Unraveling the History and Mysteries of RNA Sequencing Technology

Amitesh Chakraborty[1], Rishav Kar[2]

Department of Microbiology, Ramakrishna Mission Vivekananda Centenary College
Email: *chakrabortyamitesh429[at]gmail.com*
Mob: 8597597390

School of Biological Sciences, Ramakrishna Mission Vivekananda Educational and Research Institute (RKMVERI)
(Corresponding Author)
Mob: 9674800591

Abstract: ***RNA sequencing is a technology used extensively in transcriptomics studies. Next - generation sequencing is incorporated in the RNA - seq approach for RNA profiling to deliver unparalleled resolution for analysing and contrasting gene expression patterns. Data pre - processing is one of three main sections of the RNA - sequencing data analysis workflow. An overview of the techniques used in bulk RNA - seq and single - cell analysis is presented here, with an emphasis on alternative splicing and active RNA production study. Removal of adapters, quality checking, trimming, and filtering are essential components of data pre - processing. The data are then post - processed and subjected to a variety of analyses, such as alternative splicing, differential gene expression, and analysis of active synthesis.***

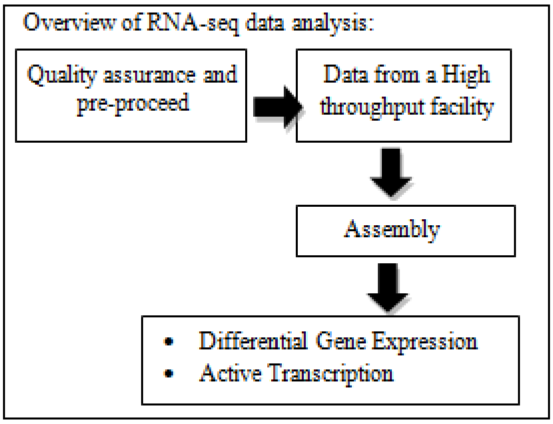
Keyword: RNA - Sequencing, Nascent mRNA analysis, Transcriptomic data analysis, Alternative splicing

1. Introduction

Watson and Crick were discovered the double helical structure of DNA in 1958 that is a new scientific branch focusing on the molecular biology of the cell (1) . All cellular functions are controlled at the molecular level by the information flow from DNA to RNA to protein (2) . We can assess how cells modify their transcriptome to respond to the surrounding environment (for instance, health and sickness) by monitoring the mRNA levels. The gene sequences of live organisms can be probed using high - throughput DNA sequencing methods, such as next - generation sequencing and the recently developed third - generation sequencing. (2) These sequencing technologies have also been modified for RNA sequencing (RNA - seq), which enables the detection and quantification of the expression of diverse RNA populations, including mRNA and total RNA.

In 1975, Sanger sequencing, the first sequencing method discovered, opened the door to studying the dynamics of genetic information (3) . Maxim and Gilbert published a unique method of DNA sequencing by chemical degradation in 1977 (4) The microarray is a technique described by Schena et al. in 1995 for measuring gene expression levels using a tiny chip. For the first time, it was possible to assess gene expression, but this method could only be applied to target genes that were already identified. The development of enormous parallel sequencing technology has dramatically decreased the cost and time needed to produce gene expression data throughout an entire transcriptome of a species. A new branch, bioinformatics was created duo to scale and power requirements of the massively parallel sequencing datasets.

Here, a summary of the background and mystique surrounding RNA is given.


Overview of RNA-seq data analysis:
Quality assurance and pre-proceed
Data from a High throughput facility
Assembly
Differential Gene Expression
Active Transcription

Transcriptome Sequencing: Transcriptomic analysis has been transformed by the development of high - throughput next - generation sequencing (NGS) (5) . This technological advancement removed several obstacles that Sanger sequencing - based methods and hybridization - based microarrays, which were previously employed to measure gene expression. To determine the nucleotide sequence of RNA molecules as well as the amounts of RNA species within populations of RNA molecules, RNA - seq uses high - throughput sequencing of nucleic acids. Specialized mathematical methods are required for RNA - seq analysis. Numerous scientific breakthroughs have resulted from computational analysis of RNA - seq data, including the identification of biomarkers and pathogenic mutations as well as the development of novel therapeutics and a thorough understanding of genetic regulatory areas.

First, the desired biological material (such as cells or tissues) is used to extract RNA. In the second step, particular protocols are used to separate subsets of RNA molecules, for as the poly - A selection technique to enrich for polyadenylated transcripts or the ribo - depletion protocol to eliminate ribosomal RNAs. Reverse transcription is used to turn the RNA into complementary DNA (cDNA), and the ends of the cDNA fragments are then joined using sequencing adaptors. The RNA - Seq library is ready for sequencing after PCR amplification. (6)

Describe of different Sequencing Technologies:

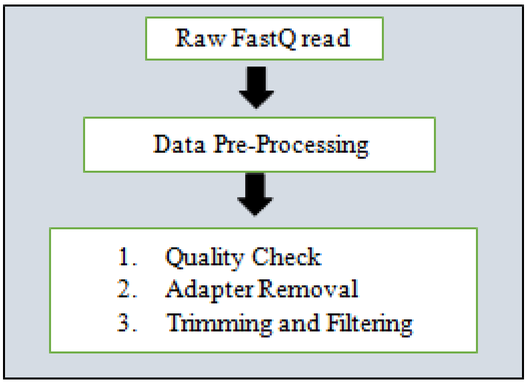
Standard RNA sequencing
Description: Evaluates the RNA concentrations in a biological sample at any given time.
Advantages: Allows it possible to analyse alternative splicing. (2)
Limitation: Sufficient initial material is required (for standard library preparation).

Nascent RNA Sequencing
Description: Techniques based on nucleotide analogues are used to evaluate RNA production. The nucleotide analogue is incorporated into the nascent RNAs, which are then either decoded computationally or enriched affinity - based methods.
Advantages: Permits analysis of mRNA degradation and half - lives and identifies DNA replication origins by sequencing nascent DNA strands
Limitation: More time - consuming and laborious than RNA - seq

Single - cell RNA sequencing
Description: These techniques are unquestionably important for revealing rhythmic and diverse gene expression levels in different types of single cells and for locating cell - specific biomarkers. (7)
Advantages: Allows for the evaluation of sample heterogeneity.
Limitation: More time - consuming and laborious than RNA - seq.

Data Processing: The sequencing facility distributes the sequencing data in the FastQ (8) format. Every read is described by four lines in this format, which is a modified version of the standard fasta - format. The first line starts with "[at]" and the sequence identifier, the second line contains the raw sequence, the third line is a "+, " and the fourth and final line contains the quality values corresponding to the raw sequence, or the "phred" score. Quality checking, adapter removal, cutting, and filtering are all steps in the pre - processing of the raw FastQ file. (2)
One of the most important steps in the workflow for analysing sequencing data is to examine the read quality in the raw FastQ files. The phred score, which is given to each nucleotide within the read, is used to assess the read quality. The quality of the data increases with the phred score. How much data needs to be refined depends on the findings of the quality control.

Long - read sequencing increases the precision of assembly or may perhaps be able to do away with assembly altogether. Compared to long - read technologies, short - read technologies can produce data with a reduced error rate and higher throughput. (6)



1) Quality Check: For more accuracy, the FastQ file quality must be reviewed. The adapter contamination on either side of the read or technological problems by sequencers can cause low - quality reads in the FastQ files. Severe problems, such as increased background and the identification of incorrect alternative splicing events, might result from failing to filter the low - quality reads from the RNA - seq data. If the reads have varied lengths, some of the reads may mistakenly map to the intronic regions in the context of splicing analysis, leading to cause splicing fault. You can get FastQC, the most popular Java - based program for assessing the quality of FastQ files, for free at https: //www.bioinformatics. babraham. ac. uk/projects/fastqc/. (2) An acceptable Phred score is 20 or higher. If the score is less than 20, additional pre - processing of the data is necessary.

2) Adapter Removal: Short oligonucleotides are called adapters. Before analysis, these short oligonucleotides (adapters) must be removed because they will prevent alignment to the reference genome. However, Adapter Remover cannot be used to simultaneously trim several adapters. It can remove adapters from 50 - and 30 - end, single - and paired - end data. On the other hand, CutAdapt is unable to handle paired - end data but can manage several adapters at once. While Trimmomatic can handle multiple adapters and trim adapters from the 30 - end of both single - and paired - end data, it is unable to remove adapters from the 50 - end. In a nutshell, the choice of tool relies on the demands of the particular task. The widely used tools for adapter trimming are: Adapter Remover https: //adapterremoval. readthedocs. io.; (9) CutAdapt https: //cutadapt. readthedocs. io/en/stable (10) and trimmomatic http: //www.usadellab. org/cms/?page=trimmomatic. (11)

3) Trimming and Filtering: Trimming and filtering are used to remove low - quality, unintelligible reads as well as the reads that contribute to read length variance. Trimming with programs like Trimmomatic helps to keep the sequencing depth while removing the low - quality reads if there are only a few nucleotides that fall below the quality barrier. (11) Software like fastp, however, is more suited when full reads must be filtered out because it enables read filtering based on quality and length thresholds. (12) Fastp can be accessed freely from here: https: //github. com/OpenGene/fastp.

Data Analysis: An unattainable high - resolution image of the world's transcriptional landscape is now possible because to the use of RNA - Seq to profile gene expression**.** Numerous new informatics problems and applications arise as a result of the ongoing development of sequencing technology and procedure approaches. RNA - seq applications including single - cell RNA - seq (scRNA - seq) data, alternative splicing, fusion transcripts, and nascent mRNA synthesis are all described, as well as analytic choices for these applications. (8) RNA - Seq can be used to find novel gene structures, alternatively spliced isoforms, and allele - specific expression (ASE) in addition to monitoring gene expression levels. (5) Furthermore, genetic investigations of gene expression have discovered genetically associated variation in expression, splicing, and ASE. The main procedures in the analysis of a typical RNA - seq experiment are covered in this section.

Workflow: A typical process for RNA - Seq data includes producing FASTQ - format files with reads that were sequenced using an NGS technology, comparing these reads to an annotated reference genome, and determining gene expression. (5) Despite the fact that, basic sequencing analysis tools are more available than ever, RNA - Seq analysis presents unique computational challenges not present in other sequencing - based research and needs specific consideration to the biases inherent in expression data.

These steps include read alignment with and without a reference genome, getting metrics for gene and transcript expression, and strategies for identifying differential gene expression, alternative splicing, nascent mRNA synthesis, and single - cell RNA - seq (scRNA - seq) data.

4) Read Alignment: Because many RNA - Seq reads map across splice junctions, mapping RNA - Seq reads to the genome is significantly more difficult than mapping DNA sequencing reads. Due to their incapacity to handle spliced transcripts, traditional read mapping algorithms like Bowtie (13) and BWA (14) are not advised for mapping RNA - Seq reads to the reference genome. The reference genome can be expanded with sequences produced from exon - exon splice junctions obtained from existing gene annotations as one method of fixing this issue. The initial stages of RNA - seq data processing involve creating the putative transcriptome and mapping the reads from the raw FastQ file. When the reference genome for the organism of interest is available, reference - based mapping is used. Without access to the reference genome, de novo assembly is used. The short reads are combined in this instance to create the contig, which is then used as "a hypothetical genome" to which the identical reads are re - mapped. The Sequence Alignment Map (SAM) (15) file contains the alignment information. To conserve storage space and hasten further processing, the alignment files are converted to their binary form, or BAM. SAM and BAM files can be read and edited with Samtools. The alignment algorithm is chosen in accordance with the accessibility of the reference genome.

GSNAP (16), MapSplice (17) , RUM (18) , STAR (19) and TopHat (20) are some of the most popular RNA - Seq alignment programs, while Trinity is a reliable approach for reference - free mapping. Performance, speed, and memory use benefits vary depending on the aligner. Based on these parameters and the general goals of the RNA - Seq investigation, the best aligner should be chosen (for instance, STAR creats big temp files while Bowtie does not). A particular alignment tool may be needed for downstream analysis, usually when more involved data analysis is of interest. The RNA - Seq Genome Annotation Assessment Project 3 (RGASP3) of GENCODE has started efforts to systematically assess the performance of RNA - Seq aligners (5). RGASP3 has discovered significant performance differences between alignment tools on several benchmarks, including alignment yield, basewise accuracy, mismatch and gap placement, and exon junction discovery.

Short read mapping Software

Tools	Link	Procedure	Ref.
1. Bowtie	http: //bowtie. cbcb. umd. edu.	BWT - based	13)
2. Maq	http: //maq. sourceforge. net/	Hash - based	21)
3. SeqMap	http: //biogibbs. stanford. edu/Bjiangh/SeqMap/	Hash - based	22)
4. BWA	http: //bio - bwa. sourceforge. net/	BWT - based	14)
5. RMAP	http: //rulai. cshl. edu/rmap/	Hash - based	23)
6. SOAP	http: //soap. genomics. org. cn/soap1/	Hash - based	24)
7. ZOOM	http: //www.bioinfor. com	Hash - based	25)
8. SHRiMP	http: //compbio. cs. toronto. edu/shrimp/	Hash - based	(26)
9. SSAHA2	http: //www.sanger. ac. uk/resources/software/ssaha2/	Hash - based	(27)

5) Assembly and Quantification: The mapped reads from RNA - Seq can be put together into transcripts after the reads have been aligned. Most computational tools use the accumulation of read alignments to the reference genome to infer transcript models. De novo reconstruction, which assembles contiguous transcript sequences using a reference genome or annotations, is an alternative method for assembling transcripts. (28–30) Estimating gene expression levels is a frequent downstream feature of transcript reconstruction software. The number of reads that correspond to full - length transcripts is used to measure expression by computational tools like Cufflinks (31), FluxCapacitor (32), and MISO (33).

Once the actual transcript quantification values have been determined, it is important to examine them for biases in GC content and gene length so that corrective normalizing techniques can be used as needed. Researchers could examine the biotype composition of the sample, which is a sign of how successfully the RNA purification process was done, assuming the reference transcriptome is comprehensively annotated.

The paired fragments per kilobase of transcript per million mapped reads (FPKM) metric for paired end - reads normalizes for sources of variance in transcript quantification and takes into consideration the dependence between paired - end reads in the RPKM estimate (31) . The mapping of reads to multiple transcripts because of genes with several isoforms or close paralogs presents another technical hurdle for transcript quantification. By excluding all reads that do not map uniquely, as in Alexa - Seq (34) , one way to address this "read assignment uncertainty" is to use Alexa - Seq. This approach is far from perfect for genes without distinctive exons, though. The mapping of reads to various transcripts resulting from genes with numerous isoforms or near paralogs presents another technical barrier for transcript quantification. To exclude all reads that do not map uniquely, as in Alexa - Seq, is one way to address this "read assignment uncertainty. " For genes without distinctive exons, this approach is far from ideal. Building a likelihood function that simulates the sequencing experiment and calculates the highest likelihood that a read maps to a specific isoform is an alternative approach utilized by Cufflinks and MISO.

Gene Expression Quantification Software

Tools	Link	Ref.
ALEXA - Seq	http: //www.alexaplatform. org/alexa_seq/index. htm	(34)
Scripture	http: //www.broadinstitute. org/software/scripture/?q=home	(35)
IsoInfer	http: //www.cs. ucr. edu/~jianxing/IsoInfer. html	(36)
MMSEQ	http: //bgx. org. uk/software/mmseq. html	(37)
MISO	http: //genes. mit. edu/burgelab/miso/	(33)
rSeq	http: //www - personal. umich. edu/~jianghui/rseq/	(38)
Cufflinks	http: //cufflinks. cbcb. umd. edu/	(31)

6) Differential Gene Expression Analysis: Studying the differences in gene expression levels between two or more conditions is one of the most frequently used RNA - seq applications. To conduct a differential expression study, gene expression values must be compared between samples. When choosing the appropriate research technology, this is a crucial element to consider because multi - exon genes can encode many functional isoforms. RNA - Seq will unavoidably and eventually replace microarrays, even though it is still significantly more expensive to sequence many samples than with microarrays. The amount of reads that are mapped onto a gene or transcript determines its expression level for RNA - Seq, whereas fluorescence levels that are retrieved after hybridization for microarrays reflect this. However, several RNA - Seq biases, including sequencing depth, sample count distribution, and the length of genes or transcripts, should be considered when doing differential expression analysis. Normally, counts will increase with increasing sequencing depth. In the meantime, there may be variations in the count distribution across samples. The most straightforward normalization technique is reads per kilobase per million reads mapped (RPKM). RPKM can be applied to single - and paired - end sequencing and corrects for variations in gene length as well as library size (39). Comparable to RPKM, FPKM (fragments per kilobase of transcript per million mapped reads) is utilized for paired - end sequencing data (40).

When the samples have different levels of sequencing depth, normalization is especially crucial. Prior to normalization, the samples can be clustered (principal component analysis), and the number of reads can be used to display this variance. Analysis of differential gene expression can be done following normalization. When identifying the genes or isoforms that are differentially expressed, these RNA - Seq biases should be considered. More and more methods are being developed to use RNA - Seq data to find those tags that are differentially expressed among the analyzed gene or transcript sets under various situations. Those approaches can be separated into two groups based on whether parametric models are used. The most widely used parametric techniques are Binomial, Poisson, and Negative Binomial (41). The negative binomial (NB) is used as the reference distribution by DESeq2, edgeR, and baySeq, each of which has its own normalization strategy. EBSeq is based on empirical Bayesian analysis, much like baySeq, but uses median normalization instead. Non - parametric methods, in contrast, make no assumptions about the distribution of the data. A potent non - parametric technique called NOISeq has recently been developed. It can withstand changes in sequencing depth and predicts the noise distribution from actual data. Only datasets without replicates are used with NOISeq. SAMseq uses an outlier - tolerant nonparametric statistical test. These are more adaptable to variations in sequencing depth than most previous parametric techniques (baySeq, DESeq, edgeR). DESeq, edgeR, and baySeq all use the Negative Binomial (NB) distribution; whereas NOISeq does not, these methods reveal a substantial dependence on sequencing depth. The estimation of the negative binomial distribution can be noisy in small - scale investigations that contrast two samples with no or few repeats. Simpler techniques based on the Poisson distribution, like DEGseq, or on empirical distributions, like NOISeq, may be an alternative in such circumstances. Studies with little sample size employ shrinkseq. The choice of method (or even the version of a software package) can significantly alter the analysis's results, and no one method is likely to perform favorably for all datasets, according to recent independent comparative studies.

Tools for differential expression analysis

Tool	Link	Ref.
EdgeR	http: //www.bioconductor. org/packages/release/bioc/html/edgeR. html	(42)
NOISeq	http: //bioinfo. cipf. es/noiseq/doku. php?id=start	(43)
baySeq	http: //www.bioconductor. org/packages/2.8/bioc/html/baySeq. html	(44)
ASC	http: //www.stat. brown. edu/Zwu/research. aspx	(45)
DEGseq	http: //bioinfo. au. tsinghua. edu. cn/software/degseq/	(46)
DESeq	http: //www - huber. embl. de/users/anders/DESeq/	(43)
Myrna	http: //bowtie - bio. sourceforge. net/myrna/index. shtml	(13)
GENE - Counter	http: //changlab. cgrb. oregonstate. edu/?q=node/view/527	(47)
Cuffdiff	http: //cufflinks. cbcb. umd. edu/	(31)
GPSeq	http: //www - rcf. usc. edu/~liangche/software. html	(48)

7) Downstream Analysis of the DEGs: The DEGs are ranked in order of their significance (p - value) and the log2 fold change; thresholds of these parameters are subjective to the study. A few of the most popular visualization techniques are provided, together with the R - scripts that produce them. The three basic visual representations of DEGs are the MA plot, the volcano plot, and the heatmap presentation. The MA plot, which stands for log ratio (M) and mean average (A), is an easy way to see the DEGs. On the x - axis, it shows the number of reads, and on the y - axis, it shows the log2 fold change. This graph, which is not usually used in papers, does not show the statistical significance of the DEGs. Volcano plots show the importance and levels of gene expression for each examined gene. The heatmap presentation is yet another way that is widely used to display RNA - seq data. The DEGs feature being highlighted determines the type of data presentation that is chosen. Easy - to - use platforms for pathway and gene ontology (GO) enrichment are offered via online web servers like Database for Annotation, Visualization and Integrated Discovery (DAVID) and Enrichr. The DEGs are grouped into clusters using the Cluego - plugin, which can be accessed using Cytoscape, based on the enriched GO terms for the categories. Since gene set enrichment analysis (GSEA) is an effective analytical technique to cluster and enrich the GO keywords for the DEGs, it has grown in popularity in recent years. Both an open - source Windows application and an R - package are available for GSEA. The Python library GeneWalk, which uses representation learning to find regulator and moonlighting genes, is a relatively recent addition to the route enrichment tools. The tools mentioned above use different backend databases and statistical tests to determine the significance. For cross - validation purposes and to gain a deeper knowledge of the impacted biological processes, several enrichment algorithms can be used to the same dataset. The enriched GO keywords will be more certain as a result, although experimental confirmation is still advised despite the computational cross - validation.

8) Transcriptome reconstruction: The entire amount of RNAs produced by a single cell or a population of cells, including different protein - coding and non - coding RNAs, is known as the transcriptome (49) . RNA - Seq is a realistic and useful option to collect an organism's whole transcriptome (50) . There are primarily two groups of methodologies for reconstructing the transcriptome at the moment (51) . The 'genome - guided' strategy is the first, and it involves mapping all the transcriptome sequencing reads to the reference genome first. The aligned reads are then put together into transcripts or fragments in accordance with the read mapping data. This genome - guided methodology is used in programs like Cufflinks (31) and Scripture (35) . Both Cufflinks and Scripture directly reconstruct the transcriptome from the spliced reads, and both have comparable computational needs. Scripture's method is based on maximum sensitivity, whereas Cufflinks' method is based on maximum accuracy (52) . The "genome - independent" approach, which does not require a reference genome and directly assembles the reads into transcripts, is another method for reconstructing the transcriptome. On this genome - independent methodology, programs like Velvet (53) , Trans - ABySS (30) , Trinity (52) , and Oases (54) are based. It's fascinating to notice that Velvet is capable of building the genome and transcriptome from scratch. The de Bruijn graphs are mostly used by the de novo assembly software to model the read - derived overlapping k - mer subsequences. The de Bruijn graph is then parsed using several methods, and the reads are then put together into contigs or scaffolds (55).

In general, genome - guided approaches are more suited for species with high - quality assembled reference genomes, whereas genome - independent methods can be applied to any species, regardless of whether they have reference sequences readily available.

Tools for Transcriptome reconstruction

Tools	Link	Category	Ref.
Trinity	http: //trinityrnaseq. sourceforge. net/	Genome - independent	(30)
Oases	http: //www.ebi. ac. uk/~zerbino/oases/	Genome - independent	(6)
Trans - ABySS	http: //www.bcgsc. ca/platform/bioinfo/software/trans - abyss	Genome - independent	(29)
Velvet	http: //www.ebi. ac. uk/~zerbino/velvet/	Genome - independent	(53)
Cufflinks	http: //cufflinks. cbcb. umd. edu/	Genome - guide	(31)
Scripture	http: //www.broadinstitute. org/software/scripture/?q=home	Genome - guide	(35)

Allele - Specific Expression: The sequenced transcript reads can provide coverage across heterozygous regions since they represent transcription from both the maternal and paternal alleles. The null hypothesis is that the ratio of maternal to paternal alleles is balanced if enough reads cover a heterozygous location within a gene (56) . Allele - specific expression (ASE) may be present if there is a significant departure from this expectation. Genetic variation (such as single - nucleotide polymorphism in a cis - regulatory area upstream of a gene) and epigenetic changes (such as genomic imprinting, methylation, histone modifications, etc.) are two potential mechanisms for ASE (57) . A statistical test, such as the binomial test or the Fisher's exact test, is applied after counting reads that include each allele at heterozygous sites in conventional processes to detect ASE (58–60).

To determine the cis - regulatory effects of genetic variations, RNA - seq can quantify allele - specific expression (ASE or allelic expression). ASE stands for independent measurements of gene expression for a gene's paternal and maternal alleles (61) . Only genes with a heterozygous single - nucleotide polymorphism (SNP) inside the transcribed region can have ASE evaluated in a normal RNA - seq experiment (62) . This SNP, also known as the aseSNP, can be used as a tag to distinguish reads that come from every gene copy. Additionally, ASE data can be utilized to map the causal regulatory variations in eQTL data and to increase statistical power for eQTL identification (63). Additionally, ASE data are naturally noise - resistant, making it possible to uncover the impacts of gene - by - environment interactions or the impact of uncommon genetic variations on gene expression to increase the accuracy of diagnosing Mendelian disorders.

Nascent RNA Sequencing Technologies: The active mRNA production is measured by emerging RNA sequencing technology, not the overall amount of mRNA. Every sample in conventional RNA - seq has the exact same quantity of RNA that is sequenced. This implies that the relative quantity of a particular mRNA is assessed whether the gene is actively being transcribed. It is possible to measure the actively transcribing genes using biosynthetic metabolic markers. Several techniques have been created for this purpose. Metabolically label RNA using 2, 4 - dithiouracil, which was published by Cleary et al. in 2005, is one of the pioneers of this strategy (64) . This approach calls for biotinylating the labelled RNA, enrichment using

Paper ID: SR24109000554 DOI: https://dx.doi.org/10.21275/SR24109000554

streptavidin - coated magnetic beads, and microarray analysis. Precision nuclear run - on sequencing (PRO - seq) is an adaption of GRO - seq in which biotinylated nucleotide triphosphates (NTPs) are used in place of the Br - UTP (65) . Biotin - NTP incorporation prevents transcription, and 30 end sequencing identifies the precise position of the RNA polymerase active site that is interacting with the developing RNA. Transient transcriptome sequencing (TT - seq), which was invented by Schwalb et al. in 2016, began with 4 - thiouridine (4sU) (66) The nucleotide analogue 4sU is used in TT - seq to mark the developing mRNA for just five minutes. Only the actively transcribed genes are sequenced because the transcripts are first broken up and enriched using streptavidin - coated magnetic beads. Thiol (SH) - linked alkylation for metabolic labelling of RNA (SLAM - seq) was introduced by Herzog et al. in 2017 (67) . The chemistry - based RNA - seq method known as SLAM - seq finds the inclusion of 4sU at the single nucleotide level. In essence, cells that incorporate the label to the actively transcribing mRNAs receive the addition of 4sU.

Technique	Description
PRO - seq (precision nuclear run - on sequencing) (65)	The incorporation of biotinylated NTPs into the developing mRNA prevents transcription. The specific position of the stopped RNA polymerase is identified by 30 end sequencing.
GRO - seq (global run - on sequencing) (68)	Nucleotides that have been tagged (Br - UTP) are integrated into the RNA. Following hydrolysis, the RNA is subsequently purified with antibody - coated beads.
TT - seq (Transient transcriptome sequencing) (66)	RNA is isolated, fragmented, biotinylated, purified, and sequenced after being tagged with 4 - thiouridine.
SLAM - seq (Thiol (SH) - linked alkylation for the metabolic sequencing of RNA) (67)	Alkylation after 4 - thiouridine labelling makes it possible to identify the nucleotide analogue as a cytosine. It is assessed how much of the T > C conversion occurs during early mRNA production.
Analysis of 2, 4 - dithiouracil labelled and enriched RNA using a microarray. (64)	Using streptavidin beads, the biotinylated and enriched RNA is labelled. A microarray is used for the analysis of the extracted RNA.

Analysis of Alternative Splicing: There have been suggested specific RNA - seq analysis techniques that focus on alternative splicing. To produce mRNA suitable for translating into proteins, mRNA must be subjected to splicing. In the process of alternative splicing, a cell produces different protein isoforms from the same gene following translation. There are five main types of alternative splicing: exon skipping (SE), retained introns (RI), mutually exclusive exons (MXE), alternative 5' splice sites (A5SS), and alternative 3' splice sites (A3SS) (69) . Data on differential gene expression can be included into upstream processing through the analysis of DEGs and alternative splicing. The three main programs utilized to identify the differentially spliced sites from the RNA - seq data are MISO (70) , rMATS (71) , and SUPPA (72) .

Programs	Description
1. Mixture of isoforms (MISO)	One of the first tools for alternative splicing analysis; developed in 2010. Using the statistical model, the expression of the alternatively spliced gene is provided along with an estimation of confidence.
2. Replicate Multivariate Analysis of Transcript splicing (rMATS)	Developed in 2014; a statistical technique for comparing two RNA - Seq data' alternative splicing patterns.
3. SUPPA	The other two algorithms are 1000 times slower than SUPPA, which is based on transcript abundance. Even though SUPPA decreases time.

2. Future Prospects

Numerous RNA - Seq applications exist, and for each application, there are typically a variety of software options from which to choose. Computational tools will need to improve synchronously with sequencing technology to address new technical issues and facilitate innovative applications. Selecting the best parameters for the software and choosing appropriate software to conduct related investigations are both crucial decisions that have a direct impact on the outcomes. We can get better findings and accomplish our study aims with the help of suitable software and excellent parameter configuration. Additionally, the algorithms used by distinct pieces of software for the same application have varied design variances and would each have a different advantage over the same dataset. To effectively provide better outcomes, it is thus vital to test the software and various factors before making a final choice. The limitations of sequencing technologies still exist, despite the rapid development of the algorithms for numerous applications to satisfy the needs of research. The sample preparation step of the sequencing procedure could introduce contaminants, and the library creation step could lose sources and miss some targets. These ambiguities may amplify data noise and result in incomplete data. Additionally, sequencing methods have bias in their sequencing, and bioinformatics techniques have their own limits. These factors might make it more challenging to analyse the data and provide unfavourable outcomes. With little doubt, advances in sequencing technologies and associated analysis methods will significantly improve data interpretations and make it easier for us to understand the transcriptomes of different species. Future developments in sequencing costs and the creation of more potent algorithms will make it possible for scientists to examine a variety of transcriptomes more quickly and thoroughly from various organisms. Additionally, these modifications will give us excellent chances to research the roles of noncoding RNAs (both short and long), which were previously thought to be transcriptional noise but may really have unidentified purposes. These contingent research findings will increase our knowledge of the transcriptome and may even challenge our preconceived notions about it as study into other transcriptomes progresses. The myriad associated investigations that will be made possible by these new discoveries will undoubtedly advance our understanding of life.

References

[1] Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.

[2] Gondane A, Itkonen HM. Revealing the History and Mystery of RNA - Seq. Vol.45, Current Issues in Molecular Biology. MDPI; 2023. P.1860–74.

[3] Sander F, Goulson AR. A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase. Vol.94, J. Mol. Bid.1976.

[4] Maxam AM, Gilbert W. A new method for sequencing DNA (DNA chenistry/dimethyl sulfate cleavage/hydrazine/piperidine). Vol.74, Biochemistry.1977.

[5] Kukurba KR, Montgomery SB. RNA sequencing and analysis. Cold Spring Harb Protoc.2015 Nov 1; 2015 (11): 951–69.

[6] Chen G, Wang C, Shi TL. Overview of available methods for diverse RNA - Seq data analyses. Vol.54, Science China Life Sciences.2011. P.1121–8.

[7] Deshpande D, Chhugani K, Chang Y, Karlsberg A, Loeffler C, Zhang J, et al. RNA - seq data science: From raw data to effective interpretation. Vol.14, Frontiers in Genetics. Frontiers Media S. A.; 2023.

[8] Conesa A, Madrigal P, Tarazona S, Gomez - Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA - seq data analysis. Vol.17, Genome Biology. BioMed Central Ltd.; 2016.

[9] Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. BMC Res Notes.2016 Feb 12; 9 (1).

[10] Cutadapt removes adapter sequences from high - throughput sequencing reads.

[11] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics.2014 Aug 1; 30 (15): 2114–20.

[12] Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra - fast all - in - one FASTQ preprocessor. In: Bioinformatics. Oxford University Press; 2018. P. i884–90.

[13] Langmead B, Hansen KD, Leek JT. Cloud - scale RNA - sequencing differential expression analysis with Myrna [Internet]. Vol.11, Genome Biology.2010. Available from: http: //genomebiology. com/content/11/8/R83

[14] Chong CF, Li YC, Wang TL, Chang H. Stratification of Adverse Outcomes by Preoperative Risk Factors in Coronary Artery Bypass Graft Patients: An Artificial Neural Network Prediction Model.

[15] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics.2009 Aug; 25 (16): 2078–9.

[16] Wu TD, Nacu S. Fast and SNP - tolerant detection of complex variants and splicing in short reads. Bioinformatics.2010 Feb 10; 26 (7): 873–81.

[17] Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: Accurate mapping of RNA - seq reads for splice junction discovery. Nucleic Acids Res.2010 Aug 28; 38 (18).

[18] Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, et al. Comparative analysis of RNA - Seq alignment algorithms and the RNA - Seq unified mapper (RUM). Bioinformatics.2011 Sep; 27 (18): 2518–28.

[19] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA - seq aligner. Bioinformatics.2013 Jan; 29 (1): 15–21.

[20] Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA - Seq. Bioinformatics.2009; 25 (9): 1105–11.

[21] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res.2008 Nov; 18 (11): 1851–8.

[22] Jiang H, Wong WH. SeqMap: Mapping massive amount of oligonucleotides to the genome. Bioinformatics.2008 Oct; 24 (20): 2395–6.

[23] Smith AD, Xuan Z, Zhang MQ. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics.2008 Feb 28; 9.

[24] Li R, Li Y, Kristiansen K, Wang J. SOAP: Short oligonucleotide alignment program. Bioinformatics.2008 Mar; 24 (5): 713–4.

[25] Lin H, Zhang Z, Zhang MQ, Ma B, Li M. ZOOM! Zillions of oligos mapped. Bioinformatics.2008; 24 (21): 2431–7.

[26] Rumble SM, Lacroute P, Dalca A V., Fiume M, Sidow A, Brudno M. SHRiMP: Accurate mapping of short color - space reads. PLoS Comput Biol.2009; 5 (5).

[27] Ning Z, Cox AJ, Mullikin JC. SSAHA: A fast search method for large DNA databases. Genome Res.2001; 11 (10): 1725–9.

[28] Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: Robust de novo RNA - seq assembly across the dynamic range of expression levels. Bioinformatics.2012 Apr; 28 (8): 1086–92.

[29] Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA - seq data. Nat Methods.2010 Nov; 7 (11): 909–12.

[30] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full - length transcriptome assembly from RNA - Seq data without a reference genome. Nat Biotechnol.2011 Jul; 29 (7): 644–52.

[31] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript assembly and quantification by RNA - Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol.2010 May; 28 (5): 511–5.

[32] Montgomery SB, Sammeth M, Gutierrez - Arcelus M, Lach RP, Ingle C, Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature.2010 Apr 1; 464 (7289): 773–7.

[33] Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods.2010 Dec; 7 (12): 1009–15.

[34] Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, et al. Alternative expression analysis by RNA sequencing. Nat Methods.2010 Oct; 7 (10): 843–7.

[35] Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, et al. Ab initio reconstruction of cell type - specific transcriptomes in mouse reveals the conserved multi - exonic structure of lincRNAs. Nat Biotechnol.2010 May; 28 (5): 503–10.

[36] Feng J, Li W, Jiang T. Inference of isoforms from short sequence reads. In: Journal of Computational Biology.2011. P.305–21.

[37] Turro E, Su SY, Gonçalves Â, Coin LJM, Richardson S, Lewin A. Haplotype and isoform specific expression estimation using multi - mapping RNA - seq reads. Genome Biol.2011 Feb 10; 12 (2).

[38] Chen G, Yin K, Shi L, Fang Y, Qi Y, Li P, et al. Comparative analysis of human protein - coding and noncoding rnas between brain and 10 mixed cell lines by RNA - seq. PLoS One.2011 Nov 30; 6 (11).

[39] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA - Seq. Nat Methods.2008 Jul; 5 (7): 621–8.

[40] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript assembly and quantification by RNA - Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol.2010 May; 28 (5): 511–5.

[41] Tarazona S, García - Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA - seq: A matter of depth. Genome Res.2011 Dec; 21 (12): 2213–23.

[42] Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics.2009 Nov 11; 26 (1): 139–40.

[43] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol.2010 Oct 27; 11 (10).

[44] Hardcastle TJ, Kelly KA. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data [Internet].2010. Available from: http: //www.biomedcentral. com/1471 - 2105/11/422

[45] Wu Z, Jenkins BD, Rynearson TA, Dyhrman ST, Saito MA, Mercier M, et al. Empirical bayes analysis of sequencing - based transcriptional profiling without replicates. BMC Bioinformatics.2010 Nov 16; 11.

[46] Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: An R package for identifying differentially expressed genes from RNA - seq data. Bioinformatics.2009 Oct 24; 26 (1): 136–8.

[47] Cumbie JS, Kimbrel JA, Di Y, Schafer DW, Wilhelm LJ, Fox SE, et al. GENE - counter: A computational pipeline for the analysis of RNA - seq data for gene expression differences. PLoS One.2011 Oct 6; 6 (10).

[48] Srivastava S, Chen L. A two - parameter generalized Poisson model to improve the analysis of RNA - seq data. Nucleic Acids Res.2010 Jul 29; 38 (17).

[49] Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA - sequencing of 922 individuals. Genome Res.2014 Jan; 24 (1): 14–24.

[50] Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA - seq. Vol.8, Nature Methods.2011. P.469–77.

[51] Martin J, Bruno VM, Fang Z, Meng X, Blow M, Zhang T, et al. Rnnotator: An automated de novo transcriptome assembly pipeline from stranded RNA - Seq reads. BMC Genomics.2010 Nov 24; 11 (1).

[52] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA - seq using the Trinity platform for reference generation and analysis. Nat Protoc.2013 Aug; 8 (8): 1494–512.

[53] Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res.2008 May; 18 (5): 821–9.

[54] Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: Robust de novo RNA - seq assembly across the dynamic range of expression levels. Bioinformatics.2012 Apr; 28 (8): 1086–92.

[55] Chen G, Li R, Shi L, Qi J, Hu P, Luo J, et al. Revealing the missing expressed genes beyond the human reference genome by RNA - Seq [Internet].2011. Available from: http: //www.biomedcentral. com/1471 - 2164/12/590

[56] Pastinen T. Genome - wide allele - specific analysis: Insights into regulatory variation. Vol.11, Nature Reviews Genetics.2010. P.533–8.

[57] Li Q, Seo JH, Stranger B, McKenna A, Pe'Er I, Laframboise T, et al. Integrative eQTL - based analyses reveal the biology of breast cancer risk loci. Cell.2013 Jan 31; 152 (3): 633–41.

[58] Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read - mapping biases on detecting allele - specific expression from RNA - sequencing data. Bioinformatics.2009 Oct 6; 25 (24): 3207–12.

[59] Wei X, Wang X. A computational workflow to identify allele - specific expression and epigenetic modification in maize. Genomics Proteomics Bioinformatics.2013 Aug; 11 (4): 247–52.

[60] Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, et al. AlleleSeq: Analysis of allele - specific expression and binding in a network framework. Mol Syst Biol.2011; 7.

[61] Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature.2007 Jul 26; 448 (7152): 470–3.

[62] Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet.2014; 10 (5).

[63] Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Vol.24, Trends in Genetics.2008. P.408–15.

[64] Cleary MD, Meiering CD, Jan E, Guymon R, Boothroyd JC. Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell - specific microarray analysis of mRNA synthesis and decay. Nat Biotechnol.2005 Feb; 23 (2): 232–7.

[65] Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, et al. Base - pair - resolution genome -

wide mapping of active RNA polymerases using precision nuclear run - on (PRO - seq). Nat Protoc.2016 Aug 1; 11 (8): 1455–76.

[66] Wright RHG, Lioutas A, Dily F Le, Soronellas D, Pohl A, Bonet J, et al. ADP - ribose - derived nuclear ATP synthesis by NUDIX5 is required for chromatin remodeling. Science (1979).2016 Jun 3; 352 (6290): 1221–5.

[67] Herzog VA, Reichholf B, Neumann T, Rescheneder P, Bhat P, Burkard TR, et al. Thiol - linked alkylation of RNA to assess expression dynamics. Nat Methods.2017 Dec 1; 14 (12): 1198–204.

[68] Hale C, Kleppe K, Terns RM, Terns MP. Prokaryotic silencing (psi) RNAs in Pyrococcus furiosus. RNA.2008 Dec; 14 (12): 2572–9.

[69] Neumann T, Herzog VA, Muhar M, Von Haeseler A, Zuber J, Ameres SL, et al. Quantification of experimentally induced nucleotide conversions in high - throughput sequencing datasets. BMC Bioinformatics.2019 May 20; 20 (1).

[70] Wu E, Nance T, Montgomery SB. SplicePlot: A utility for visualizing splicing quantitative trait loci. Bioinformatics.2014 Apr 1; 30 (7): 1025–6.

[71] Shen S, Park JW, Lu ZX, Lin L, Henry MD, Wu YN, et al. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA - Seq data. Proc Natl Acad Sci U S A.2014 Dec 23; 111 (51): E5593–601.

[72] Alamancos GP, Pagès A, Trincado JL, Bellora N, Eyras E. Leveraging transcript quantification for fast computation of alternative splicing profiles. RNA.2015 Sep 1; 21 (9): 1521–31.