https://doi.org/10.48047/AFJBS.6.7.2024.3233-3244

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

SNP and InDel Identification and Annotation from RNA-Sequencing Data

Uddipta Borthakur[1,2]*, Nibedita Sarma[2], Nickolsova Handique[1], Kanishka Purkait[1]

[1]Department of Botany, Handique Girls' College, Guwahati, Assam, India

[2] Department of Botany, Gauhati University, Guwahati, Assam, India

***Corresponding Author:** Uddipta Borthakur

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

Phone No.: +91 8822673380

Email :borthakuruddipta@gmail.com)

Volume 6 issue 7 2024

Received:15May2024

Accepted:10June2024

doi:10.48047/AFJBS.6.7. 2024.3233-3244

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.*,

2008). 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).

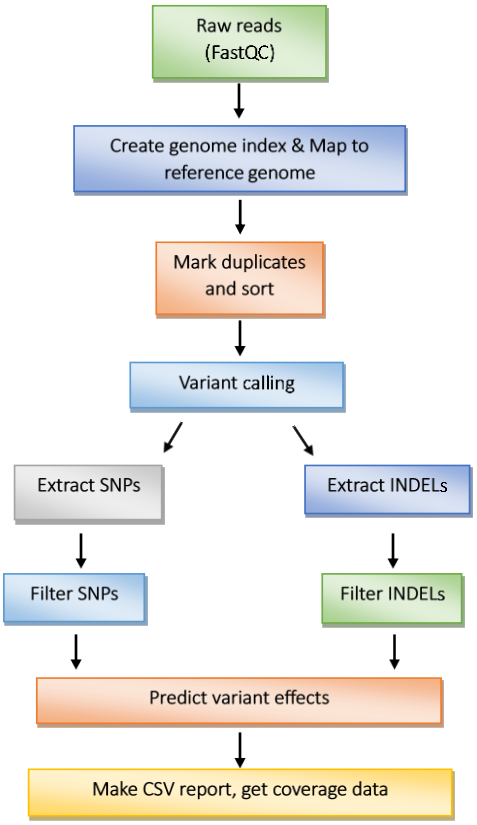

Fig 1. Workflow for SNP and InDel Identification and Annotation from RNA-Seq Data

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

1. Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature reviews genetics*, 12(5), 363-376.
2. Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

3. Bickhart, D. M., Hou, Y., Schroeder, S. G., Alkan, C., Cardone, M. F., Matukumalli, L. K., ... and Liu, G. E. (2012). Copy number variation of individual cattle genomes using next-generation sequencing. *Genome research*, 22(4), 778-790.
4. Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120.
5. Brenner, C. H. and Weir, B. S. (2003). Issues and strategies in the DNA identification of World Trade Center victims. *Theoretical Population Biology*, 63(3), 173-178.
6. Bushnell, B., Rood, J., and Singer, E. (2017). BBMerge–accurate paired shotgun read merging via overlap. *PloS one*, 12(10), e0185056.
7. Bickhart, H., Wang, N., Zhao, X., Ross, C. A., O'shea, K. S., and McInnis, M. G. (2013). Gene expression alterations in bipolar disorder postmortem brains. Bipolar disorders, 15(2), 177-187.
8. Chen, L. Y., Wei, K. C., Huang, A. C. Y., Wang, K., Huang, C. Y., Yi, D., ... and Hood, L. E. (2012). RNASEQR—a streamlined and accurate RNA-seq sequence analysis program. Nucleic acids research, 40(6), e42-e42.
9. Chen, H., Wang, N., Zhao, X., Ross, C. A., O'shea, K. S., and McInnis, M. G. (2013). Gene expression alterations in bipolar disorder postmortem brains. Bipolar disorders, 15(2), 177-187.
10. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884-i890.
11. Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics, 12(7), 499-510.
12. 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.
13. Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*, *6*(5), e19379.
14. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), 3047-3048.
15. Glaszmann, J. C. (1987). Isozymes and classification of Asian rice varieties. *Theoretical and Applied genetics*, 74, 21-30.
16. Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature reviews genetics*, 17(6), 333-351.
17. Goya, R., Sun, M. G., Morin, R. D., Leung, G., Ha, G., Wiegand, K. C., ... and Shah, S. P. (2010). SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, 26(6), 730-736.

18. Horton, M. W., Hancock, A. M., Huang, Y. S., Toomajian, C., Atwell, S., Auton, A., ... and Bergelson, J. (2012). Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature genetics*, 44(2), 212-216.
19. Huang, D. W., Sherman, B. T., Zheng, X., Yang, J., Imamichi, T., Stephens, R., and Lempicki, R. A. (2009). Extracting biological meaning from large gene lists with DAVID. *Current protocols in bioinformatics*, 27(1), 13-11.
20. Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology*, 37(8), 907-915.
21. Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., ... and Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3), 568-576.
22. Konishi, S., Izawa, T., Lin, S. Y., Ebana, K., Fukuta, Y., Sasaki, T., and Yano, M. (2006). An SNP caused loss of seed shattering during rice domestication. *Science*, 312(5778), 1392-1396.
23. Ku, C. S., Loy, E. Y., Salim, A., Pawitan, Y., and Chia, K. S. (2010). The discovery of human genetic variations and their use as disease markers: past, present and future. *Journal of human genetics*, 55(7), 403-415.
24. Kumar, S., Banks, T. W., and Cloutier, S. (2016). SNP discovery through next-generation sequencing and its applications. *Crop Breeding*, 209-244.
25. Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), 357-359.
26. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21), 2987-2993.
27. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, 25(14), 1754-1760.
28. Liu, Z. J., and Cordes, J. F. (2004). DNA marker technologies and their applications in aquaculture genetics. *Aquaculture*, 238(1-4), 1-37.
29. Mackenzie, R. J. (2018). RNA-Seq: Basics, applications and protocol. *Technology Networks Genomics Research.*
30. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1), 10-12.
31. McCarthy, M. I., and Hirschhorn, J. N. (2008). Genome-wide association studies: past, present and future. *Human molecular genetics*, 17(R2), R100-R101.
32. McNally, K. L., Childs, K. L., Bohnert, R., Davidson, R. M., Zhao, K., Ulat, V. J., ... and Leach, J. E. (2009). Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences*, 106(30), 12273-12278.

33. Montgomery, D. C., and Runger, G. C. (2010). *Applied statistics and probability for engineers*. John wiley and sons.
34. Morin, P. A., Luikart, G., Wayne, R. K., and SNP Workshop Group. (2004). SNPs in ecology, evolution and conservation. *Trends in ecology and evolution*, 19(4), 208-216.
35. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7), 621-628.
36. Mullaney, J. M., Mills, R. E., Pittard, W. S., and Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Human molecular genetics*, 19(R2), R131-R136.
37. Nachman, M. W. (2004). Haldane and the first estimates of the human mutation rate. *Journal of Genetics*, 83, 231-233.
38. Nowell, P. C. (1976). The Clonal Evolution of Tumor Cell Populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression. *Science*, 194(4260), 23-28.
39. Obenchain, V., Lawrence, M., Carey, V., Gogarten, S., Shannon, P., and Morgan, M. (2014). VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, 30(14), 2076-2078.
40. O'Roak, B. J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J. J., Girirajan, S., ... and Eichler, E. E. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature genetics*, 43(6), 585-589.
41. Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12), 1413-1415.
42. Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., ... and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289), 768-772.
43. Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R., WGS500 Consortium, ... and Lunter, G. (2014). Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics*, 46(8), 912-918.
44. Salk, J. J., Schmitt, M. W., and Loeb, L. A. (2018). Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nature Reviews Genetics*, 19(5), 269-285.
45. Sauna, Z. E., and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics*, 12(10), 683-691.
46. Schaeffer, L. R. (2006). Strategy for applying genome-wide selection in dairy cattle. *Journal of animal Breeding and genetics*, 123(4), 218-223.
47. Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6), 863-864.

48. Seddon, J. M., Parker, H. G., Ostrander, E. A., and Ellegren, H. (2005). SNPs in ecological and conservation studies: a test in the Scandinavian wolf population. *Molecular Ecology*, 14(2), 503-511.
49. Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., ... and Aparicio, S. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403), 395-399.
50. Singh, B. P., Jayaswal, P. K., Singh, B., Singh, P. K., Kumar, V., Mishra, S., ... and Singh, N. K. (2015). Natural allelic diversity in OsDREB1F gene in the Indian wild rice germplasm led to ascertain its association with drought tolerance. *Plant cell reports*, 34, 993-1004.
51. Singh, B., Singh, N., Mishra, S., Tripathi, K., Singh, B. P., Rai, V., ... and Singh, N. K. (2018). Morphological and molecular data reveal three distinct populations of Indian wild rice Oryza rufipogon Griff. species complex. *Frontiers in Plant Science*, 9, 272865.
52. Singh, N., Choudhury, D. R., Singh, A. K., Kumar, S., Srinivasan, K., Tyagi, R. K., ... and Singh, R. (2013). Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. *PloS one*, 8(12), e84136.
53. Smith, C. T., Templin, W. D., Seeb, J. E., and Seeb, L. W. (2005). Single nucleotide polymorphisms provide rapid and accurate estimates of the proportions of US and Canadian Chinook salmon caught in Yukon River fisheries. *North American Journal of Fisheries Management*, 25(3), 944-953.
54. Soemedi, R., Cygan, K. J., Rhine, C. L., Wang, J., Bulacan, C., Yang, J., ... and Fairbrother, W. G. (2017). Pathogenic variants that alter protein code often disrupt splicing. *Nature genetics*, 49(6), 848-855.
55. Subbaiyan, G. K., Waters, D. L., Katiyar, S. K., Sadananda, A. R., Vaddadi, S., and Henry, R. J. (2012). Genome-wide DNA polymorphisms in elite indica rice inbreds discovered by whole-genome sequencing. *Plant Biotechnology Journal*, 10(6), 623-634.
56. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., ... and Mering, C. V. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1), D607-D613.
57. Tan, A., Abecasis, G. R., and Kang, H. M. (2015). Unified representation of genetic variants. *Bioinformatics*, 31(13), 2202-2204.
58. Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2), 178-192.
59. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., ... and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3), 562-578.

60. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., ... and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5), 511-515.
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.
63. Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ... and DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 43(1), 11-10.
64. Vogelstein, B., and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature medicine*, 10(8), 789-799.
65. Voichek, Y., and Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nature genetics*, 52(5), 534-540.
66. Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1), 57-63.
67. Yeang, C. H., McCormick, F., and Levine, A. (2008). Combinatorial patterns of somatic gene mutations in cancer. *The FASEB journal*, 22(8), 2605-2622.
68. Yim, B., Winkelmann, T., Ding, G. C., and Smalla, K. (2015). Different bacterial communities in heat and gamma irradiation treated replant disease soils revealed by 16S rRNA gene analysis–contribution to improved aboveground apple plant growth. *Frontiers in microbiology*, 6, 1224.
69. Yu, H., Xie, W., Wang, J., Xing, Y., Xu, C., Li, X., ... and Zhang, Q. (2011). Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. *PloS one*, 6(3), e17595.