ISSN NO. 2320-5407

*Journal homepage: http://www.journalijar.com*

*INTERNATIONAL JOURNAL OF ADVANCED RESEARCH*

**RESEARCH ARTICLE**

# The Need behind Messenger RNA Sequencing Analysis

**Shilu Mathew[1], Manal Shaabad[1], Shireen Hussein[1], Lobna Mira[1]and Ishtiaq Qadri[2*]**

**1.** Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, KSA

**2.** Medical Biotechnology and TranslationalMedicine Research, King Fahd Medical Research Center, King Abdul Aziz University,

## *Manuscript Info*

***Manuscript History:***

Received: 18 April 2015
Final Accepted: 22 May 2015
Published Online: June 2015

***Key words:***

mRNA, Sequencing,Next generation sequencing, Bioinformatics

***Corresponding Author***

**Shilu Mathew**

## *Abstract*

The current human genome data analysis has revealed that only small fraction of about 1.5% codes for protein when compared to higher eukaryotic genomes available. Most of the genomic DNA is involved in the regulation of gene expression; controlling gene is transcription, post-transcriptional level including controlling the fate of the transcribed RNA molecules, including their stability, the efficiency of their translation and their localization. Messenger RNA (mRNA) has crucial roles in many aspects of gene regulation. Next-generation sequencing (NGS) technologies have enlightened advantages in terms of cost-effectiveness, extraordinary sequencing speed, high resolution and precision in genomic studies. Currently, these high-throughput sequencing technologies have been generally employed in a variable ways, such as whole genome sequencing, gene expression profiling, targeted sequencing, chromatin immunoprecipitation sequencing as well as small RNA sequencing, to facilitate biological and biomedical research. NGS enables the classification of all RNA transcripts for a given organism (e.g. human, mouse and rat), comprising both the coding mRNA and non-coding RNA, such as snRNAs and snoRNAs, longer than 170 nucleotides in length, irrespective of whether they are polyadenylated or not. However, the huge amount of data created by NGS signifies an excessive challenge. This article talks about the present guidelines for the need behind mRNA sequencing, existing applications of NGS technologies and for selecting suitable tools in genomics, transcriptomics and small RNA research analysis.



## INTRODUCTION

RNA-seq technique is the latest approach to understand transcriptome profiles using high throughput-sequencing. This advanced technique provides precise details of the transcripts as well as their isoforms compared to other methods. In general, the extracted RNA is converted to cDNA fragments attached with adaptors in both the ends. These fragments are sequenced in high throughput with or without amplification to obtain short sequences using single-end or paired-end sequencing. Depending on the size of the reads, the DNA-sequencing instrument are as follows such as Illumina IF[1], Applied Bio systems SOLiD5500[2] and Roche Life sciences[3] which are commonly used for RNA-seq analysis. The resulting reads obtained after sequencing are either compared to the reference transcripts or compared to de novo to synthesis a genome-scale transcription map.[4]

The arrival of the deep sequencing techniques have changed the way in which the genome transcription map is studied. RNA-seq involves direct sequencing of the converted cDNA using DNA sequencing followed by mapping of sequencing reads.[5] RNA-seq provides more comprehensive understanding for the identification of

introns and exons, 3' and 5' ends of genes and mapping their boundaries; identification of splicing site, transcription start sites and finally it allows understanding the precise quantification of splicing variant and exon expression.[2,6] This review article briefs about the advances made so far and the challenges faced with its usage in various eukaryotic transcriptomes.

**2. Experimental Design for RNA-Seq**

RNA sequencing methods can be used to detect RNAs in very small highly purified pools of RNA, such as those cross-linked to a protein of interest in a "CLIP-Seq" experiment. Quantity of the starting material is the major criteria of any high throughput-sequencing. For transcriptome studies, a few hundred nanograms of oligo (dT)-selected RNA can serve as the starting material. Oligo dT-selected RNA enrichment of Poly(A)+ mRNA using immobilized Oligo (dT) serves as a starting material for preparing mRNA-Seq libraries for transcriptome studies.[7] Thereby small amounts of input RNA are enough to generate millions of reads. RNA sequencing captures sequences mostly with respect to their presence in the sample as total RNA sample,which provides maximum RNA reads thereby reducing accuracy of quantification and sensitivity of detection particularly for lower-abundance transcripts.[8]

Library construction is the next stage for the preparation of small RNA libraries for high throughput-sequencing[9] and can be prepared by a variety of methods such as fragmentation of whole-transcriptome RNA using *E. coli* RNase III[8], preparation of small RNA libraries for high-throughput sequencing[5] and as well as by other commercial kits. Many transcriptome libraries are prepared from cDNA that is primed with random hexamers.[10] Library preparation can be modified to capture transcript ends selectively, to capture reads that represent polyadenylated 3′ ends[11] or capped 5′ ends.[12]

Library preparation requires amplification of an originating set of RNA-derived cDNA fragments by PCR. Amplification is done by PCR after library preparation from original set of RNA derived from cDNA fragments. This may create bias if some fragments are amplified more efficiently compared to others. This problem may be more severe for longer-fragment libraries. Size selection before amplification is an important criterion to ensure that the population of fragments is efficiently replicated, thereby avoiding over amplification or loss of fragments.[13]

Paired ends are one option available under most sequencing protocols to obtain paired end reads. It requires that the library to be constructed is done by linking with linkers. It also involves capturing read data from individual reads or fragments by sequencing with primer from one end to the other, regenerating single-stranded DNA and sequencing a second time with a primer from the other end of the fragment DNA. Thus two sequences reads are obtained from individual seed molecule from library and hence it greatly improves mapping.[14]

Barcoding allows multiple libraries to be sequenced in the same sequencing run or lane. During individual library construction, discrete linker (particular barcode which constitutes small nucleotide) sequence is attached to each library. These barcodes can be read with high confidence through high throughput-technologies.[8]

Read lengths were used in the early days of high throughput-sequencing, were short offering 25-32 high confidence base calls with which to determine the genomic origin of the sequence. In even repeated regions of the genome, the short length of reads meant that many could not be mapped to a single region. The bottom line is that all of these problems decrease away rapidly as read length increases; in particularly, the identification of splice junction reads improves greatly even in the absence of paired-end data.[14,15]

**4. Mapping and processing sequencing reads**

Most studies have sequencing conducted at a core facility, where they send RNA samples for library construction, and sequencing and receive a huge data per experiment. A high throughput-sequencing instrument uses internal computational methods to process raw read which is necessary to determine base identity. At the end of each run, each platform produces a system folder that includes a base call plus a quality score at each read position, which indicates whether the confidence of the base call at that raw read position is correct or not. Raw reads are trimmed so that low-confidence bases, usually at the end of the read, are removed. If the sequencing run has been derived from a mixture of bar-coded libraries, the next step is to sort the reads by their bar codes into separate files that contain only reads from a given input library.[8] Mapping is obtained after processing the reads which is mapped to a reference genome or assembled to de novo. Bioconductor website can be used to download various latest software updates for bioinformatics methods for reviewing the captured map reads and analyse the map splice

junction reads.[14,15] For discovery of novel transcripts, mapped reads must be assembled into transcripts which are obtained by using commonly used three programs such as Bowtie, Tophat and Cufflinks.[8,16]

Measuring the level of gene expression are also other mapping problems that are evolving. The expression and the measure in the change of expression is detected and compared, a list of gene expression or RNA-processing changes by gene or genomic location, denoting the number of the RNA processing changes by genomic location, along with the magnitude of change as well as to make sure the measure of the e likelihood that the change is not due to chance, is obtained. It is important to determine and understand these changes as how the data stands up to an orthogonal method.[13,17] Further validation of a gene expression in a high throughput data is validated only by direct experiment with the individual gene of interest. Validation is important to find the favorite gene that truly changed the experiment as it leads to discovery of specific biological process of insights and on the other hand to use bioinformatics to validate the results hold the true best scoring genes which requires high confidence set of gene for analysis.[18] Techniques such as RT-PCR or independent measurement has to be further used to predict the true changes observed by more labor intensive approach. Studies have also used microarrays for cross validation with RNA sequencing that has indicated good agreement between them for validation of the analyzed sequenced data.[8,18]

**3. Transcriptome: the key to understanding gene activity**

In order to unravel the link between genome and cells functioning, researchers sought to conduct large scale proteomics studies on proteins being products of the expressed genome. Knowing that proteomics is a comprehensive study of a proteome that gives details on the diversity and quantity of proteins, it can be easily presumed that it holds the key to the mystery. However, the dynamic nature of proteins along with the fact that they are co- and post-translationally modified and cannot be amplified easily makes proteomics a challenging approach in most cases and adds a layer of technical difficulty to it. Thus, in order to bridge the gap between the genome and the functional molecules of the cells in a less complex approach, research focus of many studies has been shifted to studying transcripts which are considered an intermediate step between the genome and the genes that subsequently encode for either a protein or a non-coding RNA. The complete set of RNA transcripts within a cell is known as transcriptome.[19-21] Figure 1 denotes the key to understand transcriptome in the RNA world.

Although almost all the cells in multicellular organisms share the same genome, yet this does not apply to their gene expression patterns. In other words, in a cell, some genes are transcriptionally active while the others are not, which emphasizes a wide range of functional, biochemical, developmental and physical variations amid different cells and tissues. Theses variations potentially differentiate between health and disease status of the cells and thus the organism. Hence, studying the transcriptome of varying cell types and tissues may explain the potential contribution of cells transcriptional activity and diseases. The proportion of transcriptome which represent the percentage of genetic code that is transcribed into RNA is estimated to comprise less than 5% of the human genome, while the proportion of transcribed sequences that does not code for protein (non-coding RNA) seems to increase in more complex organisms.[19] Moreover, process such as alternative splicing, RNA editing or alternative transcription initiation and termination sites may lead to producing multiple variants of mRNA for each gene. Thus, studying the transcriptome addresses a level of complexity that genome analysis does not.[22,23]

Studying the transcriptome may lead to understanding gene activity of various cell types (i.e. when and in which cell is a specific gene turned on or off). Furthermore, the possibility of quantifying the number of transcripts allows for measuring gene expression in a cell during a certain statues. In stem cells research, for instance, the transcriptome information may help to identify genes that contribute to stem cells unique immortality property and developmental elasticity. In addition, transcriptome analysis holds a great promise in discovering new gene's function. For example, if a gene with an unknown function was expressed in fat tissue while it was not expressed in other tissues (e.g. muscle or bone tissues), this indicates that this gene may play a role in fat metabolism or storage. Additionally, transcriptome analysis reveals gene expression changes associated with disease state that in some cases might be the driver of some severe diseases such as cancer.[20,24]

**4. Understanding transcriptome complexity through RNA structures**

The central dogma of molecular biology indicates that the genetic data is transferred from DNA to proteins through mRNA. This process is regulated via the action of multiple proteins. These proteins bind to the core and auxiliary flanking regions of the gene and their function differ in concordance with their binding location. Some

proteins are involved in events related to pre-mRNA processing whereas others are involved in gene splicing activities.

As discussed earlier, "transcriptome" indicates gene expression and transcriptional activity since it represents the RNA content of the cell which is considered as a bridge in the process of transferring genetic information between DNA and proteins.[25] The transcripts content comprising the transcriptome is classified into ribosomal RNA (rRNA) accounting for 80-90% of the transcriptome, transfer RNA (tRNA) which makes up between 5-15%, messenger RNA (mRNA) of 2-4% and only 1% of non- coding RNA (ncRNA) which represents both intronic and intergenic RNAs.[26] Non- coding sequences were originally considered to be what is termed as "junk" DNA which was assumed to be genetically inactive.[27] However, the fact that the proportion of non-coding DNA (ncDNA) increases in concordance with the complexity of the organism, i.e. the more complex an organism is, the more ncDNA its genome contains, supported the assumption that ncDNA is probably what creates biological complexity and diversity of organisms.[28] Studying possible correlation between DNA and the biological complexity of organisms has been always an important field of research.[29] Data from the ENCODE project and many other studies have shown that in eukaryotes, nearly the full length of non-repeat regions of the genome is being transcribed, thus, revealing the ubiquitous nature of transcription in eukaryotes.[30] Moreover, the discovery of small interfering RNA (siRNA), microRNA (miRNA), promoter and terminator-associated small RNA (PASR and TASR, resp.), Long interspersed noncoding RNA (lincRNA), transcription initiation RNA (tiRNA), transcription start site-associated RNA (TSSa-RNA) and many others unraveled an unexpected level of transcription complexity.[21] To add up an extra layer of complexity to the transcriptome puzzle, it has been found that most of the prevalent transcripts identified are found in particular cell lines, interestingly in mutant cell lines in most cases and in specific tissues. Thus, transcriptomics analysis is essential to understanding gene function and revealing the complexity of the genome by studying the molecular components of the cells. This in turn provides a comprehensive overview of many complex biological processes including disease onset and progression.[22,31,32]

**5. Analysing Transcriptome**

To interpret the functional features of the genome, to discover the molecular elements on a cellular level and to understand disease onset and progression it is crucial to understand the transcriptome. Transcriptomics is the study of transcriptomes, which is carried out with following aims: cataloguing all transcript species like non-coding RNAs, small RNAs and mRNAs,; determination of the splicing patterns and other post-transcriptional modifications, transcriptional genetic structure in accordance with their start sites, 5′ and 3′ ends; and finally measuring of the changes in expression levels of each transcript during development and under various conditions.

Hybridization or sequence-based approaches are some of the many technologies that have been established for understanding and quantifying the transcriptome. Usually in hybridization-based approaches the incubation of fluorescently labelled cDNA with custom-made microarrays or commercial high-density oligo microarrays is carried out. Arrays with probes spanning exon junctions are example of probes designed especially for this purpose, this probe may be used for detecting and quantifying distinct spliced isoforms.[33] Genomic tiling microarrays that are able to map transcribed regions to a very high resolution ranging up to 100 bp [34-37] i.e. microarrays that represent the genome at high density are also being created. Hybridization-based approaches are common and cost-effective with the exception of high-resolution tiling arrays used for interrogating large genomes. However this approach also entails certain drawbacks such as: dependence upon existing knowledge about genome sequence; high background levels owing to cross-hybridization[38,39] and because of background and saturation of signals a limited dynamic range of detection is observed (Figure 2 briefly describes a complete overview of RNA-seq experiment). Additionally, a difficulty is experienced in comparing expression levels across different experiments and may need complex normalization methods.

On the other hand, the cDNA sequence can be directly determined by the sequence-based approaches. Sanger sequencing of cDNA or EST libraries[40,41] was carried out at first but owing to low throughput and higher cost along with lack of the advantage of quantification of expression, new approaches had to be developed like Tag-based methods, which include cap analysis of gene expression (CAGE)[42,43], serial analysis of gene expression (SAGE)[44,45] and massively parallel signature sequencing (MPSS).[46-48]These tag-based approaches saliently provide precise, ‘digital’ gene expression levels and are high throughput yet these tag approaches are based on expensive Sanger sequencing technology, and a significant portion of the short tags cannot be uniquely mapped to the reference genome. Other limitations associated with this approach are that only one portion of the transcript is

studied and the isoform are quite similar to each other. The above mentioned reasons lead to limited use of traditional sequencing technology in interpreting the structure of transcriptomes.[4,49]

The recent development of novel high-throughput sequencing approaches has led to providing a single method that can be employed for both mapping and quantifying transcriptomes that is RNA Sequencing (RNA-Seq). This method overcomes all the limitation of old methods and is speculated to revolutionize the field of eukaryotic transcriptomes analysis. Transcriptomes of *Arabidopsis thaliana*, *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, mouse and human cells were already studied using RNA-Seq.[1,2,50-54] In the following chapter provides an overview about using next generation sequencing (NGS) technology in transcriptomics via RNA-Seq.

**6. Application of Next Generation Sequencing Technologies in Transcriptomics Analysis**

There is a growing need of rapid genome and transcriptome analysis technologies for the investigation of cellular state, physiology, and biological activity in fields of molecular biology, biotechnology, and medicine. Applying RNA-Seq using currently available NGS technologies is the premium answer to this emerging need in science as it opens doors to unlimited possibilities in modern bioanalysis.[22]

Investigation of transcriptome at a very high, unsurpassed resolution is made possible using RNA-Seq. The fact that RNA-Seq doesn't require prior knowledge of the sequence being investigated, plays as a major advantage as it enables us to study poorly characterized species such as that of Plasmodium.[55,56] Figure 3 denotes different methods that can be selected for transcriptome analysis. Additionally, gene expression analysis, studying alternative splice sites, allele specific expression and the identification and analysis of transcripts that are rare or novel all currently employ RNA sequencing being the state of the art technology.[22,57]

It is worth mentioning that RNA-Seq was developed in the year 2008 and has been initially used in analysing transcriptomes of yeast, mouse, and arabidopsis.[50-52] Short reads of transcript sequence information are generated, which are then assembled into complete transcripts (contigs) and mapped to the genome. There are varying designs for RNA sequencing experiment depending on the goals for which it was carried out. Currently RNA-Seq experiment might include various types of RNA isolated from whole cells or from specific sub-cellular compartments or biochemical classes (polyA+ RNA, polysomal RNA, nuclear ribosome-depleted RNA, various size fractions of RNA and a host of others).[1,58]

There are generally three main aims for which researchers perform RNA-Seq experiments, First of all, to know how many and which RNA transcripts are present in a cell or a sample as RNA-Seq enables us to count the relative number of transcripts made in each cell, explaining its function to some extent. Secondly, RNA-Seq enables us to identify the elements (DNA sequences) of the genome that are copied into RNA and identify their biological function by providing specific information for genome annotation. Previously, expressed sequence tags (ESTs) were used for genome annotation, this technique was based upon older Sanger-based sequencing technology, which is now being replaced with newer second-generation techniques such as those are provided by Illumina, life technologies and others. One of the basic advantages of the second generation techniques is ease of access to the information-dense transcriptome. Finally, by employing RNA-seq it is possible to characterize RNA splicing, editing and post transcriptional modifications, in which intronic sequences are removed and subsequently exons are combined. Through using RNA-seq it is possible to know the actual sequences present in spliced RNAs and the ones that are not present rather than depending on algorithmic prediction tools. The difference in RNA splicing can cause major changes in the translated proteins, which can cause functional consequences for cells and subsequently for the organism.[23,52,59]

**7. Treasure Behind mRNA Data analysis**

It's now clear to understand the various roles of mRNA in managing and regulating expression of gene, but the analysis of the integrated mRNA is a chasing computational challenge, which requires the input of various bioinformatics system and expertise to research in it. With the outcome of sequencing and multiple array based technologies, a broad variety of RNA expression profiling products have been launched. With the boom in modeling and bioinformatics field, several essential tools to read the expression data in various set of conditions such as healthy versus cancerous tissues and cell lines targeted against drugs results in understanding several types of analysis that can be determined such as identification of differentially expressed genes in various conditions,

computing the differentially expressed mRNA enrichment, potential mRNA target interaction study, analyzing the biological pathway in differentially expressed genes and understanding their direct and indirect interaction networks.

Analyzing a complete set of transcripts present in the cell for any particular physiological condition is obtained by transcriptome. It is highly essential to understand and interpret the various functions of genome and revealing the constituents of cells and tissue as well understanding both the developmental and disease stage. The major aim of transcriptomics is to categories different species of transcripts which encodes mRNA, small RNAs and non-coding RNA to understand the transcriptional structure from 5'and 3' end.[60] Patterns of splicing, post transcriptional modifications as well to quantify expression levels in various developmental stage with respect to various conditions. Though RNA-Seq is highly used in research but this technology is still under active development. Firstly RNA-Seq is not limited to determining transcripts for the known genome but also an attractive tool for non-model organism with genomic data that are yet to be analyzed.[2,54] Secondly, they are highly precise to reveal the location of transcription to single base resolution as well as RNA-Seq is also useful understand complex transcriptome and denote sequence variations for the transcribed regions.[54] Thirdly RNA-Seq has very less background signals compared to DNA sequences.[50] Fourthly, it has a wide range of expression levels by which transcripts can be detected compared to DNA microarrays as they lack sensitivity for expression of genes. Fifthly, RNA-Seq is also well known for the accurate for quantifying the expression levels of the genes.[50] The overall results for both technical and biological replicates from RNA-Seq also have shown high levels of reproducibility.[2,50]Table 1 denotes the comparison of transcriptomics methods and its advantages with various current technologies. Finally the amount of RNA sample required is very less as there is no amplification steps with the helicos technology. Therefore RNA-Seq is known as the first sequencing methods that reads the whole transcriptome with a high throughput manner (Figure 4. describes RNA pipeline).

**8. Trouble shooting stages with RNA-seq**

Though there are few steps in RNA-Seq, it does have several troubleshooting stages during the library preparation, which can affect its use in profiling different types of transcripts. During library preparation, few manipulations will lose affect the analysis of RNA-Seq results. Similarly RNA-Seq also faces various bioinformatics challenges, which includes retrieving data, efficient storage method and processing large some of data which should overcome base-calling, reduce errors and remove low-quality reads. Alignment is also complicated for large transcriptomes due to various portion of the sequenced genome match with multiple locations. Even high copy number of short reads with repetive regions of long stretch also indicates a great challenge to face. Generally mapping large differences also requires comparable reference genome annotation to both studying polymorphism as well to attain deep sequencing coverage. Even the percentage of transcripts surveyed in the sequence coverage is an important issue with respect to cost. Usually greater coverage therefore developing simple easy computational methods to identify novel splicing events between two sequences as well as between exons from completely different genes is a great challenge for the future. Despite the various challenges mentioned above, RNA-Seq has generated unprecedented global view on transcriptome to organize and analysis various number of species as well as cell types. With high resolution and sensitivity many novel transcribed sites and splicing is forms of familiar genes mapped from 5'-3'boundaries. Along with it, compared to microarrays the RNA expression levels are more accurate. The most powerful advantage is that they can capture Transcriptome dynamics across various tissues with multiple conditions without affecting the normalization of the data set.However RNA-Seq has been undoubtedly much valuable for understanding the dynamics behind transcriptomics in development stage, normal physiological changes, and in the analysis of biomedical samples, where it can result in robust comparison between diseased and normal tissues, as well as the sub classification of disease states. Soon due to the fall in price, RNA-Seq will have many inbuilt applications that involves determine structure and dynamics of the transcriptomes replacing microarrays technology.

**Future directions**

Although RNA-Seq is still in the early stages of use, it has clear advantages over previously developed transcriptomic methods. The next big challenge for RNA-Seq is to target more complex transcriptomes to identify and track the expression changes of rare RNA isoforms from all genes. Technologies that will advance achievement of this goal are pair-end sequencing, strand-specific sequencing and the use of longer reads to increase coverage and depth. As the cost of sequencing continues to fall, RNA-Seq is expected to replace microarrays for many applications that involve determining the structure and dynamics of the transcriptome

**Financial Disclosure**

Financial support of this work was provided by the STACK-Large grant 162-34 to Ishtiaq Qadri". Supported by a research grant from IQ foundation.

**Conflict of Interest**

All authors don't have any conflict of interest.

**Table 1: Comparison of transcriptomics methods with various technologies**

| Specifications | Microarray | cDNAsequencing | RNA-Seq |
|---|---|---|---|
| Principle | Hybridization | Sanger method | High throughput sequencing |
| Sample required | high | high | low |
| Resolution | Upto 100bp | Single base | Single base |
| Background signal | High | Low | Low |
| Throughput | High | Low | High |
| Reproducibility | low | Low | high |
| Sensitivity (dynamic range of expression) | Low | Low | high |
| Accuracy | Low | Low | high |
| Quantify Expression level | Upto-100 fold | Not possible | Greater than 8000 fold |
| Cost of mapping | High | High | low |

**Figure 1: RNA World: Transcriptome**

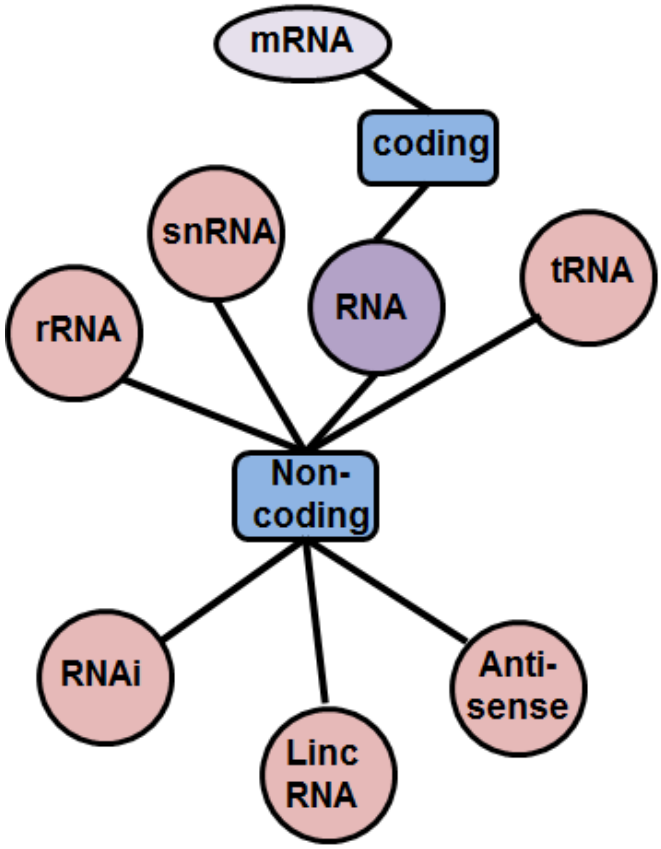



**Figure 2: Overview of RNA-seq experiment**

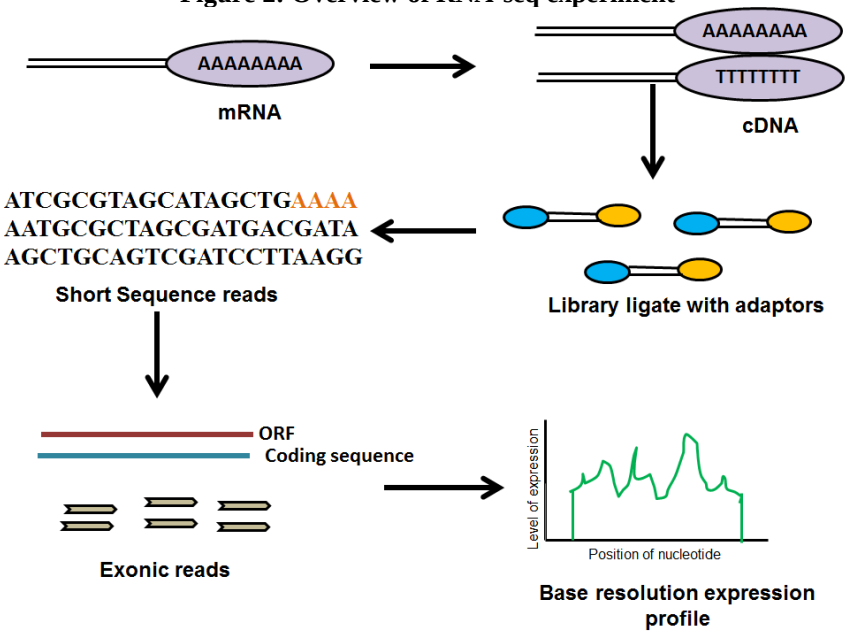


**Figure 3: Overview of RNA-Seq analysis and choosing a method**

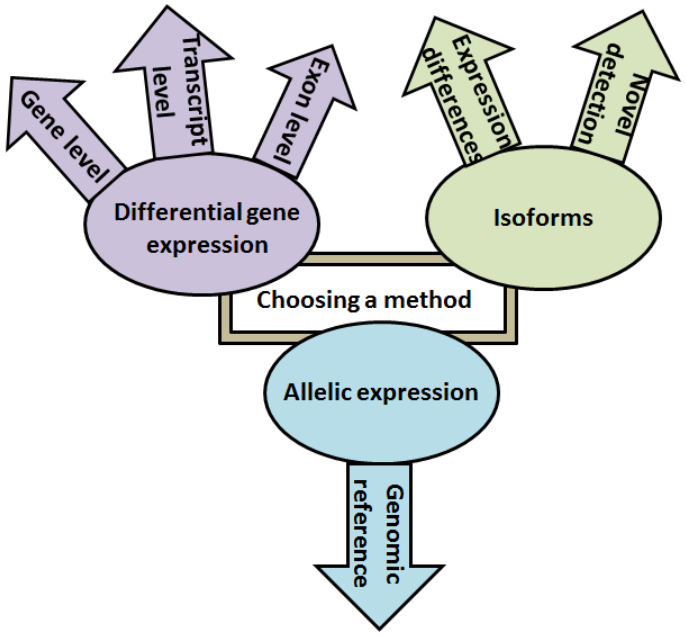
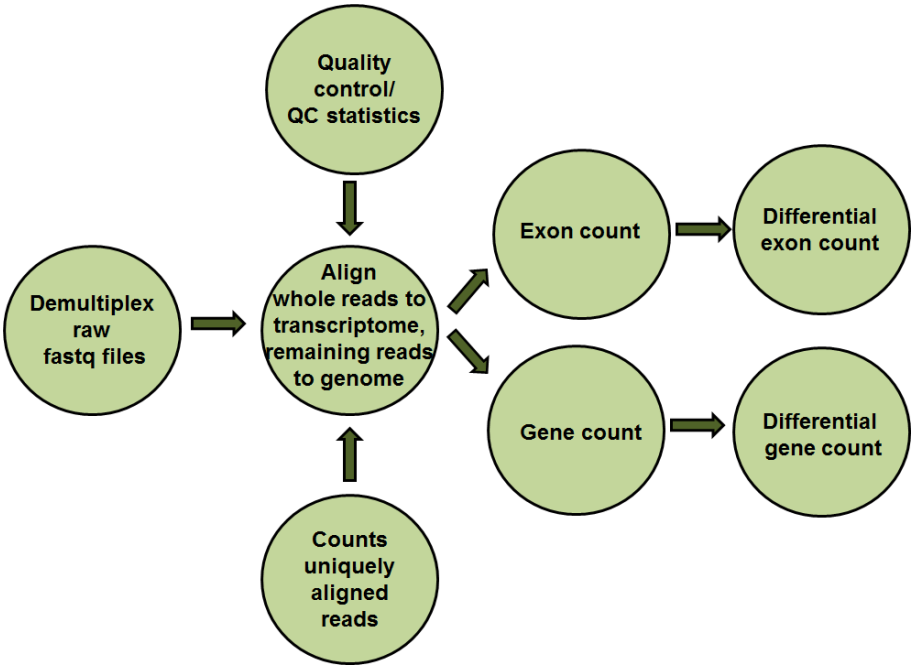



**Figure 4: RNA-seq Pipeline**



## References


1. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research.* 2008;18(9):1509-1517.
2. Cloonan N, Forrest AR, Kolle G, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature methods.* 2008;5(7):613-619.
3. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS. SNP discovery via 454 transcriptome sequencing. *Plant J.* Sep 2007;51(5):910-918.
4. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics.* 2009;10(1):57-63.
5. Nagalakshmi U, Waern K, Snyder M. RNA-Seq: a method for comprehensive transcriptome analysis. *Curr Protoc Mol Biol.* Jan 2010;Chapter 4:Unit 4 11 11-13.
6. Tsuchihara K, Suzuki Y, Wakaguri H, et al. Massive transcriptional start site analysis of human genes in hypoxia cells. *Nucleic Acids Res.* Apr 2009;37(7):2249-2263.
7. Rio DC AM, Hannon GJ, Nilsen TW. Enrichment of poly(A)+ mRNA using immobilized oligo(dT). *Cold Spring Harb Protoc.* 2010.
8. Ares M, Jr. Methods for processing high-throughput RNA sequencing data. *Cold Spring Harb Protoc.* Nov 2014;2014(11):1139-1148.
9. Malone C BJ, Czech B, Aravin A, Hannon GJ. Preparation of small RNA libraries for high-throughput sequencing. *Cold Spring Harb Protoc.* 2012.
10. Wilhelm BT LJ. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods.* 2009;49(249-257).
11. Yoon OK BR. Noncanonical transcript forms in yeast and their regulation during environmental stress. *RNA.* 2012;16:1256–1267.
12. Affymetrix ETP, Cold Spring Harbor Laboratory ETP. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature.* Feb 19 2009;457(7232):1028-1032.
13. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods.* Nov 2009;6(11 Suppl):S22-32.
14. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* May 1 2009;25(9):1105-1111.
15. Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.* Aug 2010;38(14):4570-4578.
16. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 2004;5(10):R80.
17. Reimers M, Carey VJ. Bioconductor: an open source framework for bioinformatics and computational biology. *Methods Enzymol.* 2006;411:119-134.
18. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* Sep 2008;18(9):1509-1517.
19. Carninci P, Kasukawa T, Katayama S, et al. The transcriptional landscape of the mammalian genome. *Science.* 2005;309(5740):1559-1563.
20. Adams J. Transcriptome: connecting the genome to gene function. *Nature Education.* 2008;1(1):195.
21. Jacquier A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nature Reviews Genetics.* 2009;10(12):833-844.
22. Mutz K-O, Heilkenbrinker A, Lönne M, Walter J-G, Stahl F. Transcriptome analysis using next-generation sequencing. *Current opinion in biotechnology.* 2013;24(1):22-30.
23. Consortium SM-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotech.* 2014;32(9):903-914.
24. Ginsburg GS, Willard HF. Genomic and personalized medicine: foundations and applications. *Translational Research.* 2009;154(6):277-287.
25. Velculescu VE, Zhang L, Zhou W, et al. Characterization of the yeast transcriptome. *Cell.* 1997;88(2):243-251.
26. Lindberg J, Lundeberg J. The plasticity of the mammalian transcriptome. *Genomics.* 2010;95(1):1-6.
27. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. *Nature.* 1980;284(5757):601-603.

Taft RJ, Pheasant M, Mattick JS. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays.* 2007;29(3):288-299. **28.**

Cavalier-Smith T. Cell volume and the evolution of eukaryotic genome size. *The evolution of genome size.* 1985:105-184. **29.**

Birney E, Stamatoyannopoulos JA, Dutta A, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature.* 2007;447(7146):799-816. **30.**

Costa V, Angelini C, Ciccodicola A. Uncovering the complexity of transcriptomes with RNA-Seq. *BioMed Research International.* 2010;2010. **31.**

Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY. Understanding the transcriptome through RNA structure. *Nature Reviews Genetics.* 2011;12(9):641-655. **32.**

Clark TA, Sugnet CW, Ares M. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science.* 2002;296(5569):907-910. **33.**

David L, Huber W, Granovskaia M, et al. A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences.* 2006;103(14):5320-5325. **34.**

Yamada K, Lim J, Dale JM, et al. Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science.* 2003;302(5646):842-846. **35.**

Bertone P, Stolc V, Royce TE, et al. Global identification of human transcribed sequences with genome tiling arrays. *Science.* 2004;306(5705):2242-2246. **36.**

Cheng J, Kapranov P, Drenkow J, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science.* 2005;308(5725):1149-1154. **37.**

Okoniewski MJ, Miller CJ. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC bioinformatics.* 2006;7(1):276. **38.**

Royce TE, Rozowsky JS, Gerstein MB. Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. *Nucleic acids research.* 2007;35(15):e99. **39.**

Boguski MS, Tolstoshev CM, Bassett Jr DE. Gene discovery in dbEST. *Science.* 1994;265(5181):1993-1994. **40.**

Gerhard DS, Wagner L, Feingold EA, et al. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome research.* 2004;14(10B):2121-2127. **41.**

Kodzius R, Kojima M, Nishiyori H, et al. CAGE: cap analysis of gene expression. *Nature methods.* 2006;3(3):211-222. **42.**

Shiraki T, Kondo S, Katayama S, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences.* 2003;100(26):15776-15781. **43.**

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science.* 1995;270(5235):484-487. **44.**

Harbers M, Carninci P. Tag-based approaches for transcriptome research and genome annotation. *Nature methods.* 2005;2(7):495-502. **45.**

Brenner S, Johnson M, Bridgham J, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology.* 2000;18(6):630-634. **46.**

Peiffer JA, Kaushik S, Sakai H, et al. A spatial dissection of the Arabidopsis floral transcriptome by MPSS. *BMC plant biology.* 2008;8(1):43. **47.**

Reinartz J, Bruyns E, Lin J-Z, et al. Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Briefings in functional genomics & proteomics.* 2002;1(1):95-104. **48.**

Fu X, Fu N, Guo S, et al. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC genomics.* 2009;10(1):161. **49.**

Nagalakshmi U, Wang Z, Waern K, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science.* 2008;320(5881):1344-1349. **50.**

Wilhelm BT, Marguerat S, Watt S, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature.* 2008;453(7199):1239-1243. **51.**

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods.* 2008;5(7):621-628. **52.**

Lister R, O'Malley RC, Tonti-Filippini J, et al. Highly Integrated Single-Base Resolution Maps of the Epigenome in< i> Arabidopsis</i>. *Cell.* 2008;133(3):523-536. **53.**

Morin RD, Bainbridge M, Fejes A, et al. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques.* 2008;45(1):81. **54.**

**55.** Otto TD, Wilinski D, Assefa S, et al. New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. *Molecular microbiology.* 2010;76(1):12-24.

**56.** Sorber K, Dimon MT, DeRisi JL. RNA-Seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. *Nucleic acids research.* 2011;39(9):3820-3835.

**57.** Van Keuren-Jensen K, Keats JJ, Craig DW. Bringing RNA-seq closer to the clinic. *Nat Biotech.* 2014;32(9):884-885.

**58.** Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods.* 2010;7(12):1009-1015.

**59.** Külahoglu C, Bräutigam A. Quantitative Transcriptome Analysis Using RNA-seq. *Plant Circadian Networks*: Springer; 2014:71-91.

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