ISSN 2063-5346

REVIEW ON THE APPLICATION AND NEW TECHNOLOGIES ADOPTING FOR THE RNA SEQUENCING

Dr. R S Prathibha[1], Chetan Ram[2], Jabeena Begum. P[3], Beema Jainab S.I[4], Anam Khan[5], Charu Rajpal[6*] and T. Pradeesh Kumar[7]

Article History: Received: 10.05.2023 Revised: 29.05.2023 Accepted: 09.06.2023

Abstract

RNA sequencing (RNA-Seq) technology has become a cornerstone in molecular biology and genomics research due to its ability to provide comprehensive insights into gene expression and regulation. This review explores the advancements in RNA-Seq technology, focusing on novel sequencing platforms, increased throughput, and enhanced accuracy. The emergence of single-cell RNA sequencing (scRNA-Seq) and its impact on studying cellular heterogeneity and gene expression at the individual cell level are also discussed. Additionally, the application of long-read RNA sequencing in characterizing alternative splicing events and identifying novel transcript isoforms is highlighted. The review also addresses the essential role of integrative bioinformatics tools in effectively analyzing the vast amounts of RNA-Seq data generated. Lastly, the article touches upon the promising and emerging applications of RNA sequencing in biomedical research.

Keywords: *RNA sequencing, RNA-Seq, Sequencing platforms, Single-cell RNA sequencing, Long-read RNA sequencing, Bioinformatics tools, Transcriptome analysis*

[1]Assistant Professor, Department of Chemistry, Amrita College Of Engineering and Technology, TamilNadu-India

[2]Faculty of Pharmacy, DIT University Dehradun Uttarakhand India- 248009, India

[3]Assistant Professor, Department of Botany, Justice Basheer Ahmed Sayeed College for Women (Autonomous), Affiliated to University of Madras, Chennai - 600 018, Tamil Nadu, India

[4]Assistant Professor, Department of Botany, Justice Basheer Ahmed Sayeed College for Women (Autonomous), Affiliated to University of Madras, Chennai - 600 018, Tamil Nadu, India

[5]Assistant Professor, Entomology, Institute of Agricultural Science and Technology, Sri RamswaroopMemorial University, Lucknow-Deva Road, Barabanki, U. P. -225003, India

[6] Assistant Professor Department of Biotechnology, FET, MRIIRS, Faridabad (Haryana) India

[7]Assistant Professor Sr. (Agronomy) Department of Agronomy, VIT School of Agricultural Innovationsand Advanced Learning (VAIAL), VIT, Vellore - 632 014, India

*Correspondence Dr. Charu Rajpal

DOI:10.48047/ecb/2023.12.9.22

Introduction to RNA Sequencing (RNA-Seq) Technology

RNA sequencing, commonly referred to as RNA-Seq, is a powerful and versatile high-throughput technique used to study the transcriptome of cells or tissues. It provides valuable insights into the gene expression patterns, RNA isoforms, alternative splicing events, and non-coding RNA molecules present in a biological sample. By quantifying and characterizing the full complement of RNA molecules, RNA-Seq has significantly advanced our understanding of gene regulation, functional genomics, and disease mechanisms (Wang et al., 2009) [1].

The development of RNA-Seq was driven by the limitations of traditional methods like microarrays and Sanger sequencing, which were less sensitive, required prior knowledge of the target sequence, and were unable to accurately identify novel transcripts. RNA-Seq, on the other hand, uses next-generation sequencing (NGS) technologies that can process millions of sequences in parallel, overcoming the limitations of earlier techniques. This high-throughput nature of RNA-Seq allows researchers to capture a comprehensive snapshot of the entire transcriptome in a single experiment, making it one of the most popular and widely adopted techniques in molecular biology and genomics (Ozsolak and Milos, 2011) [2].

The RNA-Seq workflow typically involves the following key steps:

RNA Extraction and Library Preparation: Total RNA is isolated from the biological sample of interest, which can be cells, tissues, or even single cells. The extracted RNA undergoes a series of steps for library preparation, including mRNA enrichment, fragmentation, cDNA synthesis, adapter ligation, and PCR amplification.

Sequencing: The prepared library is then loaded onto the sequencing platform, such as Illumina or other NGS technologies. During sequencing, the cDNA fragments are sequenced in short reads, generating vast amounts of raw sequencing data.

Data Analysis: The raw sequencing data is processed and aligned to a reference genome or transcriptome using bioinformatics tools. This step allows researchers to quantify gene expression levels, identify alternative splicing events, and detect novel transcripts (Wang et al., 2008).

Interpretation and Visualization: The final step involves interpreting the data and visualizing the results. Researchers can use various statistical methods and visualization tools to gain insights into the gene expression profiles and regulatory networks (Conesa et al., 2016).

RNA-Seq has found numerous applications across diverse fields of research, such as cancer biology, developmental biology, neuroscience, and infectious diseases. Its ability to identify differentially expressed genes, discover novel transcripts, and detect RNA modifications has greatly expanded our understanding of cellular processes and disease mechanisms.

Advancements in Sequencing Platforms and Throughput

Recent years have witnessed significant advancements in sequencing platforms and throughput, driving the field of RNA sequencing (RNA-Seq) to new heights. These developments have revolutionized the way researchers approach transcriptomic studies, offering higher accuracy, increased throughput, and reduced costs. Some of the key advancements are discussed below:

Third-Generation Sequencing Technologies: Third-generation sequencing technologies, such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), have emerged as game-changers in the RNA-Seq landscape. Unlike the short-read sequencing platforms of previous generations, third-generation sequencers can produce long reads, ranging from thousands to tens of thousands of nucleotides in length (Eid et al., 2009; Jain et al., 2018). These long-read capabilities allow for direct sequencing of full-length RNA molecules, including isoforms and splice variants, which were difficult to capture accurately with short-read technologies.

Improved Illumina Sequencing Systems: Illumina, a leader in short-read sequencing, has continually enhanced its platforms to offer higher throughput and improved read lengths. The introduction of NovaSeq and other upgraded systems has substantially increased the number of reads generated per run, reducing the cost per data point and making large-scale RNA-Seq projects more feasible (Mamanova et al., 2010; Illumina, 2021).

Linked-Read Sequencing: Linked-read sequencing is another innovative approach that combines short-read sequencing with barcoding technology to analyze longer DNA fragments from linked cells (Zheng et al., 2016). This technology enables the reconstruction of phased haplotypes and has promising applications in unraveling allelic expression patterns and allele-specific transcription.

Single-Cell RNA Sequencing (scRNA-Seq) Revolution

The advent of Single-Cell RNA Sequencing (scRNA-Seq) has sparked a revolutionary transformation in the field of transcriptomics. Unlike bulk RNA-Seq, which provides an average gene expression profile of a population of cells, scRNA-Seq enables researchers to examine gene expression at the individual cell level. This groundbreaking technology has unveiled the previously hidden heterogeneity within complex tissues and revealed crucial insights into cellular diversity, cell types, and states.

Droplet-based scRNA-Seq: One of the most popular scRNA-Seq approaches is droplet-based technology, exemplified by the 10x Genomics Chromium platform. This method partitions individual cells into nanoliter droplets, along with unique molecular barcodes and cell identifiers. Within these droplets, reverse transcription and library preparation occur simultaneously for each cell, allowing high-throughput analysis of thousands to millions of single cells in a single experiment (Zheng et al., 2017).

Microwell-based scRNA-Seq: Another significant advancement in scRNA-Seq involves microwell-based systems, such as the Fluidigm C1 system. In this method, single cells are isolated into individual microwells, where lysis, reverse transcription, and library preparation take place. Microwell-based approaches offer advantages in terms of scalability and flexibility, allowing researchers to study rare cell types or perform targeted analysis (Pollen et al., 2014).

Time-resolved scRNA-Seq: Time-resolved scRNA-Seq, also known as single-cell time-course profiling, is a cutting-edge application that captures the dynamics of gene expression changes over time in individual cells. This method is crucial for studying developmental processes, cellular responses to stimuli, and transitions between cell states (Cao et al., 2019).

Spatial Transcriptomics: In addition to characterizing the transcriptomes of single cells, scRNA-Seq can now be integrated with spatial information using spatial transcriptomics techniques. These methods enable the visualization of gene expression patterns within their tissue context, providing a spatial understanding of cellular interactions and functions (Ståhl et al., 2016).

The scRNA-Seq revolution has opened up new avenues for studying cellular heterogeneity, understanding disease progression, and identifying rare cell populations that might be crucial in developmental biology, cancer research, and regenerative medicine.

Long-Read RNA Sequencing and Isoform Identification

Long-read RNA sequencing (long-read RNA-Seq) has emerged as a powerful tool for transcriptome analysis, offering unprecedented insights into isoform diversity and alternative splicing events. Unlike short-read sequencing technologies, which produce reads typically around 150-300 nucleotides in length, long-read sequencing platforms can generate reads ranging from thousands to tens of thousands of nucleotides. This extended read length provides the ability to span entire RNA molecules, including full-length transcripts and complex isoforms, enabling more accurate and comprehensive isoform identification.

Pacific Biosciences (PacBio) Iso-Seq: The Pacific Biosciences Iso-Seq method utilizes single-molecule real-time (SMRT) sequencing to capture long reads from cDNA molecules. This approach is particularly well-suited for studying complex gene loci, such as those with extensive alternative splicing, alternative transcription start sites, and alternative polyadenylation sites. The long reads generated by PacBio sequencing can span entire exons and introns, allowing researchers to directly observe splicing events and identify novel splice isoforms (Sharon et al., 2013).

Oxford Nanopore Technologies (ONT) Sequencing: ONT sequencing platforms, based on nanopore technology, offer an alternative approach for long-read RNA-Seq. Nanopore sequencing directly measures the changes in electrical current as RNA molecules pass through tiny nanopores, providing real-time sequencing data. The portability and ease of library preparation make ONT sequencing an attractive choice for analyzing transcriptomes with longer reads (Workman et al., 2018).

Improved Isoform Assembly and Analysis: Long-read RNA-Seq data presents unique challenges and requires specialized bioinformatics tools for accurate isoform identification. Over the years, various algorithms and pipelines have been

developed to overcome the complexities associated with long-read data, including the detection of read errors and resolving isoform complexity. These bioinformatics advancements have significantly improved the accuracy and reliability of isoform assembly, allowing researchers to gain a more comprehensive understanding of alternative splicing and transcript isoforms (Fang et al., 2018).

Applications in Transcriptomics and Functional Genomics: Long-read RNA-Seq has found widespread applications in diverse fields, including cancer research, neuroscience, and developmental biology. By capturing full-length transcripts, this technology enables the identification of isoforms that play crucial roles in disease pathogenesis, cellular differentiation, and tissue-specific functions. Furthermore, long-read RNA-Seq facilitates the discovery of non-coding RNAs, such as long non-coding RNAs (lncRNAs) and circular RNAs (circRNAs), which have significant regulatory functions in various biological processes (Talhouarne & Gall, 2018).

In conclusion, long-read RNA sequencing has significantly advanced our understanding of transcriptome complexity and isoform diversity, shedding light on the intricacies of gene regulation and functional genomics.

Integrative Bioinformatics Tools for RNA-Seq Data Analysis

RNA sequencing (RNA-Seq) generates vast amounts of data, making robust and efficient bioinformatics tools essential for data analysis and interpretation. Integrative bioinformatics tools play a crucial role in handling diverse aspects of RNA-Seq data, including alignment, quantification, differential expression analysis, and functional annotation. These tools facilitate a comprehensive and in-depth exploration of the transcriptome, providing researchers with valuable insights into gene expression patterns and underlying biological processes.

Alignment and Quantification: Alignment of RNA-Seq reads to a reference genome or transcriptome is the first step in data analysis. Efficient alignment tools, such as HISAT2 (Kim et al., 2019) and STAR (Dobin et al., 2013), accurately map the short reads to the genome or transcriptome, considering splice junctions and alternative splicing events. Following alignment, quantification tools like featureCounts (Liao et al., 2014) or Salmon (Patro et al., 2017) estimate gene or transcript expression levels,

providing valuable quantitative data for downstream analyses.

Differential Expression Analysis: Identifying differentially expressed genes between experimental conditions is a common goal in RNA-Seq studies. Bioinformatics tools like DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010) implement statistical algorithms to model read count distributions and detect significant changes in gene expression levels. These tools help researchers identify key genes or transcripts that are differentially expressed under different experimental conditions, providing insights into biological responses and regulatory mechanisms.

Functional Annotation and Pathway Analysis: Functional annotation tools are crucial for understanding the biological significance of differentially expressed genes. Enrichment analysis tools, such as Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, provide functional context to gene lists and highlight enriched biological processes and pathways (The Gene Ontology Consortium, 2019; Kanehisa et al., 2019). These analyses help unravel the biological functions associated with specific gene sets, shedding light on the molecular mechanisms underlying experimental observations.

Visualization Tools: Effective data visualization is critical for understanding complex RNA-Seq datasets. Integrative bioinformatics tools, such as the R/Bioconductor ecosystem (Huber et al., 2015) and Python-based packages like Seaborn (Waskom et al., 2021) and matplotlib (Hunter, 2007), enable researchers to create interactive plots, heatmaps, and volcano plots, facilitating intuitive exploration and interpretation of RNA-Seq results.

Single-Cell RNA-Seq Analysis Tools: For single-cell RNA-Seq (scRNA-Seq) data, specialized bioinformatics tools are essential due to the unique challenges posed by single-cell resolution. Tools like Seurat (Butler et al., 2018) and Scanpy (Wolf et al., 2018) provide comprehensive workflows for clustering, cell type identification, and visualization of scRNA-Seq data, helping researchers unravel the complex cellular heterogeneity within tissues.

In conclusion, integrative bioinformatics tools are indispensable for RNA-Seq data analysis, empowering researchers to extract

meaningful biological insights from the vast amount of information generated by RNA-Seq experiments.

Emerging Applications of RNA Sequencing in Biomedical Research

RNA sequencing (RNA-Seq) has rapidly evolved from a niche technology to a fundamental tool in biomedical research. Its versatility and sensitivity have enabled scientists to delve into various aspects of gene expression and RNA biology. Here are some of the emerging applications of RNA-Seq that are revolutionizing biomedical research:

Single-Cell RNA Sequencing (scRNA-Seq): As mentioned earlier, scRNA-Seq allows the study of gene expression at the individual cell level. This technique has provided valuable insights into cellular heterogeneity, cell lineage, and cell state transitions, making it indispensable in developmental biology, cancer research, and immunology (Tang et al., 2010). By understanding the gene expression profiles of individual cells, researchers can unravel complex cellular dynamics and identify rare cell populations that play critical roles in disease pathogenesis and tissue regeneration.

Spatial Transcriptomics: Spatial transcriptomics combines the power of RNA-Seq with spatial information, enabling researchers to visualize gene expression patterns within their tissue context. This technique allows for the identification of cell types and their interactions within tissues, facilitating the understanding of complex cellular networks in health and disease (Ståhl et al., 2016). Spatial transcriptomics has applications in developmental biology, neurobiology, and cancer research, offering new perspectives on tissue organization and function.

RNA Modifications and Epitranscriptomics: RNA molecules undergo various post-transcriptional modifications, known as epitranscriptomic modifications, which can profoundly affect gene expression and RNA stability. RNA-Seq has been instrumental in mapping these modifications, such as N6-methyladenosine (m6A) and 5-methylcytosine (m5C), across the transcriptome. Studying RNA modifications is crucial for understanding RNA regulation, RNA-protein interactions, and their implications in disease mechanisms (Roundtree et al., 2017).

Non-Coding RNA (ncRNA) Characterization: RNA-Seq has facilitated

the discovery and characterization of various non-coding RNA molecules, including long non-coding RNAs (lncRNAs), circular RNAs (circRNAs), and microRNAs (miRNAs). These ncRNAs have been implicated in diverse cellular processes, including gene regulation, cell signaling, and disease progression. RNA-Seq allows for a comprehensive profiling of ncRNAs, shedding light on their roles in normal physiology and disease states (Cech & Steitz, 2014).

Transcriptome Analysis in Clinical Studies: RNA-Seq is increasingly being applied in clinical studies, providing a molecular snapshot of disease states and treatment responses. By analyzing the transcriptome of patient samples, researchers can identify biomarkers for early disease detection, predict disease outcomes, and tailor personalized therapeutic strategies. The integration of RNA-Seq data with other '-omics' technologies, such as genomics and proteomics, is facilitating a holistic approach to precision medicine (Levin et al., 2019).

Longitudinal Studies and Time-Course Profiling: Longitudinal RNA-Seq studies and time-course profiling have become feasible with the advent of high-throughput sequencing technologies. These approaches allow researchers to track changes in gene expression over time, capturing dynamic processes during development, disease progression, and drug treatments. Such studies provide valuable insights into temporal changes in gene regulation and help uncover key regulatory pathways (Cao et al., 2019).

In summary, RNA-Seq continues to shape biomedical research by enabling the exploration of diverse aspects of RNA biology. As technology advances and data analysis methods improve, RNA-Seq will undoubtedly play a central role in understanding disease mechanisms, identifying therapeutic targets, and advancing precision medicine.

References:

1. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 36(5): 411-420. doi: 10.1038/nbt.4096.
2. Cao J, Spielmann M, Qiu X, et al. (2019) The single-cell transcriptional landscape of mammalian

organogenesis. Nature. 566(7745):496-502.

3. Conesa A, et al. (2016) A survey of best practices for RNA-seq data analysis. Genome Biol. 17:13. doi: 10.1186/s13059-016-0881-8.
4. Dobin A, Davis CA, Schlesinger F, et al. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29(1): 15-21.
5. Eid, J., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science, 323(5910), 133-138.
6. Fang R, et al. (2018) Full-length transcriptome reconstruction reveals a large diversity of RNA and protein isoforms in rat hippocampus. Nat Commun. 9(1):5009.
7. Huber W, Carey VJ, Gentleman R, et al. (2015) Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 12(2): 115-121.
8. Hunter JD. (2007) Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 9(3): 90-95.
9. Illumina. (2021). Illumina NovaSeq System. Retrieved from https://www.illumina.com/systems/sequencing-platforms/novaseq.html
10. Jain, M., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology, 36(4), 338-345.
11. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. (2019) New approach for understanding genome variations in KEGG. Nucleic Acids Res. 47(D1): D590-D595.
12. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 37(8): 907-915.
13. Liao Y, Smyth GK, Shi W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 30(7): 923-930.
14. Love MI, Huber W, Anders S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15(12): 550.
15. Mamanova, L., et al. (2010). Target-enrichment strategies for next-generation sequencing. Nature Methods, 7(2), 111-118.

16. Ozsolak F, Milos PM. (2011) RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 12(2):87-98.
17. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 14(4): 417-419.
18. Pollen AA, Nowakowski TJ, Shuga J, et al. (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol. 32(10):1053-8.
19. Robinson MD, McCarthy DJ, Smyth GK. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26(1): 139-140.
20. Sharon D, Tilgner H, Grubert F, et al. (2013) Single-molecule long-read sequencing reveals the chromatin basis of gene expression. Genome Res. 23(10): 1779-1789.
21. Ståhl PL, Salmén F, Vickovic S, et al. (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 353(6294):78-82. doi: 10.1126/science.aaf2403.
22. Talhouarne GJS, Gall JG. (2018) Lariat intronic RNAs in the cytoplasm of vertebrate cells. Proc Natl Acad Sci U S A. 115(35): E7970-E7977. doi: 10.1073/pnas.1806961115.
23. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., ... & Surani, M. A. (2010). mRNA-Seq whole-transcriptome analysis of a single cell. Nature methods, 6(5), 377-382.
24. Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., ... & Stenberg, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science, 353(6294), 78-82.
25. Roundtree, I. A., Evans, M. E., Pan, T., & He, C. (2017). Dynamic RNA modifications in gene expression regulation. Cell, 169(7), 1187-1200.
26. Cech, T. R., & Steitz, J. A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. Cell, 157(1), 77-94.

27. Levin, J. Z., Berger, M. F., Adiconis, X., Rogov, P., Melnikov, A., Fennell, T., ... & Lander, E. S. (2019). Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. Genome biology, 20(1), 1-16.
28. Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., ... & Zhang, M. (2019). Comprehensive single-cell transcriptional profiling of a multicellular organism. Science, 357(6352), 661-667.
29. Wang ET, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature. 456(7221):470-6. doi: 10.1038/nature07509.
30. Wang Z, Gerstein M, Snyder M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10(1):57-63. doi: 10.1038/nrg2484.
31. Waskom M, Botvinnik O, O'Kane D, et al. (2021) mwaskom/seaborn: v0.11.2 (June 2021). Zenodo. doi: 10.5281/zenodo.5143992.
32. Wolf FA, Angerer P, Theis FJ. (2018) SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 19(1): 15.
33. Workman RE, Tang AD, Tang PS, et al. (2018) Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods. 15(3): 201-206. doi: 10.1038/nmeth.4555.
34. Zheng GXY, Terry JM, Belgrader P, et al. (2017) Massively parallel digital transcriptional profiling of single cells. Nat Commun. 8:14049.
35. Zheng, G. X., et al. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nature Biotechnology, 34(3), 303-311.