RNA SEQUENCING: AN ADVANCE TRANSCRIPTOMICS TECHNOLOGY FROM UNPROCESSED DATA TO INSIGHTFUL INTERPRETATION

Abstract:

Transcriptomics are methods for examining an organism's transcriptome, which is the totality of all of its RNA transcripts. Transcription is the process through which an organism's information is expressed and stored in the DNA of its genome. In order to gain knowledge about a cell's transcriptome, RNA sequencing method makes advantage of potential of other advanced sequencing techniques. RNA-seq analysis can be used to investigate genes and the transcripts that accompany them for a number of functions, including as finding new exons or entire transcripts and analysing gene expression. The functional complexity of transcription has been further understood thanks to recent improvements in the RNA-Seq methodology, which include sample preparation, library creation, and data analysis. We go over every significant stage in RNA-seq data analysis, including experimental design, quality assurance, read alignment, measurement of gene and transcript levels, and differential gene expression. We also emphasise the difficulties that each phase presents. We outline a general bioinformatics workflow for the quantitative analysis of RNA-seq data as well as a few recently released computational tools that can be used at different stages of this workflow. These tools are part of a pipeline for RNA-seq data quality assessment and quantification that starts with raw sequencing files and is targeted at finding and analysing genes that exhibit differential expression under various biological situations.

Keywords: RNA-seq, bioinformatics, quantitative analysis of gene expression, differentially expressed genes

Authors

Komal G Lakhani1
Department of Biotechnology
College of Agriculture
Junagadh Agricultural University
Junagadh, India.

Poojaben M Prajapati
Department of Botany
Bioinformatics
Climate Change & Impact Management
School of Science
Gujarat University
Ahemdabad, India.

Sheetal Gupta
Department of Genetics and
PlantBreeding
Maharana Pratap University of
Agriculture & Technology
Udaipur, India.

Priyanka N Timbadiya
Department of Biotechnology
College of Agriculture
Junagadh Agricultural University
Junagadh, India.

I. INTRODUCTION

The development of genetic sequencing technology has exploded during the past few decades [1]. Over the past 20 years, the number of genomic sequence databases has grown exponentially due to improvements in throughput, accuracy, and cost [2,3,4,5,6]. However, in the field of molecular biology, there remains a considerable challenge in precisely mapping the same genome to multiple phenotypes across different type of tissues, stage of developement, and environmental circumstances. It is not only difficult but also at the core of this challenge to have a deeper understanding of the transcripts and expression of gene regulation. Extensive research has been conducted on a wide range of organism in the field of transcriptomics. The study of transcriptomics field provides valuable understanding of how genes are expressed, and regulated [7,8,9].

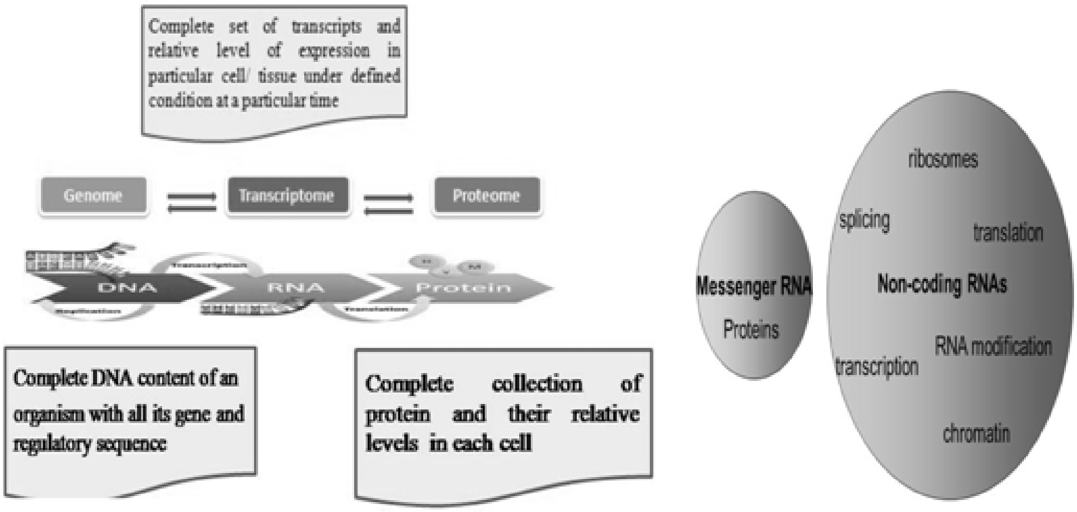

Complete set of transcripts and relative level of expression in particular cell/ tissue under defined condition at a particular time
Genome
Transcriptome
Proteome
Transcription
DNA
RNA
Protein
Replication
translation
Complete DNA content of an organism with all its gene and regulatory sequence
Complete collection of protein and their relative levels in each cell
Messenger RNA
Proteins
ribosomes
splicing
translation
Non-coding RNAs
RNA modification
transcription
chromatin

Figure 1: Central dogma of molecular biology

I. WHAT IS TRANSCRIPTOMICS

The study of the "transcriptome" is known as transcriptomics, and it refers to the entire collection of all the ribonucleic acid (RNA) molecules (also known as transcripts) expressed in a specific entity, such as a cell, tissue, or organism. The primary goal of transcriptomics involve compiling a comprehensive record of all transcript varieties, such as mRNAs, non-coding RNAs, and small RNAs. It also includes determining the genetic transcription structure, encompassing the initiation sites, 5' and 3' regions, splicing patterns, and other modifications occurring after transcription. Additionally, transcriptomics aims to monitor the fluctuations in expression levels of each transcript over time and in different conditions. Transcriptomics have ability to investigate both functional and differentially expressed genes (DEGs), and received the most attention. Transcriptome analysis will eventually provide some advice in disease diagnosis, clinical care, and crop improvement by revealing more about the network of biological processes that are regulated.

II. WHAT IS TRANSCRIPTOME

The term transcriptome refers to the entirety of all transcripts found within a cell during a specific stage of development or under a specific physiological state. In order to understand the functional parts of the genome, uncover the molecular parts of cells and tissues, and comprehend both development and illness, it is essential to possess a deep understanding of transcriptome. In plant research, it is easier and less costly to utilize the transcriptome for monitoring an organism's complete transcriptional activity even when a reference genome is not available, in comparison to the more challenging and costly genome assembly process. Besides indicating alterations in gene expression at different time points and locations, the transcriptome additionally encompasses details regarding auxiliary factors.

The total set of transcripts present in a cell for a particular developmental stage or physiological condition is known as the transcriptome. To interpret the functional components of the genome, expose the molecular components of cells and tissues, and to comprehend development and illness, one must have a thorough comprehension of the transcriptome. The transcriptome can be used to track the entire transcriptional activity of an organism without a reference genome, whereas genome assembly is more difficult and expensive in plant research. The transcriptome contains information about additional metabolic pathways inaddition to allowing researchers to track changes in gene expression over time and space.

According to research, variations in gene expression caused by differences in the plant's growth phases and cultivation conditions result in the distinct accumulation patterns of therapeutic components in various species of medicinal plants. These gene expression patterns are sensitive to both temporal and spatial changes. According to researchers, the transcriptome is better equipped for identification of genes associated with healing properties in medicinal plants. In order to study the functionality of plants [10]. The exploration of the plant functional genome, the organization of gene regulatory domains, the variability of genes (both dominant and recessive), and the identification of bioactive compounds, are neccesary for identification [11, 12].

III. HISTORY OF TRANSCRIPTOMICS

In the past, it was difficult to analyse the behaviour of the genome due to the intricacy of the transcriptome of living cells. Prior to the emergence of transcriptomics, there were multiple investigations conducted on individual transcripts aiming to study the genes. However, numerous studies of individual transcripts were being conducted before the development of transcriptomics to examine the genes. Before the era of sequencing, different approaches such as chromosome walking genetic and QTL mapping and were used for the identification of genes. These methods aid in understanding the molecular and cellular alterations that various variables cause in various tissues and cells.

In 1991, the initial effort was made to gather a section of the human transcriptome and recorded identification of a total 609 mRNA sequences in the human brain. During year 2008, two sets of human transcriptomes containing 16,000 genes and numerous sequences derived from transcripts were made available publicly. It has become common to generate transcriptomes regularly for different disease conditions, tissues, and even individual cells

[13, 14, 15].The quick advancement of technologies with increased sensitivity and economy has been the driving force behind this expansion in transcriptomics [16, 17]. In order to construct expressed sequence tags (ESTs) a short sequences of nucleotide with use of cDNA [18, 19]. The transcripts were randomly sequenced for the purpose of classifying expressed genes and gene fragments [20]. The two primary techniques for genetic analysis, namely microarray and RNA-Seq, emerged in the mid-1990 and 2000s, respectively [21, 22]. The first microarrays, which use a fixed set of transcripts hybridised to a variety of complementary probes, were reported in 1995. Massively parallel signature sequencing (MPSS), used for idetification of mRNA expression levels of by counting the number of individual mRNA molecules produced by each gene. This technique were used to confirm the expression of roughly 10,000 genes in *Arabidopsis thaliana* and about 20,000 genes in 32 different human tissues [23, 24]. According to study MPSS is based on sequencing reads of 16–20 bp [25].Subsequently, researchers began employing methods like reverse transcriptase quantitative PCR, northern blotting and nylon membrane arrays to quantify specific transcripts [26]. These techniques were challenging, though, and they only managed to record a small portion of the transcriptome. During the year 2006, the 454 technology was used for RNA-sequencing with 105 transcripts with a appropriate coverage for quantification of relative transcript abundance [27]. After the year 2008, the sequencing of RNA became in popular using new Solexa/Illumina technologies and approximately 10^9 transcript to be recorded.

IV. WHY NEED OF TRANSCRIPTOMICS

During early 2000s, to study biology of plant traditional breeding methods is mainly used and it is focused on to single genes identification of within a biological context. Numerous organisms have been extensively researched for transcriptomics, offering valuable perspectives on the structure, expression, and control of genes [7, 8,9]. The transcriptomics investigations have made significant progress due to the rapid advancement of the underlying sequencing technology [28]. Transcriptomics have become an important tool in the field of plant breeding for the assembly of crop reference genome. The traditional quantitative trait loci (QTL) analysis uses linkage maps to identify regions of the genome linked to variations in traits within a population, in contrast to conventional agricultural practices that solely select traits based on physical characteristics [29]. The whole genome information is available in only few plant spp. as well as it is very tedious and costlty. So, in the absence of reference genome, transcriptomics study is an alternative tool to identify overall transcriptional activity in plant spp. with complex genomes, including coffee, wheat, sugarcane, and several "orphan crops," including sweet potato, chickpea, and minor millets to unravel genes and regulatory regions and provides a variety of molecular markers to identify diversity [30].

The main threats to sustainable agricultural production and global food security are global climate change such as drought, salinity, high temperature and several biotic stress. In plant, metabolic activity and expression of genes is over the time at different growth period and spatial condition. However, molecular biology still faces a major challenge in accurately correlating the same genome to multiple phenotypes across different tissue, developmental stages, and environmental conditions. To deep understanding transcripts and how genes are regulated is major problem. Transcriptomics strategies used to identify the plasticity of gene expression that are activated or inactive under any specific external or developmental stimulus, and is required to reduce the negative effects of these threats to agricultural productivity, enhance crop yield and stress-tolerance in plants. Due to advancements in sequencing technology, the conventional chip hybridization platform has been replaced by RNA sequencing technology in recent years for transcriptome research [31].

V. TRANSCRIPTOMICS TECHNIQUES

To determine and quantify the transcriptome, numerous technologies have been created, such as hybridization- or sequence-based methods.

1. **Hybridization Based Methods**
   - Fluorescently labelled cDNA with custom-made microarrays
   - Commercial high-density oligo microarrays

2. **Sequence Based Methods**
   - Sanger sequencing of cDNA or EST libraries
   - Serial analysis of gene expression (SAGE)
   - Cap analysis of gene expression (CAGE)
   - Massively parallel signature sequencing (MPSS)
   - RNA-Sequencing

VI. RNA-SEQUENCING

RNA-Seq is based on next generation high-throughput sequencing of nucleic acids to determine the nucleotide sequence of RNA molecules as well as the quantities of specific RNA species within populations of RNA molecules [32]. The advancements in transcriptome techniques, such as single-molecular, single-cell, and spatial analysis, have enabled more precise identification and understanding of individual cells, alongwith the additional inclusion of spatial information, as compared to the previous technique of bulk RNA sequencing. The transition from bulk RNA sequencing to single-molecular, single-cell, and spatial transcriptome techniques has made it possible to resolve individual cells with ever-greater accuracy while also including spatial data. To address the limitations of sequencing technologies, such as sequencing errors, length biases, and fragmentation, specialized computational tools are required for RNA-seq analysis [33, 34].

The scope of RNA sequencing is vast, spanning a wide range of fields such as genetics, genomics, cancer research, developmental biology, and more. It is an essential tool for bettering our knowledge of gene regulation and how it affects different biological processes because of its ongoing progress and technological integration.

Here are some salient features about the significance and scope of RNA sequencing:

1. **In Medical And Clinical Research,** RNA-seq is an indispensable tool for investigating a wide range of diseases, such as cancer, infectious diseases, neurological disorders, and uncommon genetic abnormalities. It assists in locating biomarkers and gene expression patterns unique to a given disease that may be applied to both therapeutic and diagnostic settings.

2. **Developmental Biology:** Gene expression alterations throughout embryonic development, tissue differentiation, and organogenesis are investigated using RNA-seq. With it, a thorough understanding of the transcriptome dynamics during these processes is possible.

3. **Microbiome Research:** To examine the RNA transcripts of microorganisms in intricate microbial communities, such as bacteria, viruses, and fungus, RNA-seq is utilised. This contributes to our understanding of the microbiome's function in both health and illness.

4. **Evolutionary Studies:** RNA-seq data provides valuable insights into gene expression variations between species and evolutionary processes, as well as the functional conservation or divergence of genes. It is particularly useful for comparative genomics experiments.

5. **Gene Expression Analysis:** Using RNA-Seq, scientists may measure and compare the levels of gene expression in various samples or environments. This facilitates comprehension of the regulation of genes and the variation in their expression in reaction to distinct stimuli, illnesses, or environmental conditions.

6. **Transcriptome Assembly and Annotation:** Transcriptomes are assembled and annotated using RNA-Seq data, yielding an extensive list of expressed genes and their isoforms.

7. **Functional annotation:** Annotating the functional elements of genomes through the identification of enhancers, promoters, and other regulatory areas is made easier with the help of RNA-Seq data.

8. **Isoform Analysis and Alternative Splicing:** RNA-Seq may detect alternative splicing events, which can reveal several isoforms of a gene. The diversity and function of proteins must be understood in relation to alternative splicing.

9. **Non-Coding RNA Analysis:** By using RNA-Seq, non-coding RNAs like micro RNAs, long non-coding RNAs (lncRNAs), and circular RNAs can be found and profiled. Numerous biological activities, including the control of genes, depend on these non-coding RNAs.

10. **Pharmacogenomics:** Individual differences in drug efficacy and metabolism can be understood and drug responses predicted using RNA-Seq data.

11. **Epigenetics:** To investigate how DNA methylation and chromatin changes impact gene expression.

12. **Evolutionary Biology:** It uses RNA-Seq to investigate the variations in gene expression between species, which aids in the understanding of how gene regulation has changed over time.

13. **Single-Cell RNA Sequencing (scRNA-Seq):** This method allows researchers to explore cellular heterogeneity and cell types in complicated tissues by enabling single-cell gene expression analysis.

14. **Cancer Research:** RNA-Seq is widely utilized in cancer research to investigate tumour heterogeneity, find possible biomarkers.

15. **Biotechnology in Agriculture:** RNA-Seq is utilized in agricultural research to generate genetically modified organisms, increase crop yield, and comprehend gene expression in plants and animals.

16. **RNA Editing and Alterations:** By identifying RNA editing events and post-transnational alterations, RNA-Seq provides insight into the transcriptome's dynamic character.

VII. PRINCIPE OF RNA SEQUENCING

The adaptors are added to either one or both ends of an RNA population (whether total or fractionated, like poly(A+), to convert it into a collection of cDNA fragments. The next step involves high-throughput sequencing of each molecule to collect brief sequences from either one end (single-end sequencing) or both ends (pair-end sequencing), regardless of whether amplification is present or not. After sequencing, the obtained reads are either matched to a reference genome or transcripts, or they are constructed from scratch without the aid of the genomic sequence to create a genome-scale transcription map that includes the transcriptional organisation and/or level of expression for each gene. By switching from bulk sequencing to single-molecular, single-cell, and spatial transcriptome techniques, improvements in RNA sequencing techniques have made it possible to understand individual cells more precisely. In addition to improving cell resolution accuracy, this transition has also given spatial data. Numerous scientific discoveries have been made as a result of the use of computational analysis on RNA-sequncing data. These include the identification of biomarkers and pathogenic mutations, the development of novel therapeutics, and thorough understanding of genetic regulatorymechanisms [35].

The 454 technology from Roche, the Solexa technology from Illumina, SOLiD technology from ABI are examples of NGS, or high-throughput sequencing [36, 37, 38, 39, 40, 41, 42]. Depending on the DNA-sequencing method being employed, the reads are typically between 30 and 400 bp. RNA sequencing overcomes microarray's limitations and provides better understanding for transcriptome study.

VIII. TYPES OF RNA SEQUENCING

1. On the basis of the Formation of cDNA

- **Direct RNA Sequencing:** Bypassing the process of turning extracted RNAs into cDNA, this approach directly sequences RNAs. Compared to DNA, RNAs are more unstable, making them more challenging to work with. The process of turning RNAs into cDNAs, however, generates a number of biases, error sites, and disruptions that obstruct precise sequencing. Moreover, the process of creating cDNA is complex and involves multiple steps, and cDNAs are not suitable for sequencing shorter RNAs.
- **Indirect RNA Sequencing:** All of the RNA types present in the sample are sequenced using a process called whole transcriptome sequencing (WTS). It gives all the

necessary information on a cell's nucleotide and gene expression because it profiles the complete transcriptome.

2. **On the basis of types of RNA Sequenced**

- **Whole Transcriptome RNA-Sequencing (Total RNA-Seq)**: All of the RNA types present in the sample are sequenced using a process called whole transcriptome sequencing (WTS). It gives all the necessary information on a cell's nucleotide and gene expression because it profiles the complete transcriptome.
- **mRNA-Sequencing**: Only mRNAs are sequenced using this technique. First, mRNAs are separated using poly-A chromatography or poly-A magnetic beads, and a poly-A library is created. The library is then either directly or indirectly sequenced to obtain an mRNA sequence.
- **tRNA-sequencing and rRNA-sequencing**: This is rarely used technique of sequencing which is used for sequencing of tRNAs and rRNA.
- **Targeted RNA-Sequencing**: It is a technique for sequencing a particular transcript of interest.
- **Small RNAs Sequencing**: This kind of sequencing involves the sequencing of a non-coding RNAs such as miRNA, siRNA, and piRNA.
- **Single Cell RNA Sequencing**: This technique involves sequencing RNAs that have been isolated from just one type of cell or cell line. Transcript libraries are created, all the transcripts from a single cell are collected, and then the entire library is sequenced.

IX. STEPS IN RNA SEQUENCING

The most popular RNA-Seq approach is the indirect method, which relies on the cDNA production process. In this article, we will outline a general step in the indirect RNA-Seq method. The overall RNA-Seq workflow can be summed up as follows:

1. **RNA Isolation:** Cell lysis and full transcriptome extraction are the first steps in RNA-Seq. Cell lysis and RNA extraction frequently involve the use of RNA lysis buffer and organic solvent-based RNA isolation techniques. Following extraction, the RNAs are cleaned, purified as DNA-free RNAs, and stored in buffer or RNase-free water.

2. **RNA Selection**: A cell contains enormous amounts of RNA, including the three most prevalent types: rRNA, tRNA, and mRNA. Certain types of RNAs that we need to sequence from the extracted transcriptome are chosen using a variety of techniques, including affinity chromatography, electrophoresis, filtration (size exclusion), enzymatic depletion, target enrichment/depletion, etc. In case of whole transcriptome RNA-Seq. there is no need to choose a specific RNA. Since mRNAs are the direct transcripts of genes that contain coding sequences, they are frequently chosen and sequenced. The method that is most frequently used to separate mRNAs from the entire collection of RNAs is poly-A library creation.

3. **Synthesis of cDNA**: Reverse transcription is used to convert the isolated and chosen RNAs into first-strand cDNAs, which are more stable than sample RNAs. The second-

strand cDNAs are created by amplifying the first-strand cDNAs with Taq DNA polymerases and nucleotides.

4. **Selection**: It is an optional procedure used to get rid of globin, other smaller RNAs, and rRNA molecules, which make up around 80% to 90% of the total amount of cellular RNA.. This phase will streamline the RNA sequencing procedure, increasing its effectiveness and resulting in a reduction in costs, time, and reagent usage. The required cDNA can be chosen via enzymatic depletion, probe-based depletion, rRNA depletion, target enrichment and other approaches.

5. **Library preparation:** A collection of complete cDNA synthesised for sequencing is known as a cDNA library. A library Preparation includes following steps:

   - **Size Selection and Fragmentation:** The procedure of improving the sequencing of targeted RNAs and cDNAs is also optional. The cDNAs are broken apart, and only pieces of a particular size are chosen. Chemical, enzymatic, or physical processes, such as sonication, can fragment materials.
   - **Ligation of Adaptors:** Adaptors are ligated at the end of fragmented and/or chosen cDNAs. Short synthetic oligonucleotides called adaptors bind to transcripts and act as sequencing primer sites.
   - **Indexing and Amplification**: During PCR amplification of the cDNAs, a particular sequence (also known as a barcode) is added to the transcripts following adaptor ligation known as indexing. The concentration of the created library is raised by amplifying these adaptors and barcode-ligated cDNAs.

6. **Sequencing:** The high throughput next generation sequencing (NGS) technique is used in the high throughput NGS machine to sequence the final cDNAs in the cDNA library. The sequencing procedure is comparable to DNA sequencing. The acquired data are analyzed using bioinformatics methods after each cDNA fragment is read separately.

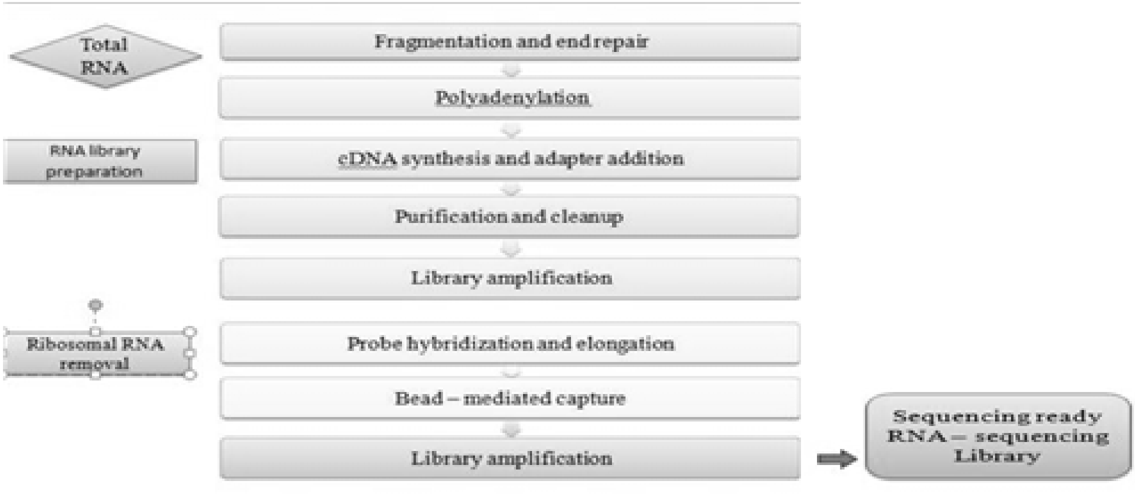


Figure 2: RNA- Sequencing library preparation workflow

X. IN SIILICO ANALYSIS OF RNA SEQUENCING

1. **From Raw Data to Smart Analysis with RNA-seq Data Science:** The RNA-seq approach is usedto identify a gene or protein that is not functioning proper and has an adverse effect on subsequent processes, leading to emergence of a disease state [43]. For better comprehension of the complex biological makeup of transcriptomes in various organisms, including humans, RNA-seq data analysis using computational methods is crucial [44]. For these processes, hundreds of computational tools and resources have been created, and it has been demonstrated that each one produces outcomes that are unique and superior to those of its predecessors. Evethough, some studies compared the primary tools available for any given step, and combinations of steps in an exhaustively *viz*., read alignment and quantification, to more clearly demonstrate choosing each pipeline and show potential features of step interaction [45, 46]. In this article, we provide a comprehensive overview of crucial stages involved in analysis of computational RNA-seq data, starting from the gathering of unprocessed data to the discovery of biological insights.

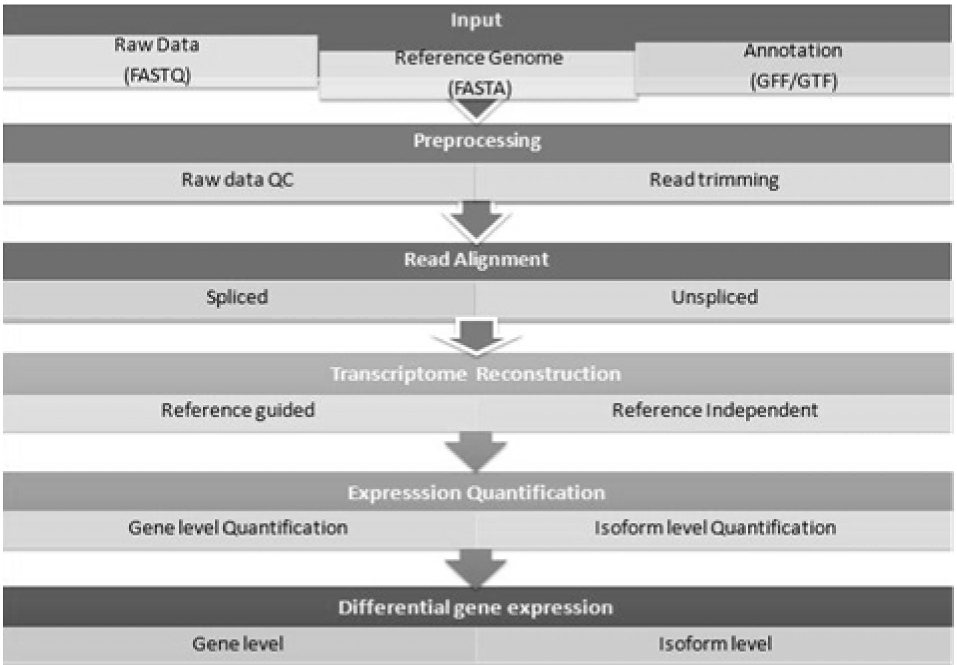


Figure 3: RNA-Sequencing Data Analysis Workflow

- **Preprocessing of Raw Data:** RNAseq data is formatted in FASTQ (sequence and base quality), which is similar to whole genome or exome sequencing. The analysis in FastQc was performed by a series of analysis modules. FastQC is a java-based tool that provides a simple way to perform quality control checks on raw sequence data coming from high throughput sequencing pipelines and give a quick impression of whether data has any problems of which should be aware before doing any further analysis.

The library preparation, sequencing, and imaging stages can introduce a lot of incorrect sequence variations, which need to be found and eliminated in the data analysis step [47]. The FASTQ format containing sequence and base quality information utilized for RNAseq data, resembling the formats used in whole genome or exome sequencing. To filter and evaluate the quality of original reads, it is crucial to perform quality control and trimming on reads after they have been generated. The read trimming process removes adaptor sequences and imperfectly accurate read fragments, which lowers PHRED quality score [48, 49, 50]

The criteria for trimming were given as under

- Removal of low-quality sequence (limit=0.05)
- When an adapter is found trim 3'end, for reads without adapters: keep the reads.

The first phase of a typical RNA-seq workflow should involve doing quality control on the raw data. The tools like FastQC and HTQC can be used in this stage to evaluate quality of original data. These tools make it easier to assess overall quality as well as quality of each base of each read within each sample. Depending on technique used to build RNA-seq library trimming the reads before aligning the data may be advised. Fast QC report spots the problems which originate either in the sequencer or in the starting library material.

The trimmomatic tool is a more flexible and efficient preprocessing tool, which could correctly handle paired-end data. The main processing steps performed by trimmomatic are identification of adapters and quality filtering.

Therefore, the fast Q file containing the filtered, high-quality reads were further pre-processed using trimmomatic for adapter removal, PCR primer sequences or fragments removal and quality filtering. The high-quality reads were assembled into contigs/transcripts by trinity assembler.

- **Read Alignment:** The read alignment stage of RNA-seq downstream analysis is crucial. The order and source of reads which could refer to the particular region, homolog, or strand of genome they come from are frequently hidden in RNA seq data. The read alignment step utilize a genome or transcriptome as a point of reference in two distinct manners. Typically, reads are assigned to either a genome or a transcriptome. The percentage of reads that are mapped is a crucial measurement of mapping quality since it serves as a general indicator of both DNA contamination and overall sequencing accuracy.

  The alignment of reads to a reference sequence also exposes the coverage, or amount of overlap each location on reference sequence with reads. The various bioinformatics tools such as GenomeScope, smudgeplot and merqury an calculate coverage without mapping reads to a reference sequence because the majority of read overlap is conserved with or without the reference sequence [51, 52, 53]. The expression levels and types of transcripts can be determined by aligning RNA-seq reads to a complementary reference sequence. To find transcripts that are absent from reference sequence, however, this method is ineffective [43]. .

- **Strategies of Read Alignment:** A genome or transcriptome can be used as a reference in two different ways for the read alignment step [54].The transcriptome is made up of all the transcripts found in a certain specimen that have undergone splicing by incorporating the exons and excluding the introns. For accurate read mapping when a transcriptome is the reference, unspliced aligners that don't permit big gaps may be the best option. In this situation, you can make use of Bowtie [55], Stampy , Mapping and Assembly with Quality (MAQ) [56], and Burrow-Wheeler Aligner (BWA) [57].

  The alignment can only be utilized to detect identified exons and junctions, as it does not identify splicing events involving new exons. If genome is used as a reference, it is advisable to employ spliced aligners that allow for a wide range of gaps. This is because reads that match at exon-exon junctions would be divided into two separate fragments. The alignment of the reads to a reference sequence also exposes the coverage, or amount of overlap each location on reference sequence wit reads. The utilization of this approach may enhance chances of uncovering new transcripts generated through alternative splicing. Several spliced aligners, such as TopHat, MapSplice, STAR, and GSNAP, have been developed [58, 59, 60, 61].

- **Challenges of Read Alignment:** It becomes difficult to align RNA-seq reads with a reference genome due to the occurrence of spliced junctions. These junctions arise when a section of an RNA-seq read aligns with the end of one exon and another section aligns with a different exon. Spliced junctions result from the removal of introns and splicing together of exons, generating multiple transcripts. Alternative splicing is advantageous for producing protein variants from the same genetic information. It is possible to monitor readings across known splice junctions using the current arrangement of exons. By precisely aligning reads across the junctions between exons and introns\in reference genome, splice alignment software packages aim to decrease the occurrence of multiple mappings. Assembling transcripts that are not annotated in reference using spliced read alignments is an important step in reference-guided assembly. Consistency of read coverage on exons and mapping orientation on complementary strand are two additional critical elements. Readings may tend to accumulate more in poly(A)-selected samples, which may be a sign of poor RNA quality in starting material. A mapped read's GC content may show PCR biases. Several tools such as Picard, RSeQC and Qualimap are among the mapping quality control tools [62, 63, 64].

  To evaluate biases in RNA-seq data, various indicators can be examined, such as the proportion of exonic or rRNA reads, the precision and biases in estimating gene expression, the presence of GC bias, the uniformity of coverage, the bias in coverage from 5' to 3', and coverage at 5' and 3' ends [65]. It is critical to check them for biases in terms of GC content and gene length once actual values for transcript quantification have been established. This assessment is required to find any potential discrepancies and make it easier, if necessary, to apply right normalization techniques. Assuming that the reference transcriptome has bee n thoroughly annotated, experts have the ability to analyze biotype makeup of sample. This analysis serves as an indication of the effectiveness of the RNA purification process. For instance, standard polyA longRNA preparations shouldn't contain rRNA or short RNAs. Several R tools, such NOISeq [66] and EDASeq offer helpful graphs for quality assurance of count data. RNA-SeQC [67], RseQC [63], and Qualimap 2 [68] for the purposes, which typically accept BAM files as input.

Furthermore, oin order to determine overall RNA -seq quality dataset, it is crucial to evaluate batch effects and gurantee consistency in repeated analysis. Technical replicates should normally have good reproducibility (Spearman R2 > 0.9), but there is no set criteria for biological replicates because they depend on the variability of the experimental system. In principal component analysis, when there are variation in gene expression observed under different experimental condition, it is anticipated that biological replicates having the ssame condition will cluster together.

- **Reconstrution of Transcript:** Transcriptome reconstruction refersto the process of identifying all transcript expressed in particular specimen. It involves the use of two methods: 1). Reference guided method and, 2). Reference independent method

➢ **Reference guided method:** If reference genome is available, RNA-seq analysis often entails mapping readings onto the transcriptome or reference genome to determine which transcripts are expressed [69]. The scope of the research is restricted to evaluating and calculating solely in relation to the recognized transcriptome of a particular species. It becomes challenging to locate recent transcripts that have not been categorized.

➢ **Reference-independent approach:** The initial step in analysis process involves bringing together the fragmented reads to form longer contigs, following which these contigs are considered as the expressed transcriptome. In this step, the reads are mapped back to these contigs to assess their quantity. The expression level of transcripts can be computed using read coverage in both situations. The decision of whether to do transcript identification and quantification sequentially or simultaneously is a fundamental one [70]. Direct consensus transcript construction from short reads without reference is accomplished using a de novo assembly algorithm. This can be achieved via Trinity, Oases and transABySS [71, 72, 73].

Table 1: RNA Seq Denovo Assembly Software

Software	Resource load	Features
Velvet Oases	heavy	Shord read assembler
SOAPdenovo-trans	Moderate	assembler early short read, and updated for transcript assembly.
Trans-ABySS	Moderate	Utilized for genome of considerable size that make it practical to process short reads, and offers MPI-parallel version available.
Trinity	Moderate	Short reads, large genomes, memory intensive

Newbler	Heavy	Specially designed to adress errors related to homo-polymers processing during analysis of Roche 454 sequences
CLC genomics workbench (Qiagen—Venlo, Netherlands	Light	Graphical user interface, hybrid data

Source: Lowe *et al.*, 2017 [9].

XI. RNA SEQ MAPPING

A quantification of alignment with the number of reads that mapped to transcript Read representation was determined by mapping the reads back to their corresponding assemblies for each of the six species individually. This mapping number depends on several factors, such as the actual expression level, the library size, percentage of reads aligning, transcript length, and GC content.

To accurately analyze differential gene expression, it is critical to have a significant portion of reads that can be linked to the transcriptome assembly. This is due to fact that increasing number of reads that can be linked back to the assembly will increase statistical power needed to carryoutthese analysis. Subsequently, a significant proportion of those readings ought to correspond with the assembly. It is necessary to also consider the number of genes detected as a measure of completeness, in addition to proportion of reads that align with an assembly. The exclusion of low-quality reads from assembly resulted in a slight decline in the overall mapping percentages when the reads were compared to the transcriptome. However, reads aligning with exons resulted in a significant increase number of reads.

XII. ANALYSIS OF DIFFERENTIAL EXPRESSION OF GENES (DEGs)

One of the goals of RNA-seq data analysis is differential expression analysis. The data are normalized by RPKM, FPKM, and TPM either by taking into account the abundance of transcripts in different samples or by accounting for their varying amounts. The depth of sequencing, a fundamental factor that is essential for comparing samples, is eliminated in the process. The performance of normalization methods based on total or effective counts is typically lower when the transcript distributions of samples are not uniform, meaning that strongly and differentially expressed features can affect count distribution in a biased manner [74, 75]. Several software packages and pipelines, such as edgeR [76], NOIseq [66], Cuffdif [77], DESeq [78], SAMseq [79], and EBSeq [80] have been created for the purpose of conducting differential expression analysis.

To identify genes that exhibit differential expression between samples, the tool DESeq2 was created. A model based on the negative binomial distribution is used by the software program DESeq2 to assess differential expression. Raw read count data gathered with HTSeq must be input with collected a heat map,volcano plot and MA plot. Transcripts that were expressed differently had an absolute change in expression of more than two times, with a p-value of 0.05 after being corrected for false discovery rate. The normalization techniques TMM, DESeq, PoissonSeq, and UpperQuartile consider this factor and exclude features that are highly variable or strongly expressed [81, 82]. According to study the difficulties in comparing samples within a group are further compounded by fluctuations in transcript length and biases in coverage along the transcript, even after accounting for their control [83].

Table 2: Software for differential Expression of Genes

Primary category	Tool name	Notes
Splice-aware read alignment	GEM	Utilize approximate string alignment methods for filtration
	GSNAP	Constructed using a complex variants handling seed and extend alignment algorithm
	MapSplice	Making use of Burrows-Wheeler Transform (BWT) algorithm
	RUM	Integrates alignment tools Blat and Bowtie to increase accuracy
	STAR	This approach involves for seed in uncompressed suffix arrays, then performing clustering of seed and stitching step, resulting fast but memory-intensive process
	TopHat	The read aligned usesing Bowtie, which based on BWT. To identify split read, exons are resolved through split reads mapping
Transcript assembly and quantification	Cufflinks	Assembles transcripts to reference annotations or de novo and quantifies abundance
	FluxCapacitor	iReckon use method of quantifies transcripts through use of reference annotations to determine abundance of recently discoverd isoforms
	iReckon	Relative quantity of novel isoforms estimated through simulation
	BaySeq	Calculate posterior probabilities using count-based method that applies empirical Bayesian method
	Cuffdiff212	Strategy based on beta negative binomial distribution and relying on Isoform

	EdgeR	Within the constraints of negative binomial model, a strategy employed that rely on Count and empirical bayes approach
	DEseq2	Application of negative binomial model with use of exon-focused strategy.
	DEGseq2	Applying isoform to the Poisson model
	MISO	Estimating posterior probabilities using isoform-specific bayes factor models.
Other tools	HCP	Utilizes prior knowledge to estimate and adjust for known and undisclosed factors in order to standardize expression data.

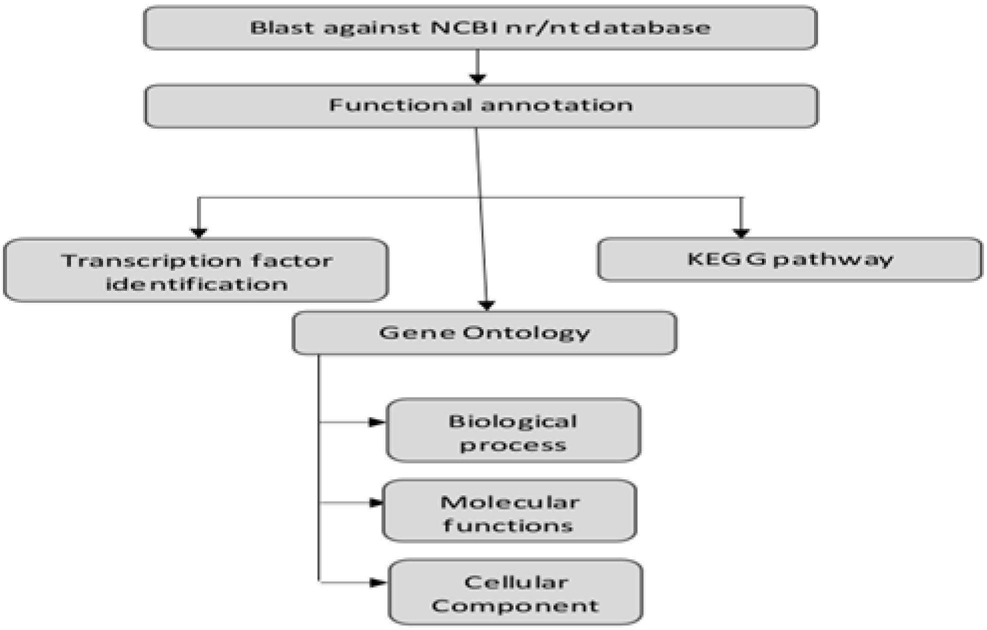
XIII. FUNCTIONAL ANNOTATION

Annotations are created to understand the biological importance of sequence data. The recognition of genes in particular sequence data that have known functions in the corresponding genomes is improved, as is understanding of biological mechanisms. The workflow shown in Figure 3.3 was used for functional annotation of all transcripts retrieved after assembly. The BLASTx algorithm was used to analyze the collected transcripts using non-redundant NCBI database [84].

The process of annotation is split into two sections: mapping and annotation. All Gene Ontology (GO) terms connected to results of BLAST search are gathered during mapping B2G plugging process. Once mapping process is completed, the mapping statistics will display the results and a summary is given in three evaluation charts.

The distribution of evidence codes for blast hits and sequences, as well as details about the database used to obtain annotations, are all included in mapping. Mapping results in development of pricing system for the pool and annotation, followed by the selection and assignment of relevant GO keywords to query sequence. For Blastx and Blast2GO, the parameters are

e-value, = 10-e6
Similarity. = 35%
Annotation cutoff >=55
GO weight cutoff >= 5



For the de novo assembly, Trinity (v2.5.1) was used to the clean readings [72]. The duplicate sequences were removed using the CD-HIT-EST programme, with the settings set to a similarity of 0.95 and a sequence length of 10 bases [85]. By choosing the parameters E-value 1e-5 and E-value 1e-10 for BLAST [84] and Hmmer [86], annotation information was eventually collected. The COG conduct a blast search in order to determine roles and identify any differentially expressed genes of newly discovered genes [87], GO [88], KEGG pathway [89], Swiss-Prot [90] and Non-redundant protein (NR) databases [91].

XIV. GENE ONTOLOGY (GO) ENRICHMENT ANALYSIS

The use of the Gene Ontology (GO) analytic approach is expanding for broad functional analyses employing genomic or transcriptome data. Genes are categorized by some shared biological characteristic in order to do this analysis. Then, it is determined which categories are overrepresented among the differentially expressed genes. There are many programmes with GO-related analytic features, new tools are still required to meet the demands for data produced by recently emerging technologies or for advanced analysis purposes [92]. It is usual practise to employ Gene Ontology (GO) categories in this technique, and there are numerous programmes available for GO analysis, including EasyGO [93], GOminer [94], GOstat [95], GOToolBox [96], topGO [97], GSEA [98], and DAVID [99].

Top DEGs were identified with GO terms and this GO analysis was performed to identify significantly enriched GO terms of differentially expressed genes. GO classification done by software such as WEGO 2.0 to identify differences in functional categories. WEGO (Web Gene Ontology Annotation Plot) is a tool for visulizing, comparing and plotting to GO annotation results [100]. The division of GO extends to three sub-ontologies, namely biological process, cellular component, and molecular functions. The text file containing merged data of dormant and non-dormant genotype having significant gene ID and its respective up and down-regulated annotated GO terms were uploaded to WEGO 2.0 and by keeping native file format enrichment was performed which classified enriched GO to different categories. Analysis of their further distribution into various sub-categories revealed the majority of GO terms were enriched from all the stages.

XV. KEGG PATHWAYS

The functional annotation of individual genes is still mostly unfinished, despite the genome sequencing project quick ability to identify gene catalogues for an expanding number of organisms. In an effort to connect genomic data with higher order functional data, KEGG (Kyoto Encyclopaedia of Genes and Genomes) standardize gene annotations and computerize current understanding of cellular process [101]. KEGG is made up of mainly three databases.The key gene involved in the biological and metabolic pathway could be identified through KEGG (Kyoto Encyclopaedia of genes and genomes database. LIGAND collect information related to chemical compounds in cells, enzyme molecules, and enzymatic reactions. The GENES collect catalogues of gene for all the fully as well as partially sequenced genomes. These genes are involved in various function and expressed at different location, at different stress condition like temperature stress, drought stress and salinity stress etc. The statistical enrichment of differentially expressed genes in KEGG pathways is investigated using KOBAS software in order to understand the roles and applications of DEGs [102].

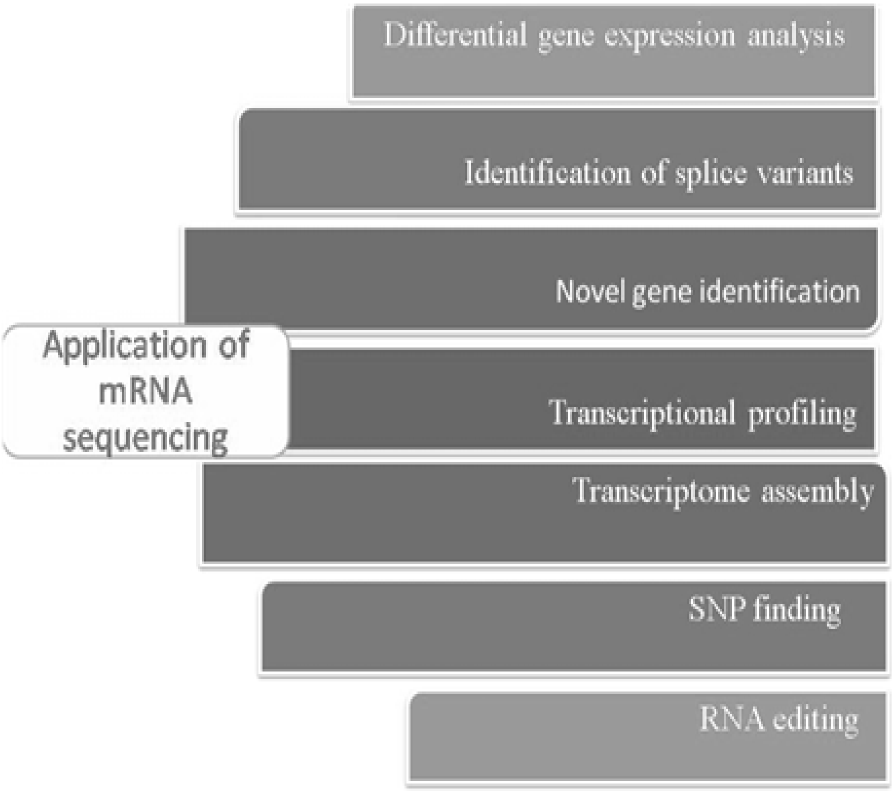

Differential gene expression analysis
Identification of splice variants
Novel gene identification
Application of mRNA sequencing
Transcriptional profiling
Transcriptome assembly
SNP finding
RNA editing

Table 1: Reported research results of RNA-Seq in different fields

Application field	**Species**	**Result**
Plants:		
Patterns of gene network & expression analysis	Finger Millet	Aluminium toxicity
	Amomum villosum	Terpene biosynthesis
	Malus domestica	Gene involved in hormone transduction and fruit ripening biosynthesis of anthocyanin
	Brassica napus	Peroxisome related pathways
	Oryza sativa	Expression of temperature and nitrogen responsive DEGS at meiosis stage
	Crocus sativus	Flavonoid Biosynthesis
	Prunus armeniaca	Drought responsive genes expression
Functional gene mining	*Euryale ferox* Salisb	Idenify 85, 006 unigenes
	Arachis hypogea	Mining of seed specific candidate genes and 377 genes identified
	Cupressus gigantea	Idenify 1,01,092 unigenes
	Dracocephalum tanguticum	Idenify 1,51,463 unigenes
Development of molecular markers	*Magnolia ashei*	10,406 SSRs
	Cephalotaxus oliveri	3900 EST-SSRs
	Zingier officinale Rosc.	16,790 SSRs
	Amomum tsaoko	55,590 EST-SSRs
	Epinephelus tukula	44,565 SSR 1,22,220 SNPs
Genetic mechanism exploring	*Gentiana rigescen*	Genetic relationship
	Capparis spinosa	Breeding programs and phylogenetic studies

Animal		
Patterns of gene network & expression analysis	*Longissimus dorsi* Chinese Red Steppe cattle	Differenrial expression genes related to biological processes such as short-chain fatty acid metabolism, regulation of fatty acid transport and PPAR signaling pathway
	Holstein Cattle	Study host immune response to vaccines & identification of DEGs related to immune response
	Holsteins & Jersey cows	Identify genes expression related to biological process of feed efficiencey
Microorganism		
Gene expression analysis	*Rosellinia necatrix*	Identification genes involved in the production of fungal toxins, detoxification and transport of toxic compound
	P chrysosporium *P. placenta*	Reulation of genes involved in lignocellulosic cell wall attack
	COVID-19	Identify biomarkers for disease severity and expression of neutrophil-related transcripts

XVI. CONCLUSION

RNA sequencing has undergone remarkable advancements, transforming transcriptomics technology from raw, unprocessed data to insightful interpretation. With its ability to capture the entire transcriptome, including coding and non-coding RNA species, RNA sequencing has revolutionized our understanding of gene expression dynamics in health and disease. Through innovative bioinformatics pipelines, it enables the extraction of meaningful biological insights from vast amounts of sequencing data, facilitating the identification of novel transcripts, alternative splicing events, and regulatory mechanisms. Moreover, the integration of RNA sequencing with other omics technologies has opened new avenues for comprehensive systems biology approaches, providing a holistic view of cellular processes. As RNA sequencing techniques continue to evolve, becoming more cost-effective, sensitive, and scalable, they promise to unravel the complexities of gene regulation with unprecedented depth and precision, ultimately driving advancements in personalized medicine and therapeutic development.

REFERENCES

[1] Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature*, 550(7676), 345–353. https://doi.org/10.1038/nature24286.

[2] Lathe, W., Williams, J., Mangan, M., & Karolchik, D. (2008). Genomic data resources: challenges and promises. *Nature Education*, 1(3), 2.

[3] Heather, J. M., & Chain, B. (2016). The sequence of sequencers: the history of sequencing DNA. *Genomics*, 107(1), 1–8 . https://doi.org/10.1016/j.ygeno.2015.11.003.

[4] Levy, S. E., & Myers, R. M. (2016). Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics*, 17, 95–115. https://doi.org/10.1146/annurev-genom-083115-022413.

[5] Ardui, S., Ameur, A., Vermeesch, J. R., & Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research*, 46(5), 2159–2168. https://doi.org/10.1093/nar/gky066.

[6] Karsch-Mizrachi, I., Takagi, T., Cochrane, G., and International Nucleotide Sequence Database Collaboration (2018). The international nucleotide sequence database collaboration. *Nucleic Acids Research,* 46, D48–D51. https://doi.org/10.1093/nar/gkx1097.

[7] Jain M. (2012). Next-generation sequencing technologies for gene expression profiling in plants. *Briefings in Functional Genomics,* 11(1), 63–70. https://doi.org/10.1093/bfgp/elr038.

[8] Casamassimi, A., Federico, A., Rienzo, M., Esposito, S., & Ciccodicola A. (2017). Transcriptome profiling in human diseases: new advances and perspectives. *International journal of molecular sciences,* 18(8), 1652. https://doi.org/10.3390/ijms18081652.

[9] Lowe R., Shirley N., Bleackley M., Dolan S., & Shafee T. (2017). Transcriptomics technologies. *PLoS Computational Biolology*, 13, e1005457. https://doi.org/10.1371/journal.pcbi.1005457.

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

[11] Tyagi, P., & Ranjan, R. (2021). Comparative study of the pharmacological, phytochemical, and biotechnological aspects of *Tribulus terrestris* Linn. and *Pedalium murex* Linn: An overview. *Acta Ecologica Sinica*, 43(2), 223-233. https://doi.org/ 10.1016/j.chnaes.2021.07.008.

[12] Tyagi P., Singh A., Gupta A., Prasad M. and Ranjan R. (2022). Mechanism and function of salicylate in plant toward biotic stress tolerance. *In* book: *Emerging Plant Growth Regulators in Agriculture,* chapter 4, 131–164. https://doi.org/10.1016/B978-0-323-91005-7.00018-7.

[13] Sandberg, R. (2014). Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods*, 11(1), 22–24. https://doi.org/10.1038/nmeth.2764.

[14] Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Molecular Cell,* 58, 610–620. https://doi.org/10.1016/j.molcel.2015.04.005.

[15] Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. V., Djebali, S., Niarchou, A., GTEx Consortium, Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., Ardille, K.G., & Guigó, R. (2015). Human genomics. The human transcriptome across tissues and individuals. *Science (New York, N.Y.)*, 348(6235), 660–665. https://doi.org/10.1126/science.aaa0355.

[16] Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Reviews. Genetics*, 12(2), 87–98. https://doi.org/10.1038/nrg2934.

[17] McGettigan, P. A. (2013). Transcriptomics in the RNA-seq era. *Current Opinion in Chemical Biology*, 17(1), 4–11. https://doi.org/10.1016/j.cbpa.2012.12.008.

[18] Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., & Moreno, R. F. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project.*Science (New York, N.Y.)*, 252(5013), 1651–1656. https://doi.org/10.1126/science.2047873.

[19] Marra, M. A., Hillier, L., & Waterston, R. H. (1998). Expressed sequence tags: ESTablishing bridges between genomes. *Trends in Genetics,* 14(1), 4–7. https://doi.org/10.1016/S0168-9525(97)01355-3.

[20] Putney, S., Herlihy, W. & Schimmel, P. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature,* 302, 718–721 (1983). https://doi.org/10.1038/302718a0.

[21] Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270(5235), 467–470. https://doi.org/10.1126/science.270.5235.467.

[22] Pozhitkov, A. E., Tautz, D., & Noble, P. A. (2007)."Oligonucleotide microarrays: widely applied—poorly understood". *Briefings in Functional Genomics & Proteomics,* 6(2), 141–148. https://doi.org/10.1093/bfgp/elm014.

[23] Meyers, B. C., Vu, T. H., Tej, S. S., Ghazal, H., Matvienko, M., Agrawal, V., Ning, J., & Haudenschild, C. D. (2004). Analysis of the transcriptional complexity of Arabidopsis thaliana by massively parallel signature sequencing. *Nature Biotechnology*, 22(8), 1006–1011. https://doi.org/10.1038/nbt992.

[24] Jongeneel, C. V., Delorenzi, M., Iseli, C., Zhou, D., Haudenschild, C. D., Khrebtukova, I., Kuznetsov, D., Stevenson, B. J., Strausberg, R. L., Simpson, A. J., & Vasicek, T. J. (2005). An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Research*, 15(7), 1007–1014. https://doi.org/10.1101/gr.4041005.

[25] Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J., & Corcoran, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, 18(6), 630–634. https://doi.org/10.1038/76469.

[26] Becker-Andre, M., & Hahlbrock, K. (1989). Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY). *Nucleic Acids Research*, 17(22), 9437–9446. https://doi.org/10.1093/nar/17.22.9437.

[27] Bainbridge, M. N., Warren, R. L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., Hickenbotham, M., Magrini, V., Mardis, E. R., Sadar, M. D., Siddiqui, A. S., Marra, M. A., & Jones, S. J. (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, 7, 246. https://doi.org/10.1186/1471-2164-7-246.

[28] Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J. C., & Ware, D. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature Communications*, 7, 11708. https://doi.org/10.1038/ncomms11708.

[29] Kumpatla, S. P., Buyyarapu, R., Abdurakhmonov, I. Y., & Mammadov, J. A. (2012) Genomics-assisted plant breeding in the 21st century: technological advances and progress. In: Ab-durakhmonov I (ed) Plant breeding. InTech publishers, Available from http://www.intechopen.com/books/plant-breeding/genomics-assisted-plant-breeding-in-the-21st-centurytechnological-advances-and-progress.

[30] Pérez-de-Castro, A. M., Vilanova, S., Cañizares, J., Pascual, L., Blanca, J. M., Díez, M. J., Prohens, J., & Picó, B. (2012). Application of genomic tools in plant breeding. *Current Genomics*, 13(3), 179–195. https://doi.org/10.2174/138920212800543084.

[31] Mironova, V. V., Weinholdt, C., & Grosse, I. (2015). "RNA-seq data analysis for studying abiotic stress in horticultural plants," in Abiotic Stress Biol. Hortic. Plants, eds Y. Kanayama and A. Kochetov Tokyo: Springer.197–220. https://doi.org/10.1007/978-4- 431-55251- 2_14.

[32] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), 376–380. https://doi.org/10.1038/nature03959.

[33] Oshlack, A. and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, 4, 14. **https://doi.org/**10.1186/1745-6150-4-14.

[34] Tuerk, A., Wiktorin, G., & Güler, S. (2017). Mixture models reveal multiple positional bias types in RNA-Seq data and lead to accurate transcript concentration estimates. *PLoS Computational Biology*, 13(5), e1005515. https://doi.org/10.1371/journal.pcbi.1005515.

[35] Han, Y., Gao, S., Muegge, K., Zhang, W., & Zhou, B. (2015). Advanced Applications of RNA Sequencing and Challenges. *Bioinformatics and Biology Insights*, 9(Suppl 1), 29–46. https://doi.org/10.4137/BBI.S28991.

[36] Barbazuk, W. B., Emrich, S. J., Chen, H. D., Li, L., & Schnable, P. S. (2007). SNP discovery via 454 transcriptome sequencing. *The Plant journal : for cell and molecular biology*, 51(5), 910–918. https://doi.org/10.1111/j.1365-313X.2007.03193.x.

[37] Cloonan, N., Forrest, A., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S.,Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., McKernan, K. J., & Grimmond, S. M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, 5(7), 613–619. https://doi.org/10.1038/nmeth.1223.

[38] Emrich, S. J., Barbazuk, W. B., Li, L., & Schnable, P. S. (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, 17(1), 69–73. https://doi.org/10.1101/gr.5145806.

[39] Holt, R. A., & Jones, S. J. (2008). The new paradigm of flow cell sequencing. *Genome Resrarch*, 18(6), 839–846. https://doi.org/10.1101/gr.073262.107.

[40] Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, 133(3), 523–536. https://doi.org/10.1016/j.cell.2008.03.029.

[41] Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., & Marra, M. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques*, 45(1), 81–94. https://doi.org/10.2144/000112900.

[42] Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I., & Marden, J. H. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, 17(7), 1636–1647. https://doi.org/10.1111/j.1365-294X.2008.03666.x.

[43] Deshpande, D., Chhugani, K., Chang, Y., Karlsberg, A., Loeffler, C., Zhang, J., Muszyńska, A., Munteanu, V., Yang, H., Rotman, J., Tao, L., Balliu, B., Tseng, E., Eskin, E., Zhao, F., Mohammadi, P., P Łabaj, P., & Mangul, S. (2023). RNA-seq data science: From raw data to effective interpretation. *Frontiers in Genetics*, 14, 997383. https://doi.org/10.3389/fgene.2023.997383.

[44] Costa, V., Angelini, C., De Feis, I., & Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-seq. Journal of*Biomedicine and Biotechnology*, 2010(5757), e853916. https://doi.org/10.1155/2010/853916.

[45] Teng, M., Love, M. I., Davis, C. A., Djebali, S., Dobin, A., Graveley, B. R., Li, S., Mason, C. E., Olson, S., Pervouchine, D., Sloan, C. A., Wei, X., Zhan, L., & Irizarry, R. A. (2016). A benchmark for RNA-seq quantification pipelines. *Genome Biology*, 17, 74. https://doi.org/10.1186/s13059-016-0940-1.

[46] Baruzzo, G., Hayer, K., Kim, E., Camillo, B. D., FitzGerald, G. A., & Grant, G. R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods*,14, 135–139. https://doi.org/10.1038/nmeth.4106.

[47] Robasky, K., Lewis, N. E., & Church, G. M. (2014). The role of replicates for error mitigation in next-generation sequencing. *Nature reviews. Genetics*, 15(1), 56–62. https://doi.org/10.1038/nrg3655.

[48] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, 17(1), 10–12. https://doi.org/10.14806/ej.17.1.200.

[49] Dodt, M., Roehr, J., 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.

[50] 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.

[51] Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics (Oxford, England)*, 33(14), 2202–2204. https://doi.org/10.1093/bioinformatics/btx153.

[52] Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communication*, 11(1), 1432. https://doi.org/10.1038/s41467-020-14998-3.

[53] Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, 21(1), 245. https://doi.org/10.1186/s13059-020-02134-9.

[54] Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8, 469–477. https://doi.org/10.1038/nmeth.1613.

[55] Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10, R25. https://doi.org/10.1186/gb-2009-10-3-r25.

[56] Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11), 1851–1858. https://doi.org/10.1101/gr.078212.108.

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

[58] Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9), 1105–1111. https://doi.org/10.1093/bioinformatics/btp120.

[59] 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), e178. https://doi.org/10.1093/nar/gkq622.

[60] Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)*, 26(7), 873–881. https://doi.org/10.1093/bioinformatics/btq057.

[61] 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 (Oxford, England)*, 29(1), 15–21. https://doi.org/10.1093/bioinformatics/bts63.

[62] García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F., & Conesa, A. (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics (Oxford, England)*, 28(20), 2678–2679. https://doi.org/10.1093/bioinformatics/bts50.

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

[64] Picard. http://picard.sourceforge.net/.

[65] Adiconis, X., Borges-Rivera, D., Satija, R., DeLuca, D. S., Busby, M. A., Berlin, A. M., Sivachenko, A., Thompson, D. A., Wysoker, A., Fennell, T., Gnirke, A., Pochet, N., Regev, A., & Levin, J. Z. (2013). Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nature methods*, 10(7), 623–629. https://doi.org/10.1038/nmeth.2483

[66] Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., 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), e140. https://doi.org/10.1093/nar/gkv711.

[67] DeLuca, D. S., Levin, J. Z., Sivachenko, A., Fennell, T., Nazaire, M. D., Williams, C., Reich, M., Winckler, W., & Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics (Oxford, England)*, 28(11), 1530–1532. https://doi.org/10.1093/bioinformatics/bts196.

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

[69] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17, 13. https://doi.org/10.1186/s13059-016-0881-8.

[70] Yang, I. S., & Kim, S. (2015). Analysis of Whole Transcriptome Sequencing Data: Workflow and Software. *Genomics Inform*,13(4), 119-125. https://doi.org/10.5808/GI.2015.13.4.119.

[71] 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., Kamoh, B., & 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.

[72] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., & Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7), 644–652. https://doi.org/10.1038/nbt.1883.

[73] 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 (Oxford, England)*, 28(8), 1086–1092. https://doi.org/10.1093/bioinformatics/bts094.

[74] Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11, 94. https://doi.org/10.1186/1471-2105-11-94.

[75] Hansen, K. D., Brenner, S. E., & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38(12), e131. https://doi.org/10.1093/nar/gkq224.

[76] Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010a). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616.

[77] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5), 511–515. https://doi.org/10.1038/nbt.1621.

[78] Anders, S. & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11, R106. https://doi.org/10.1186/gb-2010-11-10-r106.

[79] Li, J., & Tibshirani, R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical methods in medical research*, 22(5), 519–536. https://doi.org/10.1177/0962280211428386.

[80] Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., 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 (Oxford, England),* 29(8), 1035–1043. https://doi.org/10.1093/bioinformatics/btt087.

[81] Robinson, M. D., & Oshlack, A. (2010b). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25.

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

[83] 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.

[84] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids research*, 25(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

[85] Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl15.

[86] Finn, R. D., Alex, B., Jody, C., Penelope, C., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2014). Pfam: The protein families database. *Nucleic Acids Research*, 42 (D1), D222–D230. https://doi.org/10.1093/nar/gkt1223.

[87] Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution.*Nucleic Acids Research*, 28(1), 33–36. https://doi.org/10.1093/nar/28.1.33.

[88] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. The gene Ontology consortium. *Nature Genetics*, 25(1), 25–29. https://doi.org/10.1038/75556.

[89] Minoru, K., Goto, S., Kawashima, S., Okuno, Y., & Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(Database issue), D277–D280. https://doi.org/10.1093/nar/gkh063.

[90] Rolf, A., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., & Yeh, L. S. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, 32(Database issue), D115–D119. https://doi.org/10.1093/nar/gkh131.

[91] Deng, Y., Jianqi, L. I., Songfeng, W. U., Zhu, Y., Chen, Y., & He, F. C. (2006). Integrated nr database in protein annotation system and its localization. *Computer Engineering,* 32 (5), 71–72. https://doi.org/10.1109/INFOCOM.2006.241.

[92] Zheng, Q., Xiu-Jie Wang, (2008). GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Research*, 36(2), W358–W363. https://doi.org/10.1093/nar/gkn276.

[93] Zhou, X., & Su, Z. (2007). EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. *BMC Genomics*, 8, 246. https://doi.org/10.1186/1471-2164-8-246.

[94] Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C., & Weinstein, J. N. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4, R28. https://doi.org/10.1186/gb-2003-4-4-r28.

[95] Beissbarth, T., & Speed, T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics (Oxford, England),* 20(9), 1464–1465. https://doi.org/10.1093/bioinformatics/bth088.

[96] Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., & Jacq, B. (2004). GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biology*, 5(12), R101. https://doi.org/10.1186/gb-2004-5-12-r101.

[97] Alexa, A., Rahnenfuhrer, J., & Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13), 1600-1607. https://doi.org/10.1093/bioinformatics/btl140.

[98] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102.

[99] Huang, daW., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1), 44-57. https://doi.org/10.1038/nprot.2008.211.

[100] Ye, J., Zhang, Y., Cui, H., Liu, J., Wu, Y., Cheng, Y., Xu, H., Huang, X., Li, S., Zhou, A., Zhang, X., Bolund, L., Chen, Q., Wang, J., Yang, H., Fang, L., & Shi, C. (2018). WEGO 2.0: a web tool for analyzing and plotting GO annotations. *Nucleic Acids Research*, 46(W1), W71–W75. https://doi.org/10.1093/nar/gky400.

[101] Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1), 27–30. https://doi.org/10.1093/nar/28.1.27.

[102] Xiong, H., Guo, H., Xie, Y., Zhao, L., Gu, J., Zhao, S., Li, J., & Liu, L. (2017). RNAseq analysis reveals pathways and candidate genes associated with salinity tolerance in a spaceflight-induced wheat mutant. *Sci Rep*, 7, 2731 https://doi.org/10.1038/s41598-017-03024-0.