SNP and InDel Identification and Annotation from RNA-Sequencing Data

Uddipta Borthakur\textsuperscript{1,2}\textsuperscript{*}, Nibedita Sarma\textsuperscript{2}, Nickolsova Handique\textsuperscript{1}, Kanishka Purkait\textsuperscript{1}

\textsuperscript{1}Department of Botany, Handique Girls’ College, Guwahati, Assam, India
\textsuperscript{2}Department of Botany, Gauhati University, Guwahati, Assam, India

\textsuperscript{*}Corresponding Author: Uddipta Borthakur

Address: Department of Botany, Gauhati University, Guwahati, Assam, India

Phone No.: +91 8822673380
Email: borthakuruddipta@gmail.com

Abstract:

This study presents an in-silico pipeline for identifying single nucleotide polymorphisms (SNPs) and insertions or deletions (InDels) using RNA sequencing (RNA-seq) data. Genetic variations, such as SNPs and InDels, are vital for understanding genetic diversity and gene function. RNA-seq is an efficient and cost-effective method for analysing these variations, enabling detailed examination of gene expression profiles and detection of differentially expressed transcripts. The pipeline involves converting RNA samples into cDNA libraries, followed by fragmentation and adapter ligation. The RNA-seq data undergoes rigorous quality control, read alignment, and variant calling using advanced bioinformatics tools. This approach allows for precise identification of SNPs and InDels, providing critical insights into gene regulation, protein structure, and evolutionary adaptation. By detailing the workflow from RNA extraction to variant annotation, this study underscores the utility of RNA-seq in genetic variation research. The integration of high-throughput sequencing technologies and sophisticated computational methods facilitates the identification of genetic variants, with significant applications in personalized medicine, disease research, and crop improvement. This study highlights RNA-seq's potential to enhance our understanding of genetic diversity and its implications across various biological fields.

Key Words: cDNA libraries, Single nucleotide polymorphisms, Insertions and deletions, RNA-sequencing

Introduction

DNA sequence can undergo a persistent modification, which is termed as genetic variation. The term "gene variant" is now preferred over "gene mutation" because genetic changes don't always lead to disease, whereas "mutation" often carries a negative impact. Genomes frequently contain structural variations and presence or absence of polymorphisms (Voichek and Weigel, 2020) but are being generally ignored. Recently, the methods
of genetic variant detection is rapidly advancing, moving beyond the identification of single nucleotide changes to more complex variations, including insertions, deletions, repetitive sequences, and larger structural changes. (Tan et al., 2015).

Sequencing of whole-genome (WGS), exome sequencing (ES), and genotyping-by-sequencing (GBS) are well-established techniques that have greatly contributed to the analysis of genetic changes at the genome level (Bickhart et al., 2012; Davey et al., 2011; Elshire et al., 2011; Goodwin et al., 2016). These approaches are instrumental in identifying and characterizing various types of genetic variation, including single nucleotide polymorphisms (SNPs), single nucleotide variants (SNVs) and insertions or deletions (InDels) (Alkan et al., 2011; DePristo et al., 2011; Van der Auwera et al., 2013). These genetic variations leads to genetic diversity within and between populations.

Additionally, RNA-sequencing (RNA-seq) offers a cost-efficient approach to genetic variation studies, providing a powerful alternative to traditional methods (Wang et al., 2009; Trapnell et al., 2010). RNA-seq enables the examination of gene expression profiles and identify the transcripts which are differentially expressed across various conditions and tissues. Detecting SNP in a single nucleotide can provide essential information associated with a particular phenotype (Pickrell et al., 2010; Montgomery et al., 2010). Besides, SNP can be linked to a particular stress response leading to a more specific understanding of stress responses (Li et al., 2011).

Genetic variation has a key function in shaping the diversity of living organisms. single nucleotide polymorphisms (SNPs) and insertions or deletions (InDels) are mostly found genetic variation in genome. The emergence of high-throughput sequencing technologies like RNA sequencing (RNA-seq) has made it easier to identify and characterize genetic variation at a genome-wide scale.

RNA-seq is a technique that allows researchers to capture a snapshot of the transcriptome, the complete package of RNA molecules produced by the genome by transcription, in a particular cell type or tissue at a time (Wang et al., 2009; Mortazavi et al., 2008). 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 (Pan et al., 2008; Trapnell et al., 2012).

Types of Genetic variants:

Genetic variants are naturally occurring discrepancies in DNA sequence found among individuals within a specific population. These distinctions can emerge in both protein-coding and non-coding regions of the genomic sequences and have the potential to impact an array of traits and features, including susceptibility to diseases, efficacy of drug metabolism, and physical attributes. The different variations include structural variants, single nucleotide polymorphism or single nucleotide variation, insertion and deletion, copy number variants, translocation and transversion variants (Ku et al., 2010). These subtle alterations, involving the substitution of a
single nucleotide base pair, are termed single-nucleotide polymorphisms (SNPs) when observed in population-level genetic variation, and single-nucleotide variations (SNVs) when identified in individual genomes. An average individual has millions of SNPs, and plants may have many more (Kumar S, et al., 2012). InDels, which stand for "insertion" and "deletion," are base pair additions or subtractions made to a DNA segment (Mullaney JM, et al., 2010). InDels are more significant than SNPs/SNVs since they involve one to ten thousand base pairs. Copy number variants (CNVs) denote variances in the quantity of genes for a given trait within a genome. CNVs are notably widespread, often encompassing three times the number of base pairs compared to SNP/SNVs, making them the most prevalent form of structural variation.

Significance of SNPs and InDels:

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, whereas InDels can cause frameshifts 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, model organisms, and non-model organisms. 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 (Salk et al., 2018; Soemedi et al., 2017). SNPs and InDels are crucial in various biological processes, including gene regulation, protein structure, and evolutionary adaptation. They play pivotal roles in shaping gene expression levels, protein function, and interactions with other molecules. SNPs occurring within coding regions can lead to amino acid substitutions, potentially affecting protein stability, enzymatic activity, or protein-protein interactions (Sauna and Kimchi-Sarfaty, 2011). Non-synonymous SNPs, in particular, have the capacity to introduce alterations in the functional domains of proteins, potentially modifying their activity or specificity. On the other hand, although synonymous SNPs do not directly alter the amino acid sequence, they can influence protein folding, translation efficiency, or RNA stability. In the realm of InDels, they have the potential to cause substantial disruptions in gene function. Specifically, insertions or deletions within coding regions can give rise to frameshift mutations, resulting in premature stop codons and truncated proteins. This has a profound impact on protein function. Moreover, in non-coding regions, InDels possess the capability to modify regulatory elements, thereby affecting gene expression patterns (O’Roak et al., 2011). From an evolutionary perspective, SNPs and InDels contribute significantly to genetic diversity within populations. They are instrumental in genetic adaptation and speciation by introducing variations that confer selective advantages or disadvantages under different environmental conditions (Nachman et al., 2004).

Nowadays, Single Nucleotide Polymorphisms (SNPs) are the preferred since they are present in almost all groups of individuals in substantial numbers. Human forensics (Brenner and Weir, 2003) and medicine (McCarthy et al.,
SNPs have been used in various fields, including aquaculture (Liu and Cordes, 2004), marker-assisted dairy cattle breeding (Schaeffer, 2006), crop improvement (Yu et al., 2011), conservation (Seddon et al., 2005) and management of resources in fisheries (Smith et al., 2005).

SNPs are useful in interpreting breeding pedigrees, determining species genomic divergence to clarify speciation and evolution, and connecting genetic variants to phenotypic features (McNally et al., 2009). SNPs have been used to measure genetic variation, identify individuals, ascertain population structure, and ascertain parentage relatedness (Morin et al., 2004). Through a Genome Wide Association Studies (GWAS) designed to uncover the rice's evolutionary path leading up to its domestication, seed shattering (or lack thereof) has been linked to an SNP (S. Konishi et al., 2006).

Studies also say cells have numerous defences against the deadly effects of cancer-causing genetic mutations, in contrast to some other diseases that can be brought on by alterations in a single gene (Vogelstein and Kinzler, 2004). As a result, a fraction of faulty genes lead to cancer (Yeang et al., 2008) and several DNA alterations in cancer genes can influence the ultimate stage of carcinogenesis. Furthermore, it is believed that somatic mutation accumulation in tumour suppressor and oncogenes is crucial for the development of cancer and causes normal cells to transform into malignant ones throughout a few stages (Nowell, 1976).

Also, since chloroplast DNA (CpDNA) is inherited from mothers and has a stable structure, it is one of the essential parts of plant total DNA. Identification of various species will be aided by research on genetic differences found in the chloroplast genome, such as InDels and SNPs. Initially, population structure analysis, genetic diversity, and classification in the Oryza sativa L. genome were accomplished using SNP (Glaszmann, 1987, Singh et al., 2013). The InDel species-specific markers of chloroplasts were created to differentiate between 22 species of the genus Oryza sativa L. (Misra, 2019). Singh et al. (2018) employed SNP array for population structure study in wild rice accessions. The use of SNP variants in phylogenetic analysis, association studies, background selection, QTL mapping, assessment of genetic diversity, and background selection has been supported by research of Singh et al., 2015. SNPs are now the most often used marker for genetic investigations in plant species like rice (Subbaiyan et al., 2012) and Arabidopsis (Horton et al, 2012).

**Workflow of RNA Sequencing**

RNA-seq, a high-throughput sequencing technology, enables the simultaneous quantification and characterization of RNA molecules in a sample. Transcriptome analysis allows for the identification of actively transcribed genes, as well as their relative expression levels, thereby providing knowledge about the dynamic interplay between genomic information and cellular function (Mackenzie, 2018). RNA sequencing (RNA-seq) offers a valuable alternative for identifying genomic variants, as it provides information not only on gene expression but also on alternative splicing events, RNA editing, and other transcriptomic features. Recent developments in accurate mapping of RNA-seq reads and computational methods to identify SNPs in cancer (Goya et al., 2010, Chen et al., 2012)) have been able to identify disease-associated variations in RNA-seq data (Shah et al., 2012).
The initial step in the procedure involves converting the RNA sample into a cDNA library by the fragmentation of the RNA into complementary DNA pieces. This is achieved through reverse transcription, allowing the RNA to be utilized in a next-generation sequencing (NGS) process. Subsequently, the cDNA is fragmented, and adapters are attached to both ends of the resulting fragments. These adapters contain functional components necessary for sequencing, such as the primary sequencing priming site and an amplification element, which facilitate the clonal amplification of the fragments. Short sequences that partially or entirely match the segment from which the cDNA library can be created, are produced by NGS analysis of the library after it has undergone amplification, size selection, clean-up, and quality-checking procedures. Sequencing can be performed using either single-end or paired-end techniques. Single-read sequencing, which sequences cDNA fragments from one end, is faster and significantly more cost-effective (approximately 1% of the cost of Sanger sequencing). Strand-specific methods offer the advantage of providing additional information, resulting in millions of reads by the end of the workflow.

Identification and Annotation of SNP and InDels from RNA-Seq Data

Pre-processing of Raw Sequencing Data:

The initial phase of raw sequencing data analysis is fundamental for ensuring data integrity and eliminating artifacts. This encompasses critical steps such as adapter trimming, culling low-quality reads, and purging contaminants. Several software tools are available for these pre-processing procedures, affording researchers flexibility in their approach.

Trimmomatic (Bolger et al., 2014) and Cutadapt (Martin, 2011) are widely adopted tools for adapter trimming and enhancing sequence quality. Fastp is a powerful tool that seamlessly integrates adapter trimming, read filtering, and quality control (Chen et al., 2018). PRINSEQ provides extensive options for sequence pre-processing, including quality trimming, filtering, and statistical assessments (Schmieder and Edwards, 2011). Additionally, BBMap offers a versatile suite of tools for pre-processing, encompassing tasks such as adapter trimming, filtering, and even error correction (Bushnell et al., 2017).

Quality Control and Read Alignment:

Quality control is an indispensable step in evaluating the fidelity of sequencing data. FastQC (Andrews, 2010) and tools like MultiQC offer a convenient means to aggregate quality control metrics from multiple samples into a comprehensive report (Ewels et al., 2016).

Following quality control, the reads undergo alignment or mapping to a designated reference genome or transcriptome. Esteemed alignment algorithms such as Bowtie2 (Langmead and Salzberg, 2012), BWA (Li and Durbin, 2009), and HISAT2 (Kim et al., 2019) are standard choices for this endeavor.
Identification of Variants:

For SNP and InDel calling, in addition to the well-established tools like GATK, FreeBayes, and SAMtools, there are other resources available. VarScan2 (Koboldt et al., 2012) is a versatile tool for identifying somatic mutations and germline variants. Platypus (Rimmer et al., 2014) provides a robust framework for detecting variants in high-throughput sequencing data, encompassing SNPs, InDels, and structural variants.

Functional Annotation of Genetic Variants:

Functional annotation tools like ANNOVAR, SnpEff, and VEP offer comprehensive annotations. Additionally, tools like Variant Annotation in the R programming environment provide flexible solutions for variant annotation within a scripting environment (Obenchain et al., 2014).

Post-Processing and Downstream Analyses:

Once variants are identified and annotated, further analyses can be conducted to gain insights into their functional implications. These analyses may include pathway enrichment, functional enrichment, and network analysis. Tools like DAVID (Huang et al., 2009), Enrichr (Chen et al., 2013), and STRING (Szklarczyk et al., 2019) are
commonly employed for such analyses. In addition to functional analyses, visualization of the variant data can provide valuable insights. Tools like Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al., 2013) and GenomeBrowse (Golden Helix, Inc.) allow for interactive exploration of the genomic data, enabling researchers to visually inspect the variants in their genomic context.

Validation of identified variants is a crucial step to ensure their accuracy and reliability. Experimental validation methods, such as Sanger sequencing, polymerase chain reaction (PCR), or targeted sequencing, can be employed to validate specific variants of interest.

Furthermore, validation against independent datasets or comparison with previously published studies can provide additional confidence in the identified variants.

**Conclusion**

In genetic variation studies, the detailed analysis of gene expression profiles and the identification of differentially expressed transcripts are made possible by the effective technique of RNA sequencing (RNA-seq). It performs a fundamental part in identifying SNPs and InDels, offering essential insights into specific phenotypes and stress responses. Additionally, RNA-seq significantly contributes to advancing agriculture by elucidating how genes influence plant phenotypes.

Single nucleotide polymorphisms (SNPs) and InDels are of particular interest due to their functional and evolutionary implications. SNPs can lead to alterations in protein sequences, stability, or RNA processing, while InDels can cause frameshift mutations, affecting protein function. These variations are essential in gene regulation, protein structure, and evolutionary adaptation, shaping genetic diversity within populations and undergoing natural selection.

Recent studies utilizing RNA-seq have been instrumental in identifying and characterizing SNPs and InDels across diverse species, providing valuable insights into their functional consequences, including their involvement in diseases and adaptive processes. This integration of high-throughput sequencing technologies and RNA-seq holds far-reaching implications in fields ranging from personalized medicine to agriculture, driving progress in disease research and crop improvement. The ongoing refinement of these techniques promises even deeper insights into the genetic diversity of living organisms.

**References**


61. Tripathy, K. Development of Indian wild Rice Database and DNA Markers for the Identification of wild Oryza Species.

62. Tripathy, K. Development of Indian wild Rice Database and DNA Markers for the Identification of wild Oryza Species.


