich resource for development of functional markers. PloS one. 2012;7(12):52443. https://doi.org/10.1371/journal.pone.0052443

7. Liu B, Wang Y, Zhai W, Deng J, Wang H, Cui Y *et al.* Development of InDel markers for *Brassica rapa* based on whole-genome re-sequencing. TAG Theoretical and applied genetics Theoretische und angewandte Genetik. 2013;126(1):231-39. https://doi.org/10.1007/s00122-012-1976-6

8. Sahu PK, Mondal S, Sharma D, Vishwakarma G, Kumar V, Das BK. InDel marker based genetic differentiation and genetic diversity in traditional rice (*Oryza sativa* L.) landraces of Chhattisgarh, India. PloS one. 2017;12(11):e0188864-e. https://doi.org/10.1371/journal.pone.0188864

9. Moonsap P, Laksanavilat N, Sinumporn S, Tasanasuwan P, Kate-Ngam S, Jantasuriyarat C. Genetic diversity of Indo-China rice varieties using ISSR, SRAP and InDel markers. Journal of Genetics. 2019;98(3):80. https://doi.org/10.1007/s12041-019-1123-0

10. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics (Oxford, England). 2018;34(17):i884-i90. https://doi.org/10.1093/bioinformatics/bty560

11. Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S *et al.* Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice. 2013;6(1):4. https://doi.org/10.1186/1939-8433-6-4

12. Bolser D, Staines DM, Pritchard E, Kersey P. Ensembl plants: Integrating tools for visualizing, mining and analyzing plant genomics data. Methods in molecular biology (Clifton, NJ). 2016;1374:115-40. https://doi.org/10.1007/978-1-4939-3167-56

13. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature Methods. 2015;12(4):357-60. https://doi.org/10.1038/nmeth.3317

14. Keel BN, Snelling WM. Comparison of burrows-wheeler transform-based mapping algorithms used in high-throughput whole-genome sequencing: application to illumina data for livestock genomes1. Frontiers in Genetics. 2018;9(35). https://doi.org/10.3389/fgene.2018.00035

15. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al.* The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England). 2009;25(16):2078-99. https://doi.org/10.1093/bioinformatics/btp352

16. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics (Oxford, England). 2011;27(21):2987-93. https://doi.org/10.1093/bioinformatics/btr509

17. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics (Oxford, England). 2014;30(20):2843-51. https://doi.org/10.1093/bioinformatics/btu356

18. Li H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. Bioinformatics (Oxford, England). 2016;32(14):2103-10. https://doi.org/10.1093/bioinformatics/btw152

19. Chu J, Mohamadi H, Warren RL, Yang C, Birol I. Innovations and challenges in detecting long read overlaps: an evaluation of the state-of-the-art. Bioinformatics (Oxford, England). 2017;33(8):1261-70.

20. Kizhakayil D, Bhas S. Molecular marker based adulteration detection in traded food and agricultural commodities of plant origin with special reference to spices. Current Trends in Biotechnology and Pharmacy. 2010;4.

21. Voorhuijzen MM, van Dijk JP, Prins TW, Van Hoef AM, Seyfarth R, Kok EJ. Development of a multiplex DNA-based traceability tool for crop plant materials. Analytical and bioanalytical

chemistry. 2012;402(2):693-701. https://doi.org/10.1007/s00216-011-5534-x

22. Choudhury P, Kohli S, Srinivasan K, Mohapatra T, Sharma RP. Identification and classification of aromatic rices based on DNA fingerprinting. Euphytica. 2001;118:243-51. https://doi.org/10.1023/A:1017554600145

23. Zhang Q, Maroof MA, Lu TY, Shen BZ. Genetic diversity and differentiation of indica and japonica rice detected by RFLP analysis. TAG Theoretical and applied genetics Theoretische und angewandte Genetik. 1992;83(4):495-99. https://doi.org/10.1007/BF00226539

24. Shirasawa K, Shiokai S, Yamaguchi M, Kishitani S, Nishio T. Dot-blot-SNP analysis for practical plant breeding and cultivar identification in rice. TAG Theoretical and applied genetics Theoretische und angewandte Genetik. 2006;113:147-55. https://doi.org/10.1007/s00122-006-0281-7

25. Archak S, Lakshminarayanareddy V, Nagaraju J. High-throughput multiplex microsatellite marker assay for detection and quantification of adulteration in Basmati rice (*Oryza sativa*). Electrophoresis. 2007;28(14):2396-405. https://doi.org/10.1002/elps.200600646

26. Steele KA, Ogden R, McEwing R, Briggs H, Gorham J. InDel markers distinguish Basmatis from other fragrant rice varieties. Field Crops Research. 2008;105(1):81-87. https://doi.org/10.1016/j.fcr.2007.08.001

27. Yonemaru J-I, Choi SH, Sakai H, Ando T, Shomura A, Yano M *et al.* Genome-wide indel markers shared by diverse Asian rice cultivars compared to Japanese rice cultivar 'Koshihikari'. Breed Sci. 2015;65(3):249-56. https://doi.org/10.1270/jsbbs.65.249

28. The rgp. The 3,000 rice genomes project. GigaScience. 2014;3(1):7. https://doi.org/10.1186/2047-217X-3-7

29. Zeng YX, Wen ZH, Ma LY, Ji Z, Li XM, Yang CD. Development of 1047 insertion-deletion markers for rice genetic studies and breeding. Genetics and molecular research: GMR. 2013;12:5226-35. https://doi.org/10.4238/2013.October.30.7

30. Wu DH, Wu HP, Wang CS, Tseng HY, Hwu KK. Genome-wide InDel marker system for application in rice breeding and mapping studies. Euphytica. 2013;192. https://doi.org/10.1007/s10681-013-0925-z

31. Yamaki S, Ohyanagi H, Yamasaki M, Eiguchi M, Miyabayashi T, Kubo T *et al.* Development of INDEL markers to discriminate all genome types rapidly in the genus *Oryza*. Breed Sci. 2013;63(3):246-54. https://doi.org/10.1270/jsbbs.63.246

Additional information

Peer review information: *Plant Science Today* thanks Sectional Editor and the other anonymous reviewers for their contribution to the peer review of this work.

To cite this article: Huynh K, Quoc G V, Thanh T N C, Loc H N, Thanh V C. Whole-genome sequencing of three local rice varieties (*Oryza sativa* L.) in Vietnam. Plant Science Today. 2021;8(3):437–444. https://doi.org/10.14719/pst.2021.8.3.1047

Plant Science Today, published by *Horizon e-Publishing Group*, is covered by Scopus, Web of Science, BIOSIS Previews, Clarivate Analytics, etc. See https://horizonepublishing.com/journals/index.php/PST/indexing_abstracting