Review of RNA Sequencing Methods

Kuo Fang[1, a], Lihua Ding[2, b], Ruijie Lin[3, c], Sike Yao[4, d], Yikun Liu[5, e], Yiyan Luo[6, f]

[1] The Chinese University of HongKong (CUHK), Shenzhen, China, 2001 Longxiang Boulevard, Longgang District, Shenzhen, 518172, China

[2] Grier School, 2522 Grier School Rd, Tyrone, PA, PA 16686, USA

[3] Crean Lutheran High School, 12500 Sand Canyon Ave, Irvine, CA, CA 92618, USA

[4] Harrow International Hong Kong, 38 Tsing Ying Road, Tuen Mun, New Territories, Hong Kong, 999077, China

[5] Crean Lutheran High School, 12500 Sand Canyon Ave, Irvine, CA, CA 92618, USA

[6] University of Manchester, Oxford Rd, Manchester, M13 9PL, UK

[a] 121050014@link.cuhk.edu.cn, [b] emily.ding0831@outlook.com, [c] cliolin.us@gmail.com,

[d] Kabulelan929@gmail.com, [e] 4ndrew778@gmail.com, [f] yiyan.luo@student.manchester.ac.uk

Abstract

Since its inception, RNA sequencing (RNA-seq) technology has revolutionized our ability to analyze gene expression, discover novel transcripts, and identify differential splicing events at the whole-transcriptome level. This review systematically outlines the development of RNA-seq technologies, focusing on comparing the principles, workflows, advantages and disadvantages, and application scenarios of three mainstream technical approaches: short-read cDNA sequencing, long-read cDNA sequencing, and long-read direct RNA sequencing. Short-read sequencing remains the gold standard for quantitative analysis due to its high accuracy and low cost; long-read cDNA sequencing (e.g., PacBio and Nanopore cDNA sequencing) can seamlessly span repetitive regions and homologous genes, enabling precise reconstruction of transcript isoforms; while the emerging long-read direct RNA sequencing (e.g., Nanopore) preserves native RNA modification information, opening new avenues for epitranscriptomics research. The article also reviews the latest applications of RNA-seq in disease research, developmental biology, and single-cell analysis, and discusses future directions, such as multi-omics integration, improvement of long-read quantification accuracy, and development of bioinformatics tools.

Keywords

RNA sequencing, transcriptome, long-read sequencing, differential expression, alternative splicing.

1. Introduction

Ribonucleic acid (RNA) serves as a critical bridge connecting genomic genetic information with the functional proteome. Its types, quantities, and structural dynamics determine cell fate and function. Systematic analysis of RNA at the transcriptome level is crucial for understanding gene regulatory mechanisms, cellular responses to internal and external environments, and the pathogenesis of diseases [1]. Before the advent of RNA sequencing technology, gene expression analysis primarily relied on microarray technology, which offered high throughput but depended on pre-designed probes, could not discover novel transcripts or splice variants, and had a limited dynamic range [2].

The emergence of RNA sequencing (RNA-seq) technology marked the beginning of a new era in transcriptome research. Its fundamental concept involves massively parallel sequencing of reverse-transcribed cDNA libraries on high-throughput sequencing platforms, enabling unbiased capture of nearly all transcript information [3]. Since its commercialization around 2008, RNA-seq has rapidly become a core tool for studying differential gene expression (DGE), alternative splicing, gene fusion, allele-specific expression, and de novotranscript discovery, owing to its high sensitivity, broad dynamic range, and independence from prior gene annotations [4, 5].

Over the past decade, RNA-seq technology itself has evolved rapidly. Its development has primarily revolved around two core dimensions: read length and library preparation strategies. The field has progressed from being dominated by early short-read sequencing to the current complementary landscape where long-read sequencing coexists with short-read sequencing. This review will systematically elaborate on the technical principles, workflows, and applications of different RNA-seq methods, and future development trends.

2. Short-read cDNA Sequencing

Short-read cDNA sequencing technology originated around 2008, propelled and validated by multiple research teams, marking a definitive shift from the microarray era to high-throughput sequencing for transcriptome analysis [3-5]. Its core principle involves first fragmenting the entire RNA population into small segments of hundreds of bases, then converting them into double-stranded cDNA libraries through reverse transcription, and finally generating massive amounts of short reads (typically 50-300 bp) on sequencing platforms like Illumina using sequencing-by-synthesis technology [3, 6]. This method remains the undisputed gold standard for whole-transcriptome differential gene expression analysis due to its exceptional quantitative accuracy, high detection throughput, and relatively low cost.

A complete short-read RNA-seq workflow begins with the extraction of high-quality RNA. Obtaining total RNA with high integrity (typically indicated by an RNA Integrity Number (RIN) > 8) is the primary prerequisite for ensuring data reliability. Subsequently, poly(A)-tailed mRNA in eukaryotes is specifically enriched using oligo(dT) beads, or total RNA including non-coding RNA is enriched using ribosomal RNA removal kits (rRNA-depletion). The enriched RNA is fragmented, and reverse transcription is performed using random hexamer primers or oligo(dT) primers to synthesize first-strand cDNA, followed by synthesis of double-stranded cDNA. After steps such as end repair and A-tailing, sequencing adapters containing sample-specific indices (barcodes) are ligated, and the library is finally constructed via PCR amplification for sequencing [6]. The raw data generated from sequencing undergoes rigorous quality control (e.g., using FastQC), adapter trimming, and is then aligned to a reference genome (Alignment, using tools like STAR, HISAT2) or subjected to de novoassembly without a reference genome (using tools like Trinity). The aligned sequences can then be used for gene/transcript-level quantification (e.g., using featureCounts, HTSeq), and finally, statistical tools (e.g., DESeq2, edgeR) are employed for differential expression analysis, alternative splicing analysis, etc. [6].

The main advantages of this technology lie in its long-validated quantitative precision and cost-effectiveness. Extremely high sequencing depth ensures effective detection of low-abundance transcripts, while a mature bioinformatics analysis ecosystem provides powerful and reliable data interpretation support. However, its inherent limitations cannot be ignored. Short reads are difficult to map uniquely to repetitive regions of the genome, leading to a certain proportion of multi-mapping reads in the results. More importantly, since transcripts are physically fragmented at the beginning of library construction, full-length transcript information cannot be directly obtained. The analysis of complex transcriptional phenomena such as alternative

splicing, gene fusion, and allele-specific expression heavily relies on computational assembly and inference algorithms, whose accuracy and completeness have a ceiling [7].

To overcome these limitations, research progress in this field has mainly focused on optimizing library preparation methods. Among these, the application of Unique Molecular Identifiers (UMIs) is a major breakthrough. By introducing random molecular tags into the reverse transcription primers, duplicate molecules generated by amplification bias can be tracked and corrected during subsequent PCR amplification and sequencing, thereby achieving ultra-high precision quantification of the absolute number of transcript molecules. This is crucial for single-cell RNA-seq (scRNA-seq) and low-frequency mutation detection [8]. Furthermore, strand-specific library construction strategies, by preserving the information of the transcript's strand of origin, can accurately distinguish between sense and antisense strand transcripts, providing strong support for research on long non-coding RNAs (lncRNAs), among others. Despite the rapid development of long-read sequencing technologies, short-read cDNA sequencing will remain the mainstream choice for large-scale transcriptome studies, especially disease cohort studies, pharmacogenomics, and applied basic research, in the foreseeable future, due to its unparalleled quantitative robustness and cost-effectiveness [7].

3. Long-read cDNA Sequencing: Studying Transcript Isoforms

To fundamentally address the bottleneck of short-read sequencing in resolving transcript isoforms, long-read sequencing technologies emerged. These technologies primarily rely on two major platforms: Pacific Biosciences' (PacBio) Single Molecule Real-Time (SMRT) sequencing and Oxford Nanopore Technologies' (ONT) nanopore sequencing. Their core innovation lies in the ability to directly sequence near-full-length cDNA molecules, bypassing the fragmentation and assembly steps required by short-read sequencing and enabling direct observation of complete transcript isoforms [9, 10].

The workflow for long-read cDNA sequencing is largely consistent with short-read sequencing in the RNA extraction and mRNA enrichment steps. The key difference is the avoidance of RNA or cDNA fragmentation. Instead, full-length first-strand cDNA is synthesized using template-switching reverse transcription technology (e.g., SMARTer technology) and directly constructed into a sequencing library. On the PacBio platform, a DNA polymerase uses the cDNA molecule as a template for synthesis reactions in zero-mode waveguides (ZMWs), with real-time sequencing achieved by detecting incorporated fluorescently labeled nucleotides. On the ONT platform, full-length cDNA molecules with adapters are pulled through a nanopore by a motor protein, and the characteristic changes in electrical current caused by different nucleotides are decoded into sequence data [9, 10]. PacBio's latest Circular Consensus Sequencing (CCS) mode generates highly accurate HiFi reads by sequencing the same molecule multiple times. The raw data requires a series of quality control steps specific to long reads, error correction (e.g., using LoRDEC, NanoComp), and alignment to a reference genome (e.g., using Minimap2) or transcript isoform identification (e.g., using StringTie2, FLAIR) [11].

The most significant advantage of long-read cDNA sequencing is its ability to generate continuous reads that span entire transcripts, thereby enabling the inference-free direct resolution of complex transcriptional phenomena such as alternative promoters, exon skipping, intron retention, alternative polyadenylation (APA) site selection, and gene fusions. It greatly improves the accuracy and completeness of transcript annotation and can even directly discover novel transcripts [11]. However, this technology has long been hampered by two major shortcomings: firstly, a relatively high raw read error rate (particularly insertions and deletions), although CCS and repeated sequencing can effectively correct errors, it comes at the cost of reduced throughput; secondly, lower throughput and higher cost compared to short-read sequencing, limiting its application in large cohort studies. Furthermore, its bioinformatics

toolchain, while developing rapidly, still lacks the standardization and maturity of short-read analysis.

Nonetheless, long-read cDNA sequencing has become a powerful tool for constructing high-quality reference transcriptomes and in-depth study of transcript isoform function. It is widely used in precision medicine research, for example, identifying disease-causing specific splice variants in cancer, studying splicing dysregulation mechanisms in neurodegenerative diseases, and performing accurate genotyping of highly homologous gene families (e.g., HLA genes) [11]. As the cost of PacBio HiFi sequencing continues to decrease and the throughput of ONT sequencing increases, long-read sequencing is transitioning from a cutting-edge technology to routine application, forming a complementary and symbiotic landscape with short-read sequencing.

4. Long-read Direct RNA Sequencing: The New Frontier of Epitranscriptomics

Long-read direct RNA sequencing is a revolutionary leap in transcriptome sequencing technology, currently primarily enabled by the ONT platform. Its fundamental difference from all cDNA sequencing methods is that it completely bypasses reverse transcription and PCR amplification steps, sequencing the native RNA molecules directly [10, 12]. This not only preserves RNA sequence information but, more importantly, natively captures post-transcriptional RNA modifications, opening a new door for epitranscriptomics research.

Its workflow is more direct compared to cDNA sequencing. After total RNA extraction, RNA molecules with poly(A) tails are directly ligated to sequencing adapters. One adapter sequence is covalently linked to a motor protein that drives the RNA molecule through the nanopore. Upon application of a voltage, the RNA molecule is pulled through the nanopore at a controlled speed by the motor protein. Different bases (and their modifications) within the pore cause unique disruptions in the ionic current. The device records these signals, and algorithms (e.g., the base-calling algorithm Guppy) decode them directly into RNA sequence [12]. Analysis of the resulting data includes not only routine alignment and quantification but also utilizes the raw electrical signal information (e.g., using tools like Tombo, EpiNano) to detect and locate RNA modifications such as N6-methyladenosine (m6A), 5-methylcytidine (m5C), and pseudouridine (Ψ).

The ultimate advantage of direct RNA sequencing is its ability to simultaneously read sequence and modification information, enabling the study of "RNA epigenetics" at the single-molecule level. Simultaneously, by avoiding reverse transcription and PCR steps, it eliminates potential sequence biases and amplification artifacts inherent in those steps, allowing for a more authentic representation of the original transcriptome state. Additionally, it possesses all the advantages of long-read sequencing, enabling the resolution of full-length transcript isoforms [12]. However, this technology still faces several challenges: its raw error rate remains higher than cDNA sequencing, especially at the 5' end of reads; throughput and data output are still lower than cDNA modes; it requires higher quantity and quality of input RNA; most crucially, its data analysis pipeline, particularly the identification and quantification of modified bases, is extremely complex and still rapidly evolving, placing higher demands on bioinformatics analysis[13].

Despite these challenges, direct RNA sequencing has already become a powerful engine for epitranscriptomics research. Researchers have used it to map RNA modification landscapes across the entire transcriptome in various organisms under different physiological and pathological states, dynamically studying the regulatory roles of these modifications in cell differentiation, stress response, viral infection, and cancer progression [14-15]. Furthermore, it has been used to directly sequence viral RNA genomes (e.g., SARS-CoV-2) to study their

structure, evolution, and modifications. With continuous innovations in nanopore chemistry, sequencing reagent accuracy, and bioinformatics algorithms, direct RNA sequencing is expected to become more efficient, accurate, and accessible. This will ultimately advance our understanding of RNA biological function to a new level, moving from a paradigm of "sequence determines fate" to one of "sequence and modification determine fate."

5. Conclusion

Since its inception, the development of RNA sequencing technology has consistently revolved around the core goal of enabling more comprehensive, precise, and in-depth analysis of the transcriptome. This review has systematically outlined three mainstream technological routes: short-read cDNA sequencing, long-read cDNA sequencing, and long-read direct RNA sequencing. They do not represent a simple iterative replacement relationship but rather constitute a powerful technological ecosystem where functions complement and reinforce each other. Short-read sequencing (represented by the Illumina platform), relying on its over-a-decade of validated ultra-high quantitative accuracy, high throughput, and cost-effectiveness, still occupies an unshakable dominant position in differential expression analysis requiring large sample cohorts, remaining the "gold standard" for transcriptome quantitative research. Long-read cDNA sequencing (represented by PacBio HiFi and ONT cDNA modes) has made a revolutionary contribution by its ability to seamlessly span entire transcripts, fundamentally solving the inherent inferential uncertainty of short-read technology due to fragmentation and assembly. This allows it to demonstrate incomparable advantages in the precise identification of transcript isoforms, resolution of complex alternative splicing events, and construction of high-quality reference transcriptomes. The most cutting-edge direct RNA sequencing (led by the ONT platform) further expands the dimension of sequencing from "sequence" to "modification," achieving the simultaneous reading of sequence information and base modifications at the single-molecule level. It provides an unprecedented powerful tool for the emerging field of epitranscriptomics, allowing us to glimpse another layer of regulatory codes hidden in the RNA world.

References

[1] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57-63.

[2] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270(5235):467-470.

[3] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621-628.

[4] Nagalakshmi U, Wang Z, Waern K, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320(5881):1344-1349.

[5] Wilhelm BT, Marguerat S, Watt S, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008;453(7199):1239-1243.

[6] Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.

[7] The Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-1120.

[8] Kivioja T, Vähärautio A, Karlsson K, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. 2011;9(1):72-74.

[9] Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015;13(5):278-289.

[10] Garalde DR, Snell EA, Jachimowicz D, et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods. 2018;15(3):201-206.

[11] Amarasinghe SL, Su S, Dong X, et al. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21(1):30.

[12] Liu H, Begik O, Lucas MC, et al. Accurate detection of m6A RNA modifications in native RNA sequences. Nat Commun. 2019;10(1):4079.

[13] Marx V. Method of the Year: spatially resolved transcriptomics. Nat Methods. 2021;18(1):9-14.

[14] Heitzer E, Haque IS, Roberts CES, Speicher MR. Current and future perspectives of liquid biopsies in genomics-driven oncology. Nat Rev Genet. 2019;20(2):71-88.

[15] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83.