Journal of Scientific Reports 2025, 10(1), 37-51
ISSN: 2708-7085 (Online) & 3079-9317 (Print)
Journal: ijsab.com/jsr

The Quick Guide to RNA-Seq: From Data Acquisition to Functional Analysis

Morjina Solaiman [1]

[1] Department of Biological Sciences, University of Idaho, Moscow, Idaho, United States.
Email: sola4859@vandals.uidaho.edu

Abstract

RNA sequencing (RNA-Seq) has revolutionized transcriptomics by enabling comprehensive and unbiased gene expression profiling across diverse biological contexts. This review summarizes the key stages of RNA-Seq, from RNA isolation and library preparation to sequencing technologies and computational analyses. We discuss the evolution of sequencing platforms, including second-generation short-read and third-generation long-read technologies, highlighting their applications and limitations. Critical steps such as quality control, read alignment, transcript assembly, normalization, and differential gene expression analysis are detailed, emphasizing the importance of tailored methodological choices and robust statistical approaches. Despite advances, challenges remain, including sequencing biases, errors, and variability introduced by diverse bioinformatic pipelines. Visualization techniques such as MA and volcano plots enhance the interpretation of expression changes. Overall, RNA-Seq remains a powerful tool for understanding gene regulation, alternative splicing, and disease mechanisms, with ongoing improvements needed for standardized, reproducible workflows to maximize biological insight.

ARTICLE INFO

Review paper

Received: 24 April 2025
Accepted: 03 June 2025
Published: 06 June 2025
DOI: 10.58970/JSR.1114

CITATION

Solaiman, M. (2025). The Quick Guide to RNA-Seq: From Data Acquisition to Functional Analysis, *Journal of Scientific Reports*, 10(1), 37-51.

***Keywords**: RNA seq, Transcriptome, Sequence Analysis, RNA seq methods.*

Introduction to RNA-Seq:

The beginning of molecular biology is marked by Watson and Crick by discovering DNA, a mystery of life, revealing how information is transferred from DNA to RNA to protein to regulate cellular functions (Ginsberg, 2005)The first step in correctly understanding an organism's regulatory mechanism is to identify RNA molecules by RNA sequencing. This RNA sequencing offers advantages over microarrays by not requiring prior genome information about the organism. (Stark et al., 2019). Since its inception, RNA-Seq has enabled large-scale sequencing of the complete transcriptome, facilitating the study of gene expression in various biological contexts, including disease, cancer, and treatment responses (Marguerat & Bähler, 2010). Moreover, it facilitates the detection of novel genes, splicing transcripts, and variants, providing a more comprehensive and dynamic view of an organism (Z. Wang et al., 2009). High-throughput sequencing platforms like PacBio, Illumina, and Nanopore enable the extraction of millions of nucleotide sequences from a transcriptome, providing deep coverage of gene expression and alternative transcript quantification. Illumina sequencing uses an SBS method, also called sequencing by synthesis, where RNA is fragmented, adapters are ligated, and fluorescence is

detected to identify the nucleotide base (Thermes, 2014a). It offers high throughput but requires amplification and chemical modifications. RNA-Seq can be performed using either single-end or paired-end sequencing. While single-end sequencing is cost-effective and sufficient for basic transcript quantification, paired-end sequencing is generally preferred because it provides better alignment accuracy (Hrdlickova et al., 2017). Short-read sequencing (e.g., 50–150 bp) is ideal for profiling small RNAs (e.g., miRNAs, piRNAs) due to their compact size, while longer reads (≥150 bp or paired-end) are better suited for coding mRNAs and long non-coding RNAs (lncRNAs), as they improve transcript assembly, isoform resolution, and splice junction detection (Ozsolak & Milos, 2011).

The sharp drop in sequencing costs has made RNA-seq widely accessible, powering breakthroughs from single-cell studies to spatial genomics. As datasets grow (e.g., 200,000+ in GEO), thousands of new analysis tools have emerged, expanding possibilities but also complicating reproducibility (Costa et al., 2010). In addition, artificial intelligence is also used nowadays to process large-scale bulk RNA seq and single-cell RNA seq data, especially in pharmaceutical and cancer research. (Solaiman et al., 2025). Despite significant advancements, challenges remain, such as biases in library preparation, short lengths, sequencing errors, and the need for high-quality reference genomes, which many species lack. Although replicates can mitigate some of these variations, a key challenge remains the lack of a standardized RNA-seq processing pipeline. Different software tools and parameters can lead to variations in quantification, referred to as 'in silico design noise (Wilhelm & Landry, 2009). RNA-seq analysis requires tailored methodological choices at each step—from transcript quantification to differential expression—with the optimal approach depending on experimental design, biological context, and data characteristics (Saliba et al., 2014). Quality control is vital throughout the process for reproducibility and reliable results. This review provides an overview of Sequencing technologies, the RNA seq procedure, data analysis, and bioinformatic studies, prioritizing read mapping, quality control, expression quantification, and differential gene expression analysis. It also discusses standard practices, software, and tools, and the importance of advanced statistical methods, databases, and visualization techniques for improving result accuracy.

Sequencing technologies:

The foundation of modern DNA sequencing was established in 1975 when Frederick Sanger developed the first practical method for reading nucleotide sequences (Sanger, 1975). His initial approach employed DNA polymerase with radiolabeled nucleotides to generate sequence ladders, later refined into the rapid 'plus and minus' technique. This breakthrough was soon complemented by Maxam and Gilbert's chemical degradation method (1977), which used base-specific cleavage reactions and formed the basis for modern sequencing methods (França et al., 2002). By 2003, automated DNA sequencers enabled the sequencing of key genes and the human genome (Hood et al., 1987). A breakthrough occurred in 2005 when 454 Life Sciences, founded by biotechnology pioneer Jonathan Rothberg, introduced the world's first commercially available next-generation sequencing (NGS) platform (Patrick, 2007). This revolutionary technology transformed whole genomic research fundamentally by enabling parallel sequencing of hundreds of thousands of DNA fragments and lowering sequencing costs (Thermes, 2014b). Second-generation NGS platforms, such as Roche, Illumina, and SOLiD, use various sequencing techniques like pyrosequencing and sequencing by ligation, producing billions of bases in a single run. Third-generation NGS platforms, like Helicos' Heliscope (2007) and Pacific Biosciences' SMR, offer advantages like higher throughput, longer read lengths, faster turnaround times, and the ability to sequence minimal amounts of material, making genomic sequencing even more cost-effective and efficient (Weirather et al., 2017).

RNA-Seq overview:

The RNA-Seq workflow comprises three core stages: Experimental, Computational, and Systems Biology.

Experimental Biology: The procedure begins with extracting high-quality, intact RNA from the biological samples to ensure accurate downstream analysis. The RNA is then fragmented into smaller pieces to facilitate efficient sequencing. This is followed by reverse transcription, where RNA fragments are converted into complementary DNA (cDNA). Subsequently, the cDNA undergoes library preparation, involving adapter ligation and amplification steps, to generate sequencing-ready libraries (Ginsberg, 2005). Millions of short reads are then produced using next-generation sequencing (NGS) platforms such as Illumina and SOLiD, or long-read technologies like PacBio and Oxford Nanopore, enabling comprehensive transcriptome coverage (Luecken & Theis, 2019).

Computational Biology: After sequencing, the raw data undergoes quality control to assess read quality, remove adapters, and filter out low-quality or contaminated sequences, which is called preprocessing. The clean reads are then aligned to the reference genome or assembled de novo for further analyses (Bacher & Kendziorski, 2016). Subsequent analysis also includes pathway enrichment analysis, co-expression network construction, and transcriptome structure characterization, which provide insights into gene expression and transcript behavior.

Systems Biology: This stage derives biological insights through differential gene expression (DEG) analysis. By exploring gene expression patterns, researchers can understand cellular functions, regulatory mechanisms, and how genes respond to different conditions (Dieterich & Stadler, 2013). The analysis can be extended to identify alternative splicing events, investigate non-coding RNA regulation, and explore gene-environment interactions.

RNA Extraction:
The initial and most crucial step in an RNA-Seq experiment is the extraction of high-quality intact molecules of the desired RNA. The integrity and purity of the isolated RNA significantly impact downstream results (Han et al., 2015). RNA quality is commonly evaluated using instruments such as the Agilent BioAnalyzer, which provides an RNA Integrity Number (RIN) as a standardized measure of RNA degradation (Soverchia et al., 2005). Typically, a RIN value of 8 or higher is considered indicative of good-quality RNA suitable for reliable sequencing and accurate gene expression analysis (Tan & Yiap, 2009). However, obtaining a reliable RIN or accurately measuring RNA can be challenging in samples with limited tissue, such as brain samples. Techniques like laser capture micro-dissection (LCM) have been coupled with RNA-Seq for targeted gene expression analysis, including studies on specific neuron populations (Datta et al., 2015). For small amounts of RNA, amplification methods are employed, with selective enrichment of mRNA (Ginsberg, 2005).

RNA Sequencing:
RNA-Seq is a powerful technique for comprehensively studying all RNA molecules within a cell. Still, the workflow involves multiple intricate steps, which can potentially introduce errors or biases (Tan & Yiap, 2009). Following RNA isolation, the next critical step is cDNA synthesis, where the RNA is first fragmented into smaller pieces to improve sequencing efficiency and coverage. These RNA fragments are then reverse-transcribed into complementary DNA (cDNA). Subsequently, the single-stranded cDNA is converted into double-stranded (ds) cDNA, which forms the basis for library preparation (Rapaport et al., 2013). Adapters are ligated to the ds cDNA fragments during library prep, enabling amplification and sequencing. Careful optimization at this stage is essential to preserve the representation of transcripts and maintain data quality throughout the sequencing process. This process minimizes secondary RNA structure interference, enabling the addition of sequencing adapters after cDNA formation. Small RNAs, like miRNAs and siRNAs, can be sequenced directly, while larger RNAs cannot be sequenced directly without being fragmented into around 200–500 bp for deep sequencing (Head et al., 2014). High-quality RNA (RIN ≥ 8) is essential for reliable results, and using barcodes allows the sequencing of multiple samples together, reducing costs. Amplification during library construction can

generate identical short reads, representing genuinely abundant RNA transcripts or PCR duplicates introduced as artifacts. Distinguishing between these requires careful experimental design, including using biological replicates and computational methods to identify and remove duplicate reads, ensuring more accurate quantification of transcript abundance (Kozarewa et al., 2009). Most of the Library construction in RNA-Seq focuses on capturing applicable transcripts while reducing unwanted ones. Poly(A)-selection is commonly used to detect mRNAs, removing highly abundant transcripts (rRNA) in eukaryotes. Hybridization-based rRNA depletion methods (e.g., RiboMinus, Ribo-Zero) can instead focus on mRNA and other non-coding RNAs (Van Dijk et al., 2014). For most RNA-Seq experiments focusing on gene expression profiling, a sequencing depth of 20 to 40 million reads per sample is generally sufficient to achieve reliable and comprehensive transcriptome coverage. However, deeper sequencing is necessary for exon-level measurements, differential expression, and detecting non-coding RNAs. Sequencing depth is critical for detecting allelic expression differences and SNPs. The Illumina HiSeq platform is widely used, generating up to 3 billion reads per run. Short, single-end reads (50 bp) are suitable for simple genomes, while longer, paired-end reads (100–150 bp) are better for complex genomes and alternative splicing studies (Van Dijk et al., 2014).

RNA Reads Preprocessing:

Accurately identifying and quantifying RNA molecules from sequenced reads is a complex and critical process, making thorough quality assessment the essential first step in any RNA-Seq bioinformatics pipeline. Sequencing data are most commonly delivered in FASTQ format, where each read consists of four lines: a unique sequence identifier, the raw nucleotide sequence, a "+" separator, and a line of quality scores encoded as Phred values. These Phred scores represent the probability of incorrect base calls, with scores above 20 generally considered high-quality and suitable for reliable downstream analysis (Ji & Sadreyev, 2018). An example has been shown in (Fig. 1) on how read quality looks before and after preprocessing. Long- and short-read sequencing platforms differ in yield, accuracy, read length, and error rates. High-quality reads are crucial for accurate transcript assembly and quantification. Initial quality assessment tools like FASTX-toolkit (Gordon, 2010), FASTQC (Andrews, 2016), and NGS QC (Patel & Jain, 2012) help detect adapter contamination, low-quality reads, and overall score. However, they do not perform cleaning themselves. Trimming and filtering are essential to improve read quality. Some of the popular tools, like AdapterRemover (Schubert et al., 2016), Trimmomatic (Bolger et al., 2014), Flexbar (Dodt et al., 2012), Cutadapt (Martin, n.d.), BBMerge (Bushnell et al., 2017), and Fastp (S. Chen et al., 2018) are used for keeping high-quality reads. Although Trimmomatic and Cutadapt tools are the most widely used, they can only perform the trimming step. Regarding speed and multi-threading support, Fastp is one of the recent ultra-fast preprocessors for doing all operations (quality control, trimming, filtering) together.

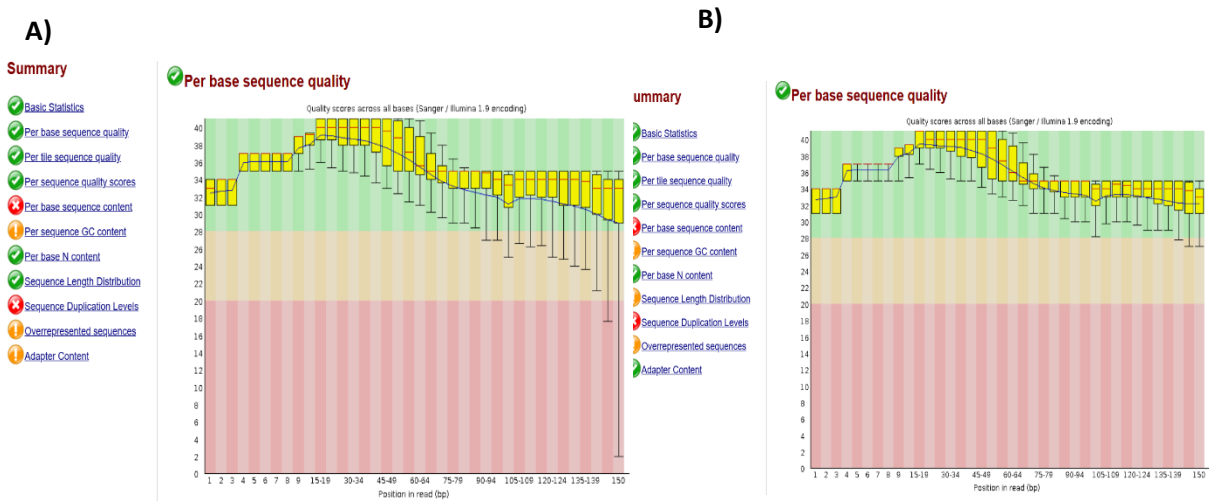

Figure 1: Figure 1 displays the raw read quality assessment from FastQC reports. Panel A shows the read quality before trimming, while Panel B illustrates the improved quality after trimming. FastQC evaluates multiple quality metrics, with Phred scores above 20 generally indicative of good-quality bases across the reads.

Alignment:

After preprocessing, the next crucial component in RNA-Seq analysis is mapping the transcripts to a reference genome or assembling them into contigs when no reference is available. Accurate mapping allows reading assignments to their genomic locations, while de novo assembly reconstructs transcript sequences from overlapping reads (Schurch et al., 2016). Both mapping and counting RNA-Seq reads are essential for quantifying gene expression, as they determine the number of reads aligned to each gene or transcript, providing the foundational data for downstream differential expression and functional analyses (Mortazavi et al., 2008). If a reference genome is available, reference-based mapping is used; otherwise, de novo assembly is used to assemble the contigs (Robertson et al., 2010). Alignment data are initially stored in SAM (Sequence Alignment/Map) files, which are plain text and can be large. For improved efficiency and faster processing, these files are converted into compressed BAM (Binary Alignment/Map) format using tools like Samtools (H. Li et al., 2009a). BAM files are more compact and enable quicker access, sorting, and indexing, which are essential for downstream analyses (H. Li et al., 2009b). Reference-guided assembly can be performed with alignment tools described in Table 1.

Table 1: Frequently Utilized RNA Alignment Software.

Aligner	Description	Reference
TopHat2	A splice-aware aligner built on Bowtie2 was designed for RNA-seq to identify exon-exon splice junctions.	(Kim et al., 2013)
STAR	Ultrafast RNA-seq aligner that handles spliced reads efficiently; excellent for large genomes and high-throughput data.	(Dobin et al., 2013)
Bowtie2	The improved version of Bowtie2 allows gap alignments and supports longer reads, but it is still not splice-aware.	(Langmead & Salzberg, 2012)
HISAT2	Fast and sensitive spliced alignment program. It has low memory usage and supports large genomes and hierarchical indexing.	(Kim et al., 2015)
BWA	Popular aligner for DNA-seq; fast and accurate for short to medium-length reads.	(H. Li & Durbin, 2009)
RUM	It combines Bowtie and BLAT to align spliced/unspliced reads; it is more comprehensive but slower.	(Grant et al., 2011)

Choosing suitable alignment tools depends on the desired study and different parameters. For example, Bowtie2 is much faster than BWA and RUM, which are usually used for mapping shorter reads between 32 and 100 base pairs. STAR is better for accurately mapping longer reads

generated from third-generation sequencing. However, improvements in TopHat2 have made it better than STAR by giving more accurate alignments while using less computing power. Moreover, it can also detect multiple splice junctions and genetic mutations in genes. Even though TopHat2 gives excellent results, tests with real data showed that HISAT2 is the fastest tool for alignment. HISAT2 uses a special indexing method that makes mapping fast and requires only a moderate amount of memory, which makes it more practical than the others.

While aligning RNA-Seq reads to a reference genome facilitates accurate quantification of known transcript expression, this approach is limited in identifying missing, novel, or species-specific transcripts absent from the reference. To overcome this limitation, de novo assembly of RNA-Seq reads reconstructs transcript sequences directly from the data without relying on a reference genome. Tools like Trinity (Haas et al., 2013) enable this assembly process by piecing overlapping reads into full-length transcripts, uncovering novel isoforms, rare transcripts, and potentially new genes. This approach is especially valuable for non-model organisms or when the reference genome is incomplete or unavailable. Other tools such as SOAPdenovoTrans (Xie et al., 2014), Oases (Schulz et al., 2012), Velvet (Zerbino & Birney, 2008), ABySS (Simpson et al., 2009), and SOAPdenovo2 (Luo et al., 2012) are also helpful for read mapping. However, one should be careful during De Novo assembly due to some of the challenges, such as isoform redundancy (multiple assembled transcripts correspond to the same gene or isoform), fragmentation (often producing fragmented contigs), and repetitive regions (repetitive sequences cause misassemblies).

After mapping, reads are summarized into expression counts, normalized (e.g., RPKM, TMM), and tested for differential expression. Read mapping can be affected by sequencing errors, repetitive sequences, and gene similarities, leading to ambiguities and multi-mapped reads. Solutions include handling multi-mapped reads proportionally or using statistical models to address mapping uncertainty. Quality control tools like RSeQC (L. Wang et al., 2012), and Qualimap2 (Okonechnikov et al., 2016) Assess mapping quality and potential biases.

Pseudoalignment:
Some RNA-seq alignment methods prioritize computational performance by using pseudo-alignment strategies. Methods like Salmon (Patro et al., 2015) Moreover, Kallisto (Bray et al., 2014) utilizes innovative quasi-mapping or pseudo-alignment approaches to rapidly quantify transcript abundance without needing complete read alignment to a reference genome. These tools significantly reduce computational time and resource usage while maintaining high accuracy. By focusing on matching k-mers from reads to transcript sequences, they efficiently estimate transcript-level expression, making them well-suited for large-scale RNA-Seq datasets. Additionally, both tools incorporate advanced bias correction techniques to account for sequence-specific, positional, and GC-content biases, further improving quantification reliability (Dillies et al., 2013). They estimate expression levels through simplified mapping, initial estimation, and refinement phases. Pseudoalignment produces normalized quantification values rather than raw read counts, which is important for choosing the appropriate method for differential expression analysis.

Isoform & Splice Junctions Detection:
Isoform expression quantification can be broadly categorized into three approaches: (1) genome-based alignment using intron-aware aligners that accurately map reads spanning exon-intron boundaries, (2) transcriptome-based alignment in which sequence reads are aligned directly to a reference transcriptome, and (3) pseudo-alignment or quasi-mapping techniques designed for high-performance quantification by rapidly associating reads with transcripts without complete alignment. Depending on the dataset and research goals, each approach offers trade-offs between accuracy, computational efficiency, and applicability. A benchmarking analysis of six popular methods—kallisto (Bray et al., 2014), Salmon (Patro et al., 2015), RSEM (B. Li & Dewey, 2011), Cufflinks (Trapnell et al., 2012), HTSeq (Anders et al., 2015), and featureCounts (Liao et al.,

2014)—was conducted using mouse liver and hippocampus tissue datasets, revealing differences in splicing complexity. Splice junctions, where exons are joined after intron removal, pose a challenge for alignment. Alternative splicing generates multiple isoforms from a single gene, significantly increasing transcriptomic complexity and functional diversity. Reference-guided methods can identify and quantify known splice junctions by aligning reads to annotated splice sites. However, they may fail to detect novel or rare splicing events not present in existing annotations. Specialized splice alignment tools minimize multi-mapping and enable reference-guided assembly of novel transcripts (Kelemen et al., 2013). Some reads may contain mismatches due to sequencing errors or mutations, and tools exist to distinguish between these and true biological variations, aiding in gene identification and fusion detection. Accurate detection of splice junctions is critical for identifying isoforms and quantifying gene expression. SpliceMap (Au et al., 2010) detects novel splice junctions without annotation, while MapSplice (K. Wang et al., 2010) and SOAPsplice (Huang et al., 2011) Offer high sensitivity and specificity for splice junction detection.

Transcript Quantification:

Counting mapped reads quantifies how many sequencing reads or transcripts align to specific genomic features, providing a fundamental dataset for downstream differential gene expression analysis. To accurately assign reads to genes or transcripts, an annotation file—typically in GFF or GTF format—is required to define gene locations and boundaries. Several widely used tools facilitate this counting process, including RSubread (Liao et al., 2019), QuasR (Gaidatzis et al., 2015), HTSeq-count (Anders et al., 2015), BEDTools (Quinlan & Hall, 2010), RSEM (B. Li & Dewey, 2011), and featureCounts (Liao et al., 2014). These tools vary in methodology and features but collectively enable accurate quantification of read counts mapped to genomic regions. Also, these can handle different alignment types, such as fully aligned, partially aligned, or junction-specific reads. Regarding speed and accuracy, the most popular tools are HTSeq and featureCounts; both accept SAM and BAM files as input and require annotation files. In contrast, RSEM does not require any reference genome for transcript quantification. However, regarding large data sets, HTSeq and RSEM fall behind and get slower than featureCounts, which is fully optimized for efficiency and speed. Although quantification is the final step for getting gene counts, these counts require normalization to account for factors like transcript length and sequencing biases. Standard normalization methods include RPKM, FPKM, and TPM (Bullard et al., 2010).

Normalization:

Data normalization is a critical stage in RNA-Seq analysis that ensures consistent and reliable inference of gene expression levels across samples. It adjusts for technical biases and systematic variations such as differences in sequencing depth, transcript length, and GC content, which can otherwise confound accurate biological signals (Ginsberg, 2005). Effective normalization allows for meaningful comparison of expression values between samples or conditions. Two key RNA seq normalization methods are:

- **TMM (Trimmed Mean of M-values)**, used in edgeR, is a normalization method that assumes most genes are not differentially expressed across samples. It adjusts library size and RNA composition differences by calculating scaling factors based on the trimmed mean of log-fold changes (M-values) between samples. This approach effectively corrects for sequencing depth biases and compositional differences, enabling a more accurate comparison of gene expression levels across diverse samples. TMM is especially useful in datasets where highly expressed genes or sample-specific biases could skew normalization. (Dillies et al., 2013).
- **DESeq2 employs geometric mean normalization, which** calculates the geometric mean of counts for each gene across all samples to serve as a reference. It then computes size factors for each sample by comparing gene counts to these geometric means. This method effectively corrects differences in sequencing depth and accounts for distributional variations, allowing for robust normalization even when samples vary

widely in library composition or RNA abundance. Geometric mean normalization helps ensure that observed expression differences reflect actual biological variation rather than technical artifacts (Zhao et al., 2021).

Although these normalization methods are helpful for bulk RNA seq, they cannot adequately address biases in single-cell RNA (scRNA) seq data. One of the most prominent features of scRNA is zero inflation or a high amount of zero read counts, so these methods often face challenges in detecting the actual biases. That is why some of the popular approaches used to reduce noise or bias in scRNA are SAMstrt (Katayama et al., 2013), GRM (Ding et al., 2015), (BASiCS) (Vallejos et al., 2017), (SCnorm) (Bacher et al., 2017), and Linnorm (Yip et al., 2017). However, BASiCS and SCnorm performed noticeably best among other tools in a comparison study of mouse embryonic datasets due to their ability to separate data variation into gene-specific constants and biological variation (Lytal et al., 2020).

Gene Expression Analysis:

Gene expression levels between different sample groups or experimental conditions are quantified and known as Differential gene expression (DGE) analysis, which helps to identify genes involved in biological processes, diseases, or treatments (Anders & Huber, 2010). For accurate DEG analysis, count data must be modeled using statistical approaches suited for discrete data, like the Poisson distribution (Rapaport et al., 2013). However, the Poisson model often fails to adequately capture biological variability, especially when multiple biological replicates are involved, resulting in an increased rate of false positives. The negative binomial distribution is commonly employed to model better the overdispersion observed in RNA-Seq data, as it accounts for variability beyond what the Poisson model can explain. Additionally, several advanced extensions, including generalized Poisson models and two-stage Poisson frameworks, have been developed to further improve the accuracy of differential expression analyses by modeling the complex variance structures inherent in biological data (Srivastava & Chen, 2010).

Gene expression differences are commonly estimated using tools like DESeq2 and edgeR, which rely on read counts mapped to genes. These counts are compiled into count matrices, where each row corresponds to a gene (or transcript) and each column represents the read counts from a specific sample. FeatureCounts (Liao et al., 2014) efficiently generates count matrices by assigning sequencing reads to genomic features. For normalization, DESeq2 (Love et al., 2014) employs the median-of-ratios method to correct for library size and composition biases, while edgeR (Robinson et al., 2009) utilizes the trimmed mean of M-values (TMM) normalization approach. Additional differential expression analysis tools include BaySeq (Hardcastle & Kelly, 2010) and EBSeq (Leng et al., 2013), which apply empirical Bayesian frameworks to improve detection accuracy by modeling biological variability and uncertainty in gene expression data. NOISeq is suited for datasets without replicates (Tarazona et al., 2015). The best tool choice depends on the dataset's characteristics. Transformation algorithms like PoissonSeq and the voom function in limma convert count data into statistical distributions for significance testing (Ritchie et al., 2015). However, biological replication is crucial for reliable analysis (Schurch et al., 2016). Method comparisons suggest that no single approach works universally, so it is recommended to document methods and consider multiple tools.

Visualization:

Identifying differentially expressed genes (DEGs) involves careful selection of significance thresholds to balance false positives and false negatives effectively. Statistical significance is commonly evaluated using raw p-values and adjusted p-values (padj) to account for multiple tests. The false discovery rate (FDR), often controlled by methods like the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) Provides a practical balance between maximizing true discoveries and minimizing false positives. In contrast, more stringent corrections, such as the

Bonferroni method, reduce the likelihood of false positives but can be overly conservative, potentially increasing false negatives by missing biologically relevant genes. Selecting appropriate cutoffs for fold change and adjusted p-values depends on the study design and biological context, ensuring more reliable identification of meaningful DEGs. The log2 fold change (l2FC) threshold helps distinguish biologically meaningful differences from random variations, with commonly used cutoffs like |l2FC| > 1 and padj < 0.05 varying based on experimental context (J. Li et al., 2012). Since no universal cutoff exists, selection should be guided by data distribution and reproducibility.

MA and volcano plots are commonly used to visualize differential gene expression (DGE) analysis results. The MA plot displays the log2 fold change (y-axis) against the average expression level or mean normalized counts (x-axis), highlighting the magnitude of expression changes relative to gene abundance (Fig. 2A). However, MA plots do not directly convey statistical significance, which limits their ability to distinguish biologically meaningful changes from noise. Consequently, MA plots are less frequently used alone for interpretation. In contrast, volcano plots combine the magnitude of change (log2 fold change) and statistical significance (usually –log10 p-value) in a single plot, enabling easier identification of significantly differentially expressed genes. Together, these visualization tools provide complementary perspectives on DGE results (L. Wang et al., 2009). Volcano plots, in contrast, plot statistical significance (-log10 p-value) against l2FC, allowing researchers to visualize both the magnitude and reliability of gene expression changes (Fig. 2B) (W. Li, 2011). Additionally, heat maps and Venn diagrams provide complementary visualizations. Heatmaps use color gradients to display expression patterns across samples (Gu et al., 2016). Venn diagrams, on the other hand, illustrate the overlap and uniqueness of DEGs across conditions, aiding in comparative analyses (H. Chen & Boutros, 2011). These graphical tools enhance the interpretability of RNA-seq results, facilitating the identification of biologically significant gene expression patterns.

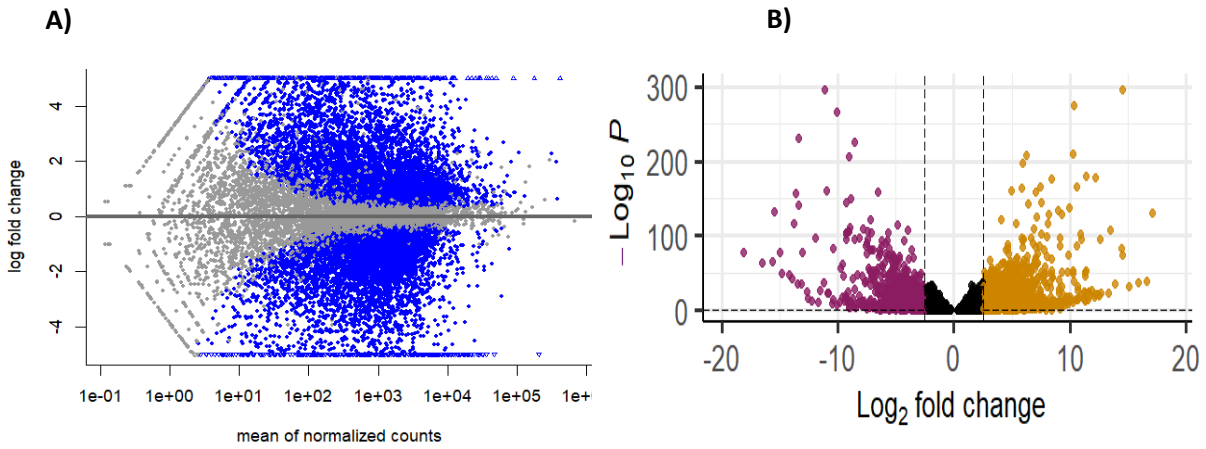

Figure 2: Figure 2 illustrates typical RNA-seq differential expression visualization methods. Panel A shows an MA plot, where each point represents a single gene, plotting mean expression counts against the $\log_2$ fold change to highlight expression differences relative to average abundance. Panel B presents a volcano plot, displaying $\log_2$ fold change versus the $-\log_{10}$ p-value, effectively combining the magnitude of change with statistical significance to identify genes of interest. These plots are fundamental for interpreting RNA-seq differential expression results by visually distinguishing biologically meaningful gene expression changes.

Conclusions:

RNA-Seq technology has significantly advanced our ability to study the transcriptome by providing a comprehensive and unbiased approach to measuring gene expression, detecting novel transcripts, and understanding transcript isoforms. The evolution of sequencing

platforms—from second-generation short-read technologies like Illumina to third-generation long-read platforms such as PacBio and Oxford Nanopore—has enhanced the depth, accuracy, and resolution of transcriptomic analyses. Despite these technological improvements, several challenges persist. These include biases introduced during RNA extraction and library preparation, sequencing errors, limitations due to incomplete or absent reference genomes, and variability from the diverse computational pipelines used for read alignment, quantification, and differential expression analysis. To overcome these issues, rigorous quality control, normalization methods, and replicates are essential. Moreover, no single computational tool suits all experimental designs, highlighting the need for careful method selection and transparent reporting of analytical steps to improve reproducibility. As RNA-Seq datasets grow in size and complexity, the ongoing development of standardized, robust, reproducible workflows will be crucial for maximizing biological insights. RNA-Seq remains a powerful and versatile technique transforming molecular biology, enabling a deeper understanding of gene regulation, disease mechanisms, and cellular function.

References:

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*. https://doi.org/10.1038/npre.2010.4282.1

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Andrews, S. (2016). *A Quality Control Tool for High Throughput Sequence Data*.

Au, K. F., Jiang, H., Lin, L., Xing, Y., & Wong, W. H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Research*, *38*(14), 4570–4578. https://doi.org/10.1093/nar/gkq211

Bacher, R., Chu, L. F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., Newton, M., & Kendziorski, C. (2017). SCnorm: Robust normalization of single-cell RNA-seq data. *Nature Methods*, *14*(6), 584–586. https://doi.org/10.1038/nmeth.4263

Bacher, R., & Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. In *Genome Biology* (Vol. 17, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-016-0927-y

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple. In *Source: Journal of the Royal Statistical Society. Series B (Methodological)* (Vol. 57, Issue 1).

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments*. http://www.biomedcentral.com/1471-2105/11/94

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *PLoS ONE*, *12*(10). https://doi.org/10.1371/journal.pone.0185056

Chen, H., & Boutros, P. C. (2011). VennDiagram: A package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-35

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Costa, V., Angelini, C., De Feis, I., & Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-Seq. In *Journal of Biomedicine and Biotechnology* (Vol. 2010). https://doi.org/10.1155/2010/853916

Datta, S., Malhotra, L., Dickerson, R., Chaffee, S., Sen, C. K., & Roy, S. (2015). Laser capture microdissection: Big data from small samples. In *Histology and Histopathology* (Vol. 30,

Issue 11, pp. 1255–1269). Histology and Histopathology. https://doi.org/10.14670/HH-11-622

Dieterich, C., & Stadler, P. F. (2013). Computational biology of RNA interactions. In *Wiley Interdisciplinary Reviews: RNA* (Vol. 4, Issue 1, pp. 107–120). https://doi.org/10.1002/wrna.1147

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, N. S., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., Le Gall, C., Schaëffer, B., Le Crom, S., Guedj, M., & Jaffrézic, F. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, *14*(6), 671–683. https://doi.org/10.1093/bib/bbs046

Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., Wildberg, A., & Wang, W. (2015). Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics*, *31*(13), 2225–2227. https://doi.org/10.1093/bioinformatics/btv122

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dodt, M., Roehr, J. T., Ahmed, R., & Dieterich, C. (2012). FLEXBAR-flexible barcode and adapter processing for next-generation sequencing platforms. *Biology*, *1*(3), 895–905. https://doi.org/10.3390/biology1030895

França, L. T. C., Carrilho, E., & Kist, T. B. L. (2002). A review of DNA sequencing techniques. In *Quarterly Reviews of Biophysics* (Vol. 35, Issue 2, pp. 169–200). https://doi.org/10.1017/S0033583502003797

Gaidatzis, D., Lerch, A., Hahne, F., & Stadler, M. B. (2015). QuasR: Quantification and annotation of short reads in R. *Bioinformatics*, *31*(7), 1130–1132. https://doi.org/10.1093/bioinformatics/btu781

Ginsberg, S. D. (2005). RNA amplification strategies for small sample populations. *Methods*, *37*(3), 229–237. https://doi.org/10.1016/j.ymeth.2005.09.003

Gordon, A. (2010). *Fastx-toolkit. FASTQ/A short-reads preprocessing tools*. ,*5*.

Grant, G. R., Farkas, M. H., Pizarro, A. D., Lahens, N. F., Schug, J., Brunk, B. P., Stoeckert, C. J., Hogenesch, J. B., & Pierce, E. A. (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, *27*(18), 2518–2528. https://doi.org/10.1093/bioinformatics/btr427

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., Macmanes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Han, Y., Gao, S., Muegge, K., Zhang, W., & Zhou, B. (2015). Advanced applications of RNA sequencing and challenges. *Bioinformatics and Biology Insights*, *9*, 29–46. https://doi.org/10.4137/BBI.S28991

Hardcastle, T. J., & Kelly, K. A. (2010). *baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data*. http://www.biomedcentral.com/1471-2105/11/422

Head, S. R., Kiyomi Komori, H., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, *56*(2), 61–77. https://doi.org/10.2144/000114133

Hood, L. E., Hunkapiller, M. W., & Smith', L. M. (1987). Automated DNA Sequencing and Analysis of the Human Genome REVIEW. In *GENOMICS* (Vol. 1).

Hrdlickova, R., Toloue, M., & Tian, B. (2017). RNA-Seq methods for transcriptome analysis. In *Wiley Interdisciplinary Reviews: RNA* (Vol. 8, Issue 1). Blackwell Publishing Ltd. https://doi.org/10.1002/wrna.1364

Huang, S., Zhang, J., Li, R., Zhang, W., He, Z., Lam, T. W., Peng, Z., & Yiu, S. M. (2011). SOAPsplice: Genome-wide ab initio detection of splice junctions from RNA-Seq data. *Frontiers in Genetics*, *2*(JULY). https://doi.org/10.3389/fgene.2011.00046

Ji, F., & Sadreyev, R. I. (2018). RNA-seq: Basic Bioinformatics Analysis. *Current Protocols in Molecular Biology*, *124*(1). https://doi.org/10.1002/cpmb.68

Katayama, S., Töhönen, V., Linnarsson, S., & Kere, J. (2013). SAMstrt: Statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics*, *29*(22), 2943–2945. https://doi.org/10.1093/bioinformatics/btt511

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., & Stamm, S. (2013). Function of alternative splicing. In *Gene* (Vol. 514, Issue 1, pp. 1–30). https://doi.org/10.1016/j.gene.2012.07.083

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4). https://doi.org/10.1186/gb-2013-14-4-r36

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., & Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods*, *6*(4), 291–295. https://doi.org/10.1038/nmeth.1311

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M. G., Haag, J. D., Gould, M. N., Stewart, R. M., & Kendziorski, C. (2013). EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, *29*(8), 1035–1043. https://doi.org/10.1093/bioinformatics/btt087

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-323

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009b). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J., Witten, D. M., Johnstone, I. M., & Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, *13*(3), 523–538. https://doi.org/10.1093/biostatistics/kxr031

Li, W. (2011). *Application of Volcano Plots in Analyses of mRNA Differential Expressions with Microarrays*. https://doi.org/10.1142/S0219720012310038

Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Liao, Y., Smyth, G. K., & Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, *47*(8). https://doi.org/10.1093/nar/gkz114

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12). https://doi.org/10.1186/s13059-014-0550-8

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, *15*(6). https://doi.org/10.15252/msb.20188746

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., … Wang, J. (2012). *SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler*. https://academic.oup.com/gigascience/article/1/1/2047-217X-1-18/2656146

Lytal, N., Ran, D., & An, L. (2020). Normalization Methods on Single-Cell RNA-seq Data: An Empirical Survey. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00041

Marguerat, S., & Bähler, J. (2010). RNA-seq: From technology to biology. In *Cellular and Molecular Life Sciences* (Vol. 67, Issue 4, pp. 569–579). https://doi.org/10.1007/s00018-009-0180-6

Martin, M. (n.d.). *TECHNICAL NOTES*. http://www-huber.embl.de/users/an-

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2016). Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*(2), 292–294. https://doi.org/10.1093/bioinformatics/btv566

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities. In *Nature Reviews Genetics* (Vol. 12, Issue 2, pp. 87–98). https://doi.org/10.1038/nrg2934

Patel, R. K., & Jain, M. (2012). NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE*, *7*(2). https://doi.org/10.1371/journal.pone.0030619

Patrick, K. L. (2007). 454 Life Sciences: Illuminating the future of genome sequencing and personalized medicine. In *YALE JOURNAL OF BIOLOGY AND MEDICINE* (Vol. 80).

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2015). Salmon *provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference*. https://doi.org/10.1101/021592

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rapaport, F., Khanin, R., Liang, Y., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., & Betel, D. (2013). *Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data*. http://genomebiology.com/2013/14/9/R95

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R., … Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, *7*(11), 909–912. https://doi.org/10.1038/nmeth.1517

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Saliba, A. E., Westermann, A. J., Gorski, S. A., & Vogel, J. (2014). Single-cell RNA-seq: Advances and future challenges. In *Nucleic Acids Research* (Vol. 42, Issue 14, pp. 8845–8860). Oxford University Press. https://doi.org/10.1093/nar/gku555

Sanger, F. (1975). Nucleotide sequences in DNA. In *Proc. B. Soc. Lond. B* (Vol. 191). https://about.jstor.org/terms

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, *9*(1). https://doi.org/10.1186/s13104-016-1900-2

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*(8), 1086–1092. https://doi.org/10.1093/bioinformatics/bts094

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., & Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, *22*(6), 839–851. https://doi.org/10.1261/rna.053959.115

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123. https://doi.org/10.1101/gr.089532.108

Solaiman, M., Tisha, S. J., & Dawood, M. (2025). ARTIFICIAL INTELLIGENCE IN THE FIELD OF PHARMACOGENOMICS. *International Journal of Engineering Technology Research & Management*, *9(4)*. https://www.ijetrm.com/IJETRM

Soverchia, L., Ubaldi, M., Leonardi-Essmann, F., Ciccocioppo, R., & Hardiman, G. (2005). Microarrays - The challenge of preparing brain tissue samples. *Addiction Biology*, *10*(1), 5–13. https://doi.org/10.1080/13556210412331327803

Srivastava, S., & Chen, L. (2010). A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research*, *38*(17). https://doi.org/10.1093/nar/gkq670

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. In *Nature Reviews Genetics* (Vol. 20, Issue 11, pp. 631–656). Nature Publishing Group. https://doi.org/10.1038/s41576-019-0150-2

Tan, S. C., & Yiap, B. C. (2009). DNA, RNA, and protein extraction: The past and the present. In *Journal of Biomedicine and Biotechnology* (Vol. 2009). https://doi.org/10.1155/2009/574398

Tarazona, S., Furió-Tarí, P., Turrà, D., Di Pietro, A., Nueda, M. J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, *43*(21). https://doi.org/10.1093/nar/gkv711

Thermes, C. (2014a). Ten years of next-generation sequencing technology. In *Trends in genetics : TIG* (Vol. 30, Issue 9, pp. 418–426). https://doi.org/10.1016/j.tig.2014.07.001

Thermes, C. (2014b). Ten years of next-generation sequencing technology. In *Trends in genetics : TIG* (Vol. 30, Issue 9, pp. 418–426). https://doi.org/10.1016/j.tig.2014.07.001

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562–578. https://doi.org/10.1038/nprot.2012.016

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., & Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: Challenges and opportunities. In *Nature Methods* (Vol. 14, Issue 6, pp. 565–571). Nature Publishing Group. https://doi.org/10.1038/nmeth.4292

Van Dijk, E. L., Jaszczyszyn, Y., & Thermes, C. (2014). Library preparation methods for next-generation sequencing: Tone down the bias. In *Experimental Cell Research* (Vol. 322, Issue 1, pp. 12–20). Academic Press Inc. https://doi.org/10.1016/j.yexcr.2014.01.008

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F., & Liu, J. (2010). MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, *38*(18). https://doi.org/10.1093/nar/gkq622

Wang, L., Feng, Z., Wang, X., Wang, X., & Zhang, X. (2009). DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, *26*(1), 136–138. https://doi.org/10.1093/bioinformatics/btp612

Wang, L., Wang, S., & Li, W. (2012). RSeQC: Quality control of RNA-seq experiments. *Bioinformatics*, *28*(16), 2184–2185. https://doi.org/10.1093/bioinformatics/bts356

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. In *Nature Reviews Genetics* (Vol. 10, Issue 1, pp. 57–63). https://doi.org/10.1038/nrg2484

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., & Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, *6*, 100. https://doi.org/10.12688/f1000research.10571.1

Wilhelm, B. T., & Landry, J. R. (2009). RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. In *Methods* (Vol. 48, Issue 3, pp. 249–257). https://doi.org/10.1016/j.ymeth.2009.03.016

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T. W., Li, Y., Xu, X., Wong, G. K. S., & Wang, J. (2014). SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, *30*(12), 1660–1666. https://doi.org/10.1093/bioinformatics/btu077

Yip, S. H., Wang, P., Kocher, J. P. A., Sham, P. C., & Wang, J. (2017). Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Research*, *45*(22), E179. https://doi.org/10.1093/NAR/GKX828

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/gr.074492.107

Zhao, Y., Li, M. C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., Williams, P. M., Evrard, Y. A., Doroshow, J. H., & McShane, L. M. (2021). TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *Journal of Translational Medicine*, *19*(1). https://doi.org/10.1186/s12967-021-02936-w

Published by

IJSAB International