J. Appl. Res Plant Sci. Vol. 3(2), 236-241, 2022 eISSN 2708-3004; pISSN: 2708-2997
www.joarps.org https://doi.org/10.38211/joarps.2022.3.2.29

RNA-Seq as an Effective Tool for Modern Transcriptomics, A Review-based Study

Mekibib Million[1], Tileye Feyissa[1]

Addis Ababa University Institute of Biotechnology
***Corresponding author:** mekibib.million@gmail.com
Article Received 04-07-2022, Article Revised 12-07-2022, Article Accepted 30-07-2022

Abstract

Transcriptome analysis is a useful method for identification and understanding of genes. Finding genes that are differentially expressed between conditions is a crucial aspect of transcriptomics. The discovery of RNA seq has revolutionized next-generation sequencing technology. The fact that RNA sequencing does not requires gene probes and provides a precise measure of gene expression over a much wider range proved its credibility over other common techniques. The expressed gene profile and transcriptome data are stored in a database and could be accessed freely. During RNA seq short read mapping to the reference transcriptome (the set of all known transcript RNA sequences for a species) or genome in the database, a variety of database search tools and alignment methods become visible. There are a variety of applications that help align short reads generated by fragment sequencing. The study of expressed genes is aided by quantifying reads that align to the reference genome or transcriptome. RNA sequencing gives crucial information regarding alternative splicing and gene isoforms, in addition to differential gene expression.

Keywords: expressed gene analysis, high throughput transcriptomics, read mapping and RNA-seqmethod,

Introduction

Global profiling of mRNA expression of a specific tissue yields information about the transcriptional changes between defined states of tissues (Fakrudin *et al*., 2012). The transcriptome, a fundamental connection between DNA and phenotype is the first product of genome expression. It refers to a species' entire repertoire of transcripts (Rani and Sharma, 2017). The information contained in a gene within the nucleus is used to synthesize polypeptides in the cytoplasm (ribosome). The DNA of a gene on the other hand, does not directly participate in the synthesis of a polypeptide; instead, another molecule transports the gene's information or message out of the nucleus and into the cytoplasm. The initial stage in this information transfer from DNA to polypeptide is transcription, which involves synthesizing this messenger (mRNA) from the gene. It was a difficult task to investigate the genome behaviors in previous days because of the intricacy of the transcriptome of living cells. Prior to the development of high-throughput transcriptomics technologies, several individual transcript studies were carried out to investigate the genes (Lowe *et al*., 2017). History of complementary DNA data back to 1970s. The silk moth was the first novel organism from which mRNAs were successfully extracted, reverse transcribed getting cDNA libraries (Pandit *et al*., 2018). Low-throughput Sanger sequencing took over in the 1980s to sequence random individual transcripts from these libraries, referred to as expressed sequence tags (ESTs), which were used to swiftly categorize expressed genes and gene fragments (Lowe *et al*., 2017). The method's high sequencing cost, however, hindered its application in expression analysis, and Velculescu *et.al* (1995) created the Serial Analysis of Gene Expression (SAGE) approach, which greatly reduced the price. Only a brief tag section per cDNA (15 bp for the short SAGE approach and 21 bp for the long SAGE approach) is sequenced in this method (Hrdlickova *et al*., 2016). However, in the mid-1990s, DNA microarray technology surpassed EST and SAGE approaches for gene expression research, owing to its substantially lower cost for large-scale studies. Since their discovery, DNA microarrays have been widely used to determine the amount of mRNA corresponding to various genes. DNA microarrays have become common instruments for profiling of expressed gene because they enable the study of the mRNA levels of a large number of genes in a single test (Stahl *et al*., 2011). A huge number of genes are organized in a compact and uniform manner on a single microarray chip. Thousands of distinct oligo nucleotides can be immobilized on a single slide due to the small size of the spots. Each of these "probes" binds to a complementary nucleic acid ("target") isolated from the test and/or reference sample. In a single experiment, comparing the binding efficiencies of two samples allows for an easy and fast assessment of transcript level variations of gene for a large number of genes (Stahl *et al*., 2011). A recent technology known as RNA-Seq, similar to microarrays, is a tool for simultaneously detecting and quantifying all of the transcripts in a given sample. This method is based on next-generation sequencing, also known as deep sequencing. The RNA seq techniques enable quick, parallel sequencing of millions of DNA fragments and can thus be used to sequence all reverse-transcribed RNAs' in a given sample, in addition to genomic DNA. RNA-Seq is more sensitive than microarrays and has a considerably broader spectrum of gene expression that can be correctly assessed.

RNA –Seq tool for analysis of gene expression: RNA

sequencing (RNA-Seq) or next-generation sequencing (NGS) has appeared as a transformative tool in genetics, genomics, and epigenomics, with great sensitivity for finding transcription from scratch /splice junctions and short RNAs (Rani & Sharma, 2017). RNA-Seq is a newly emerged approach with excellent reproducibility and precision (Rani and Sharma, 2017), yet it has already revealed previously unknown details about the transcriptional complexity of a wide range of organisms, including yeast (Nagalakshmi *et al.*, 2008), mice (Mortazavi *et al.,* 2008), Arabidopsis (Eveland *et al.*, 2008), and humans (Sultan *et al.,* 2008). All of the mRNA would be retrieved and reverse-transcribed into cDNA to determine all of the protein-coding genes that were expressed in a given set of cells under specified physiological conditions. This phase is carried out in a similar manner as sample preparation for microarrays. Complementary DNAs, on the contemporary, should be fragmented into smaller pieces first, with small sequencing adapters attached to both ends. The fragments are then submitted to sequencing at a high rate in order to obtain short sequences from them. The data from these reads is matched against the genome sequence and used to determine the degree of expression of certain genes. SAGE and RNA-seq are examples of sequence-based transcriptome investigations that employ various sequencing systems (Ballereau *et al.*, 2013). SAGE was an EST methodology modification that increased the throughput of the tags created while also allowing for some quantification of transcript abundance (Lowe *et al*., 2017). The cDNA is digested into tags by restriction enzyme, and these tags link together head-to-tail to form long strands of >500 bp are sequenced using low throughput but long read length methods like Sanger sequencing (Lowe *et al.,* 2017). Individual tags separate from constituents after sequencing and align with the reference genome to identify genes. If the reference genome is unavailable, the tags can be utilized as diagnostic markers if they are shown to be differently expressed. RNA-Seq is a method for capturing and quantifying transcripts contained in an RNA extract that combines high-throughput sequencing with computational approaches (Lowe *et al.*, 2017). The nucleotide sequences generated are typically around 100 bp in length, although depending on the sequencing technology employed, they can range from 30 bp to over 10,000 bp. The accuracy, throughput, and read length of RNA-seq have all improved over time, thanks to the advent of NGS systems (Lowe *et al.*, 2017). Unlike Sanger sequencing, which relies on capillary electrophoresis, NGS methods rely heavily on large parallel sequencing, high-resolution imaging, and complex algorithms to deconvolute signal data into sequence data (Hrdlickova *et al.*, 2016). Based on the number of bases that can be sequenced in a single sequencing reaction, current NGS technologies can be divided into two categories: long and short read length technologies. The majority of NGS methods monitor millions of sequencing reactions in parallel, resulting in enormous amounts of sequencing data (Fakrudin *et al.*, 2012). Regardless of the ability to execute parallel sequencing reactions and huge amounts of data output, all next-generation sequencing technologies have their own limitations. Preferences for NGS technologies are mostly determined by the final data sets and analysis goals. Several technologies are currently available for high-throughput DNA molecule sequencing. Applied Biosystems (ABI SOLiD), Roche (454), Illumina (Genome Analyzer I/II and Hiseq), and Pacific Biosciences (single molecule real time) sequencing are just a few examples. Different technologies necessitate different experimental methods, but the most common one, which is employed with Illumina machines, typically includes the following steps: RNA isolation, fragmentation, cDNA synthesis, adaptor ligation, PCR amplification, and synthesis sequencing (Teresa & Gon, 2012).

Isolation of mRNA: Isolation of high-quality mRNA from RNA pools takes precedence over the subsequent downstream cascade of RNA sequencing. However, because the relative population of mRNA in the pools is only about 1-5 percent, special selection approaches are required. Because mature protein coding mRNA contains a poly-A tail, polyadenylated RNA selection is arguably the most prevalent use (L. Wang *et al*., 2010). Magnetic or cellulose beads coated with oligo dT molecules can be used to select poly-A + RNA. Polyadenylated RNAs can also be chosen for reverse transcription utilizing oligo-dT priming(L. Wang *et al*., 2010). Depletion of rRNA using sequence-specific probes that can hybridize to rRNAs' is another method of mRNA enrichment. Biotinylated DNA or locked nucleic acid (LNA) probes are used to hybridize unwanted rRNAs or their cDNAs, and then streptavidin beads are used to deplete them. Antisense DNA oligoscan also be used to target rRNAs, which is known as probe-directed degradation (PDD) (Archer *et al.*, 2015). While less time-consuming than hybridization, this method necessitates constant rRNA coverage and distinct probe sets for each species.

Library preparation: The generation of the library is an important step in RNA-seq since it influences how closely the cDNA sequence data reflects the original RNA population. The simplest method is to make double-stranded cDNA and ligate the adaptor to it (He *et al*., 2008). It is crucial to start with a population of intact mRNAs in order to make high-quality cDNAs. Several hundred Adenine bases are found at the 3' terminus of most eukaryotic mRNAs. This poly-A tail can trap these RNAs while also removing contaminated rRNAs, tRNAs, and other small cytoplasmic and nuclear RNAs. Reverse transcriptase can be used with an oligo dT primer to produce a DNA copy of the mRNA strand (Gong *et al*., 2020). A specific mRNA or class of mRNAs' can be detected by using random primers. There are two common RNA-Seq experimental protocols: (a) single

end and (b) paired end sequencing procedures. Nucleotide molecules of 50 to 100 bp in length or 200 to 400 bp in length can be sequenced from one end or both ends (Z. Wang *et al*., 2010). Figure 1 depicts every possible method for cDNA library preparation for NGS (L. Wang *et al.*, 2010).

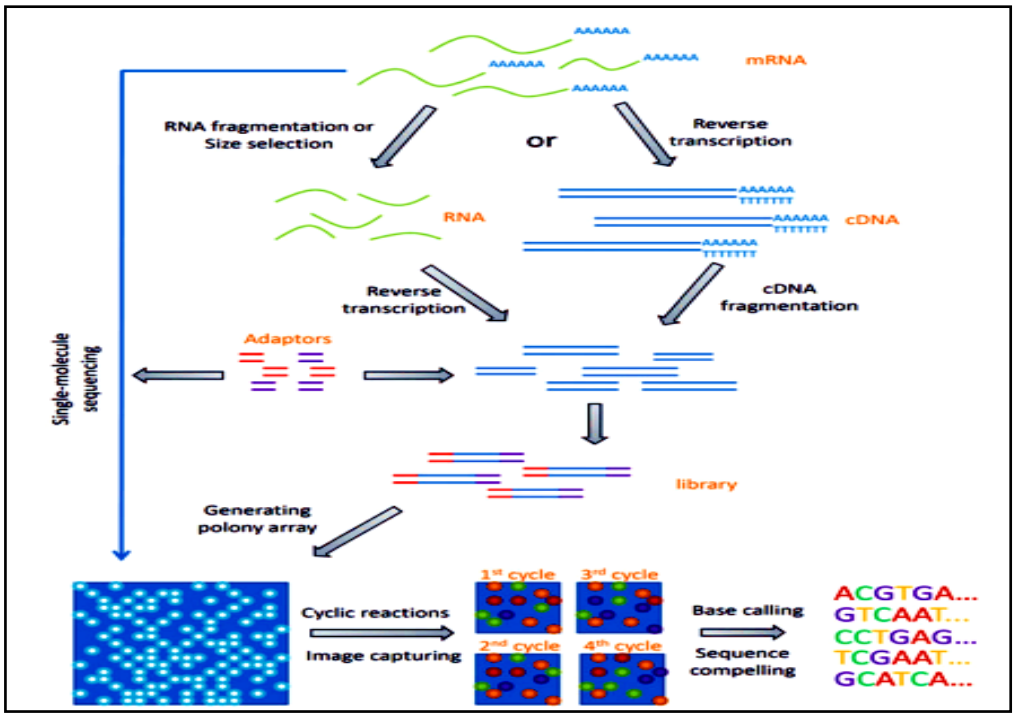

Figure 1. RNA Seq experimental procedure for NGS:

Amplification: Because most sequencers have a detection limit, cDNA libraries must be amplified using PCR before being sequenced. PCR amplification may be used to amplify the cDNA library for multiple reads, depending on the type of RNA-seq used. Except for Pacific Biosciences' (PacBio) single molecule sequencer; most NGS systems necessitate library amplification before sequencing. PCR amplification provides quantitative results of RNA expression; however, quantification of expression levels can be due to PCR as well as increased RNA expression. Thus comparison between multiple samples must be available in order to determine the effects of PCR on measured expression levels relative to the actual levels in the cell. Helicos sequencing does not require amplification of cDNA, and RNA sequencing methods are improving to reduce PCR effects. However, it's crucial to examine if PCR amplification will have an impact on the experimental construct.

Sequencing by synthesis: The final step is to identify each individual nucleotide that will be used in the sequences. Illumina uses the sequencing by synthesis approach that employs reversible terminators in which the four modified nucleotides, sequencing primers and DNA polymerases are added as a mix, and the primers are hybridized to the sequences (Kchouk *et al.,* 2017). After that, the primers are extended with polymerases using the modified nucleotides. To distinguish each nucleotide type, each is labeled with a fluorescent specific. Because the nucleotides have an inactive 3'-hydroxyl group, only one nucleotide is incorporated (Kchouk *et al*.,2017). A laser excites clusters, causing them to emit a light signal unique to each nucleotide, which is detected by a coupled charge device CCD camera and translated into a nucleotide sequence by computer programs. The process is repeated by removing the terminator with the fluorescent label and starting a new cycle with a new incorporation until the required size reads are synthesized (Stahl *et al*., 2011)

RNA-Seq data analysis: The initial step in data processing is to map the short reads from RNA-Seq to the reference genome, or to assemble them into contigs before aligning them to the genomic sequence to show transcription structure (Hrdlickova *et al.*, 2016). The most common goal of the RNA-Seq method of transcriptome analysis is to estimate expression of specific genomic areas, which could include genes, isoforms, exons, splice junctions, or newly transcribed regions. It is critical to map the reads received from synthesised sequences in order to achieve these goals.

Read mapping: The identification of which features are present in the sequencing library is the initial stage in mapping (Teresa and Gon, 2012). Because mapping

these short reads to the genome is difficult, three methods have been used: *de novo* assembly of reads, read alignment to the genome, and then assembly and read alignment to the transcriptome. Because there is no one method for read alignment, the choice of aligners will be influenced by the reference utilized (genome or transcriptome), the data type (short vs. longer reads), and the computational capacity available, among other things. Short read alignment is currently performed by a variety of techniques, and the aligner chosen is usually determined by the analytic goals and needs (Parada, 2018).

***De novo* assembly:** By utilizing read overlaps, the goal of *de novo* read assembly is to find a set of the longest possible contiguous expressed regions (contings) (Teresa and Gon, 2012). In recent years, three algorithmic solutions have been used to overcome the challenge of *de novo* assembly: prefix tree, overlap-layout-consensus, and de Bruijn graph (Wajid and Serpedin, 2012). Despite the fact that *de novo* read assembly is the most difficult of the three mapping procedures, it is the method of choice when a reference genome is not available or the annotation for the species in question is of poor quality (Dida & Yi, 2021).

Read alignment to reference genome or transcriptome: The readings can be mapped to either a genome or a transcriptome as a reference (the set of all known transcript RNA sequences for a species). Read alignment to the reference genome, on the other hand, provides the benefit of permitting the discovery of new genes and isoforms. Alternative splicing occurs concurrently with transcription and is mostly controlled by splicing factors (proteins) that bind RNA motifs (short lengths of RNA) in the pre-mRNA (Harri, 2020). The problem is solved by using read mapping tools that can detect alternative splices from the original transcript. TopHat (Trapnell *et al*., 2009), GSNAP (Wu and Nacu, 2010), QPALMA, and SOAPSplice are some of the alignment tools that can align spliced junctions of short reads(Nagalakshmi *et al.*, 2010).

Differential expression: Read counts in genes/exons are used to determine gene expression. The expression of genes in different samples is compared using gene expression analysis. The alignment result from RNA-seq provides the chromosome/position of each aligned read. There are reads aligned to the gene body for each gene; they can be summarized into a number for the expression by counting the number of reads aligned and normalizing by the total number of reads in the experiment, as well as gene lengths, if desired (L. Wang *et al*., 2010). The fragments per kilobase of transcript per million fragments mapped (FPKM) method is used to quantify expression(Mortazavi *et al.*, 2008), (Cseke *et al.*, 2003). The number of reads mapped to a gene determines the amount of expression. The amount of reads obtained from an expressed gene is proportional to the length of the transcript, with longer transcripts resulting in more fragments and hence more reads. In order to examine means and variances, differential expression normally requires numerous replicates per sample. Normalization for Library size is essential when comparing gene expression between groups of samples (number of reads obtained). Normalization for gene length and library size is performed by transforming counts to fragments per kilo base per million mapped reads (FPKM). As a result, the longer the transcript, the more reads in the library and the lower the likelihood of getting reads by accident (Smid *et al.*, 2018).

Alternative splicing: Alternative splicing is a transcript processing method that involves connecting distinct exons to produce distinct mRNA products from the same pre-mRNA. It has been reported that RNA-Seq was used for the first time to quantify splicing (Clark *et al*., 200), using a method similar to splicing junction arrays (Alamancos *et al.*, 2012). Exon arrays, on which the probes are tailored to target the junction areas, can be used to detect and quantify it. At the junctions of two exons, RNA-Seq detects "junction reads," which are reads that overlap two exons(Z. Wang *et al.*, 2010). When an RNA-Seq read crosses an exon border, however, a portion of the read does not map contiguously to the reference, causing the mapping step to fail. Reads that do not align to the genome but map to these synthetic pieces are evidence for splice junctions between recognized exons (Trapnell *et al*., 2009). Splice sites can be detected from the start by looking for reads that bridge exon junctions, however matching short reads to the genome is computationally difficult. Concatenating known neighboring exons and then synthesizing sequence fragments from these spliced transcripts solves this challenge (Orton *et al.*, 2016). Spliced junctions of short readings can be aligned using a variety of alignment algorithms (Trapnell *et al.*, 2013). TopHat, on the other hand, is a software program that uses large-scale mapping of RNA-Seq reads to identify splice sites ab *initio*. At a rate of 2.2 million reads per CPU hour, TopHat maps reads to splice sites in mammalian genomes (Trapnell *et al*., 2009). Figure 2 depicts all possible splice junction searches using splice map software(Au *et al.,* 2010).

Gene fusion: Two or more genes can be “fused” to generate a new gene. Microarrays are ineffective in detecting the phenomena. Reads from “paired-end” RNA-Seq, on the other hand, provide a wealth of information. Provided that pair of reads that are quite far apart on the reference genome indicates gene fusion. Because the DNA segments are selected based on the sizes, and there should not be long cDNA segments (Ozsolak & Milos, 2011).

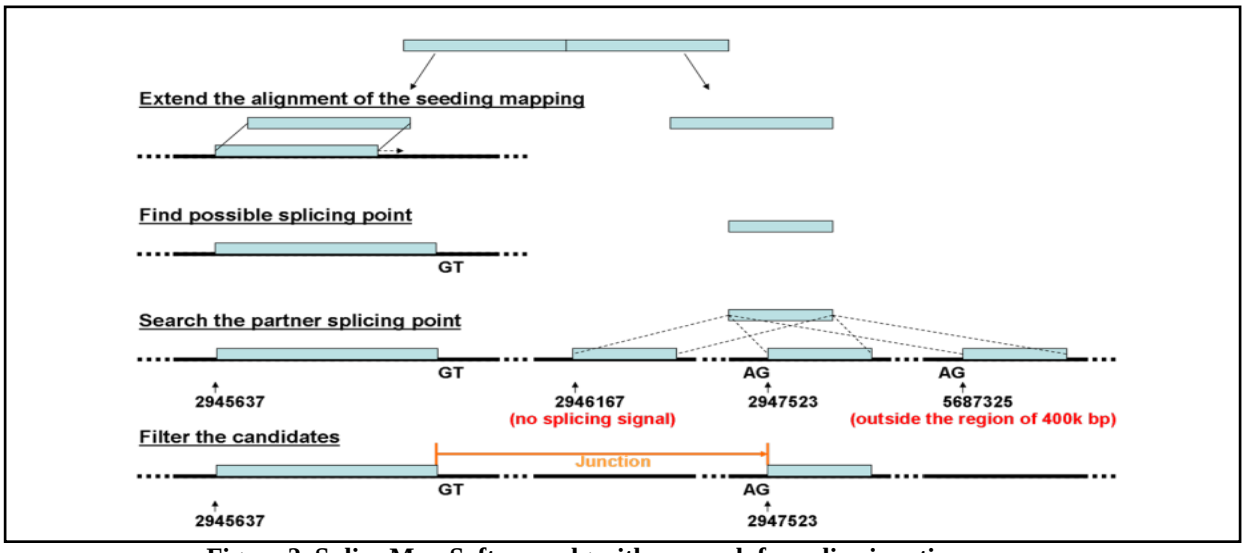

Figure 2. Splice Map Software algorithm search for splice junction.

Conclusion

Until Watson and Crick discovered the double helix structure of DNA in 1953, nothing was known about the genome. This discovery created the theoretical ground work for the human genome project. Before the advent of next-generation sequencing technologies, understanding the genome structure and function was a difficult task for scientists. The advancement of DNA and cDNA sequencing technologies from the transcriptome has aided in gene analysis. Major advances in sequencing generation evolution have resulted in the emergence of the most recent gene analysis technology, the RNA-Seq method. Beyond differential gene expression, the emergence of the RNA-Seq method has resulted in a significant paradigm shift in transcriptomics. It enabled scientists to comprehend and analyze various forms of gene expression, such as alternative splicing, isoforms, and fusions. The major challenges with the technology is, mapping the NGS 'read' to the reference transcriptome, even for well-studied species like the human and mouse, because transcriptomes are incomplete. As a result, RNA-Seq analyses must map to the reference genome as a transcriptome proxy. The ambiguous processes in preparation and protocols are the technology's limits. If any experimental or technological biases were introduced, a multi-level quality control check would be performed. As a result, technological advancements in the future; point to a very complete and simple method for high-throughput sequencing and data translation. However, bioinformatics and computational components, particularly short read aligner programs and algorithms for converting biological data into information, require further development in this area.

Acknowledgment

I would like to thank and appreciate Dr. Demisachew Guade for his invaluable, critical, and constructive comments and guidance during the preparatio. I also admire his meticulous review system that he used until the paper arrived at its current sound seminar body.

References

Alamancos, G. P., Agirre, E., & Eyras, E. (2012). *Methods to study splicing from high-throughput RNA Sequencing data*. *Table* 9, 1–31.

Archer, S. K., Shirokikh, N. E., & Preiss, T. (2015). Probe-directed degradation (PDD) for flexible removal of unwanted cDNA sequences from RNA-Seq libraries. Current Protocols in Human Genetics, 85(1), 11-15.Au, K. F., Jiang, H., Lin, L., Xing, Y., & Wong, W. H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Research*, **38**(14), 4570–4578.

Ballereau, S., Glaab, E., Kolodkin, A., Chaiboonchoe, A., Biryukov, M., Vlassis, N., ... & Auffray, C. (2013). Functional genomics, proteomics, metabolomics and bioinformatics for systems biology. In Systems Biology (pp. 3-41). Springer, Dordrecht.

Cseke, L. J., Wu, W., & Kaufman, P. B. (2003). DNA sequencing and analysis. *Handbook of Molecular and Cellular Methods in Biology and Medicine, Second Edition*, **2015**(11), 237–270.

Dida, F., & Yi, G. (2021). Empirical evaluation of methods for de novo genome assembly. PeerJ Computer Science, **7**, e636.

Fakrudin, B., Tuberosa, R., & Varshney, R. K. (2012). Omics techniques in crop research: An overview.

Gong, A. D., Lian, S. B., Wu, N. N., Zhou, Y. J., Zhao, S. Q., Zhang, L. M., ... & Yuan, H. Y. (2020). Integrated transcriptomics and metabolomics

analysis of catechins, caffeine and theanine biosynthesis in tea plant (Camellia sinensis) over the course of seasons. BMC plant biology, **20**(1), 1-14.

Hrdlickova, R., Toloue, M., & Tian, B. (2017). RNA-Seq methods for transcriptome analysis. Wiley Interdisciplinary Reviews: RNA, **8**(1), e1364.

Kchouk, M., Gibrat, J. F., & Elloumi, M. (2017). Generations of sequencing technologies: from first to next generation. Biology and Medicine, **9**(3).

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. PLoS computational biology, **13**(5), e1005457.

Mortazavi, A., Williams, B. A., & McCue, K. Schaffe er, L., and Wold, B.(2008). Mapping and quantifying mammalian transcriptomes by rna-seq. Nature methods, **5**(7), 621628.

Nagalakshmi, U., Waern, K., & Snyder, M. (2010). RNA-Seq: a method for comprehensive transcriptome analysis. Current protocols in molecular biology, **89**(1), 4-11.

Orton, R. J., Gu, Q., Hughes, J., Maabar, M., Modha, S., Vattipally, S., & Davison, A. (2016). Bioinformatics tools for analysing viral genomic data. Revue scientifique et technique (International Office of Epizootics), **35**(1), 241-285.

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. Nature reviews genetics, **12**(2), 87-98.

Pandit, A. A., Shah, R. A., & Husaini, A. M. (2018). Transcriptomics: A time-efficient tool with wide applications in crop and animal biotechnology. J Pharmac Phytochem, **7**, 1701-1704.

Rani, B., & Sharma, V. K. (2017). Transcriptome profiling: methods and applications-A review. Agricultural Reviews, **38**(4). 271-281

Smid, M., Coebergh van den Braak, R. R., van de Werken, H. J., van Riet, J., van Galen, A., de Weerd, V., & Sieuwerts, A. M. (2018). Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample comparisons. BMC bioinformatics, **19**(1), 1-13.

Stahl, F., Hitzmann, B., Mutz, K., Landgrebe, D., Lübbecke, M., Kasper, C., & Scheper, T. (2011). Transcriptome analysis. Genomics and Systems Biology of Mammalian Cell Culture, 1-25.

Teresa, A., & Gon, F. (2012). *RNA sequencing for the study of gene expression regulation*. *September*.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, **25**(9), 1105-1111.

Wang, L., Li, P., & Brutnell, T. P. (2010). Exploring plant transcriptomes using ultra high-throughput sequencing. Briefings in functional genomics, **9**(2), 118-128.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews genetics, **10**(1), 57-63.