Recent Trends in 'Computational Transcriptomics'

Chapter · July 2020

CITATIONS
0

READS
281

4 authors:

Pramod Katara
University of Allahabad
52 PUBLICATIONS **301** CITATIONS
SEE PROFILE

Neelam Krishna
University of Allahabad
7 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Anamika Yadav
University of Allahabad
10 PUBLICATIONS **48** CITATIONS
SEE PROFILE

Shraddha Vishwakarma
University of Allahabad
6 PUBLICATIONS **2** CITATIONS
SEE PROFILE

In: Recent Trends in 'Computational Omics' ISBN: 978-1-53617-941-5
Editor: Pramod Katara

Chapter 4

RECENT TRENDS IN 'COMPUTATIONAL TRANSCRIPTOMICS'

Pramod Katara*, Neelam Krishna, Anamika Yadav and Shraddha Vishwakarma
Computational Omics Lab, Centre of Bioinformatics,
University of Allahabad- Prayagraj, India

ABSTRACT

Transcripts are the product of transcription process, and the total transcript content of the cell is known as transcriptome. Unlike the genome, which is static, transcriptome is dynamic, and it's varying cell to cell, even within a cell in different conditions. In general, transcriptome reflects the gene expression of the cell in given conditions, thus, utilize as a tool for gene expression and functional genomics studies. Transcriptomics is now a mature field and has various techniques to study transcriptome. The most reliable and high throughput techniques for transcriptomics are cDNA microarray, and NGS based RNA-Seq. High throughput nature of these techniques provides competence to analyze genome wide gene expression, which further can utilize to perform functional genomics studies. Both

* Corresponding Author's Email-pmkatara@gmail.com.

cDNA and RNA-seq generate a huge amount of data which require bioinformatics resources for storage and analysis purpose. The current chapter is focused on concept, techniques, databases and data analysis pipelines, along with the scope of transcriptomics.

Keyword: cDNA, microarray, RNA-Seq, transcript, transcriptome, transcriptomics

1. INTRODUCTION

Cells have a different biological process, and one of the most important of them is transcription, which is responsible for the expression of the gene (coding as well as a non-coding gene). The products of this process are known as a transcript (i.e., mRNA, rRNA, tRNA and miRNA). Like the term genome, the total content of transcript (all RNA) molecules in a cell or population of the cell at a particular given time is known as transcriptome. Thus, by definition, all transcripts (RNAs) are the part of the transcriptome, but sometimes this term depends on the particular experiment where they only consider mRNA as the content of transcriptome. Transcriptome is a mirror of the sequence of the DNA (gene) from which it has been transcribed, thus by analyzing the entire collection of RNA sequences in a cell (transcriptome), researchers can determine when and where each individual gene is turned on or off in a particular cell or tissue of an organism. Overall, we can say that unlike the genome, which is static in nature, transcriptome of any cell or tissue is dynamic in nature which shows variability that depends on the requirement of the cell. By collecting and comparing transcriptomes of different types of cells, researchers can gain a deeper understanding of what constitutes a specific cell type, how that type of cell normally functions and how changes in the normal level of gene activity may reflect or contribute to disease. In addition, transcriptome may enable researchers to generate a comprehensive, genome-wide expression profiling of the genes that provide a picture of ‘what genes are active in which cells’. Such studies where we analyze the complete transcriptome are called ‘transcriptomics’.

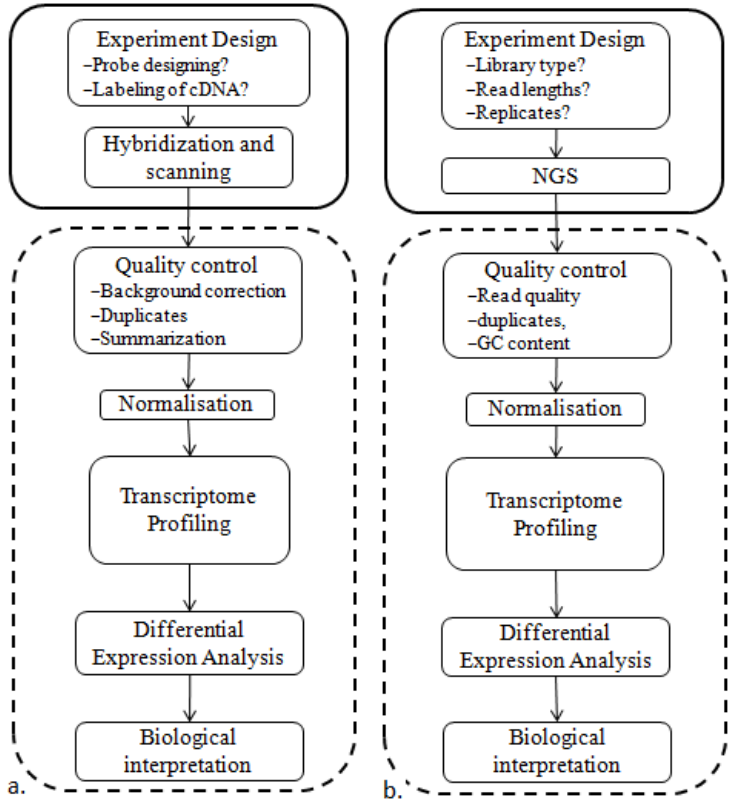


Figure 1. Schematic representation of basic steps of a) cDNA microarray and b) RNA-seq technology. Dashed line portion indicates computational transcriptomics; it deals with data storage, processing and biological interpretation.

As discussed, transcriptome is dynamic in nature, and its content is time and cell specific (e.g., normal, diseased), thus transcriptomics needs high throughput techniques to analyze and create a clear picture of genome wide gene expression. By realizing the importance of transcriptomics, especially in the field of functional genomics and medical sciences (behavior of cell or tissue in diseased conditions), various transcriptomics technologies were developed in past, including EST, SAGE, cDNA microarray along with recent NGS-based RNA-seq technology [1, 2]. All of these techniques have some pros and cons, but cDNA microarray and RNA-seq shows potential at high throughput level with some extra advantages over the rest of the

techniques, thus currently utilized preferably for transcriptomics studies [3-5].

Both cDNA and RNA-seq, due to their high throughput nature, produce voluminous raw data that require high-end computing facilities, sophisticated bioinformatics resources and approaches for the purpose of storage, processing and to analyze data more accurately and efficiently to get an inherent meaning full conclusion from that. In this chapter, we mainly focus on the data analysis process and bioinformatics resources for cDNA and RNA-seq data analysis.

2. cDNA MICROARRAY TECHNOLOGY

cDNA microarray is one of the potential techniques of microarray-based technology, as its name suggests this technique works on cDNA which is prepared from mRNA of the cell of interest, through the reverse transcription process. This reverse transcription takes place without any quantitative alteration thus the quantity of cDNA remains directly proportional to extract mRNA, and therefore, provides qualitative as well as quantitative observation of gene expression. Once cDNA formation took place they get labelled with different dyes, preferably with Cy3 (Green)/Cy5 (Red). Normal cDNA microarray, which analyzes one sample is labelled with one dye thus also known as one color cDNA microarray and those experiments which analyze two samples in the same experiment uses two dyes to differentiate them, one for each sample, are known as two-color cDNA microarray. After labelling cDNA sample is applied to microarray chip which is prefixed with oligonucleotide probes and provide sites for hybridization with complementary cDNA. These probes are gene-specific thus only provide specific hybridization with targeted cDNA. The hybridization process followed by washing and at last, scanning. For scanning, purpose-specific scanners are there which scan dye specific light intensity for each spot, which is directly proportional to the quantity of cDNA = mRNA = Transcription (Gene expression). These scanners convert

this light intensity in mathematical values which finally get stored in computers [6, 7].

2.1. Representation of Gene Expression Data

To carry out any significant biological analysis, microarray data are represented in a specific manner, some important representations are as follows (Table 1):

A. Absolute measurement: In absolute measurement representation, each cell in the matrix will represent the expression level of the gene in abstract units.
B. Relative expression ratio: In relative representations, the expression level of genes in abstract units is normalized with respect to its expression in a reference condition. This gives the expression ratio of the gene in relative units (30/20), thus there is a chance that absolute measurement will be lost in such representation because the ratio of 300/200 = 30/20 will lead to the same result. Relative expression representations have its own advantage; by using this comparison across different conditions can be made as long as the same reference condition is used to get the expression ratio.
C. Log_2 expression ratio: Log_2 representation is very useful and attractive, because it provides information on up regulation and down regulation in a very symmetric manner, for example, 4 fold up-regulation maps to Log_2 (4) = 2 and a 4 fold down-regulation maps to Log_2 (1/4) = -2. The Scientist preferably used Log_2 representation for fold change analysis to identify differentially expressed/regulated genes under given conditions.
D. Discrete Values: On the basis of the requirement, gene expression data can be converted into discrete value. Binary expression matrixes (1, 0) are preferentially utilized to convert the data into discrete values, where 1 means that the gene is over-expressed than user-defined threshold and 0 means that the gene is expressed below

the threshold. Discrete values can also be used to develop a relative expression matrix or Log_2 matrix. In such cases values are classified into one of three classes: +1, 0 and -1, where +1 represents a positively regulated gene (over-expressed), 0 represent constantly expressed genes and -1 represent under-expressed (repressed) genes. Though conversion of gene expression values in discrete values is useful in a certain analysis where real values cannot work, use of discrete value losses quantitative information about gene expression.

Table 1. (A-D) Representations of gene expression data matrix in four different ways that contain rows representing genes and columns representing expression values in a different format in particular experimental conditions

Table A. Absolute measurement

	R1	R2	R3	R4
Gene1	30	240	60	60
Gene2	100	209	400	200
Gene3	10	80	40	20
Gene4	20	161	80	80

Table B. Relative measurement

	R1/Rr	R2/Rr	R3/Rr
Gene1	0.50	4.00	1.00
Gene2	0.50	1.00	2.00
Gene3	0.50	4.00	2.00
Gene4	0.25	2.00	1.00

Table C. Log_2 (relative measurement)

	Log_2 (R1/Rr)	Log_2 (R2/Rr)	Log_2 (R3/Rr)
Gene1	-2	2.0	0.0
Gene2	-1	0.0	1.0
Gene3	-1	2.0	2.0
Gene4	-2	1.0	0.0

Table D. Discrete values

	D [Log_2 (R1/Rr)]	D [Log_2 (R2/Rr)]	D [Log_2 (R3/Rr)]
Gene1	-1	1	0
Gene2	0	0	0
Gene3	0	1	-1
Gene4	-1	0	0

2.2. Microarray Database

After the extraction of quantitative information from the images resulting from the readout of fluorescent or radioactive hybridization (image analysis) the scanned output, which work as a gene expression information, stored in the database, from where user can use it for various analyses. Presently, lots of databases for microarray are available which store data in various forms and for a range of organisms (Table 2). The key features of a microarray database are; they store the measurement data, manage a searchable index, and make the data available to other applications for analysis and interpretation (either directly or via user downloads).

Microarray Gene Expression Data (MGED) Society: The Microarray Gene Expression Data (MGED) Society is an international organization established in 1999 for facilitating sharing of microarray data. To facilitate data sharing, society established relevant data standards. The three main components of MGED standards are – *i*) Minimum Information about a Microarray Experiment (MIAME), *ii*) Microarray Gene Expression (MAGE) and *iii*) MGED Ontology (MO). Overall MGED society established the data standards for sharing and defines sets of common terms and annotation rules for microarray experiments which has been enabled proper annotation, data analysis and data exchange, without loss of meaning of the data [8].

As mentioned, there are various databases which show the collection of cDNA microarray data from different source and platform, few of these databases are organism specific, e.g., ExpressDB, few are condition-specific, e.g., Oncomine, and rest are universal, e.g., GEO (Table 3). All databases have their indispensable utilities, few of these databases are also linked with data analysis and visualization facilities. GEO, which is maintained by NCBI, is a widely used transcriptome data repository, it provides an extensive collection of all types of transcriptomics data, i.e., SAGE, cDNA microarray, RNA-seq [9]. For cDNA microarray, GEO mainly provides four different file accessions, all of them contain specific information (Table 3).

Table 2. Frequently used cDNA microarray database

S. No	Database	Description
1	ArrayExpress	It is a public database for high throughput functional genomics data hosted at European Bioinformatics Institute (EBI), https://www.ebi.ac.uk/arrayexpress/.
2	GEO	Public gene expression data from various platforms at the National Center for Biotechnology Information (NCBI), www.ncbi.nlm.nih.gov/geo.
3	YMD	The Yale Microarray Database (YMD) is a university-wide database for archiving and retrieving microarray data generated by different labs using different platforms, https://medicine.yale.edu/keck/ymd/.
4	ExpressDB	Yeast RNA expression data, http://arep.med.harvard.edu/ExpressDB/.
5	Oncomine	It is a cancer microarray database with data mining facilities to facilitate genome-wide gene expression based discoveries, www.oncomine.org.

Table 3. Details of different cDNA data files provided by GEO

S. No	Accession	Description
1	GPL	A compendious description of the array or sequencer platform. A platform is a reference of various samples that have been submitted by various submitters (GPLxxx).
2	GSE	It describes the conditions under which an individual's sample (GSExxx).
3	GSM	It records the links together a group of related samples from a single platform and may be included in multiple series (GSMxxx).
4	GDS	It is an original submitter supplied record that summarizes an experiment (GDSxxx).

2.3. Microarray Data Analysis

cDNA Microarray technology allows us to assess gene expression patterns of the hefty number of genes under multiple conditions, these conditions may be a time series during a specific biological process (i.e., cell cycle, after specific treatment) or a collection of different tissue samples (normal versus treated tissue). To conclude anything from microarray data, data need to process through various data analysis phase, including - preprocessing (quality control), normalization, transformation, clustering and data analysis (DE, Annotation, etc.).

2.3.1. Data Preprocessing and Normalization (Quality Control)

The raw microarray data, which are the intensity read for each component, are generally infected with various sources of variation (Table 4). When starting a new microarray analysis, raw data need to be preprocessed to remove artifact and undesirable effects [10]. Preprocessing mainly includes the following five major tasks:

1) Background Correction: The aim of this step is to correct for what is usually known as the *background effect*. That is any source of technical variation reflected in a spatial pattern of the intensity measurements.
2) Within Array Correction: This is more important in the case of two-color array, it removes the effect of differences in sample quantity due to the differences in the processing of the two samples.
3) Between Array Scaling (Normalization): It attempts to normalize the non-biological variations between different experiments so that they can analyze for the same purpose.

Table 4. Source of data noise and variations

S. No.	Type of variation	Source
1	Biological Variations	• RNA is extracted from individuals or cell cultures, so it should be at least from the same strain. • Different cells might be in a different developmental stage.
2	Experimental Variations	• The laboratory equipment may vary. • The expertise may vary. • Variation during hybridization and washing. • Variation during scanning.
3	Background Noise	• Non-specific binding of probes. • Substrate reflection. • Slide reflection. • Buffer effect.

4) Summarization: In this, array intensities are summarized in a final measurement relating each biological feature of interest in the study. It attempts to normalize the presence of duplicates and control spots

which originally designed for quality checking, background signal estimation or to measure cross-hybridization.

5) Quality assessment: At the end, we need to check if the modified data, the normalized data in microarray terminology, are free of the original artifacts that.

2.3.2. Differential Gene Expression

After preprocessing and normalization, data shows absolute (baseline) expression of individual genes. To get the differential gene expression (expression of genes in different conditions) we rely on gene expression ratio (treated/reference), also known as fold change, which gives an idea about the relative expression of genes in two different conditions. The expression ratio is a relevant way of representing expression differences in a very intuitive manner.

Gene expression ratio = Treated/Reference = T/R

Here:

T is the gene expression level in the testing sample.

R is the gene expression level in the reference sample.

Transformation: As mentioned below, the gene expression ratio that is also known as fold change doesn't provide a clear picture (Box 1). Gene expression ratio needs a transformation to give a clear picture of gene expression variations. Generally, to provide better relative measurement log base 2 transformation is in practice (i.e., $\log_2$ expression ratio)). This has a major advantage that it treats differential up-regulation and down-regulation equally and also has a continuous mapping space.

Though, Log_2 transformation provides comparable expression patterns; it is associated with serious risk; it removes all information about absolute expression levels of the genes. It may miss differentially expressed genes with large differences (T-R) but small ratios (T/R), leading to a high miss rate at high intensities.

Box 1. Impact of Log_2 transformation; here in case-2 and case-3, it is clear that up-regulation is blown up and mapped between 1 and infinity, whereas down-regulation is compressed and mapped between 0 and 1. Log_2 transformation of these values eliminates these inconsistencies and provides a comparable range

Case	Gene expression ratio	Log2 transformation
1).	T/R = 4T/4R = 1	= Log2 (1)= 0
2).	T/R = 4T/1R = 4	= Log2 (4)= 2
3).	T/R = 1T/4R = ¼ (0.25)	= Log2 (1/4)= -2

3. RNA-SEQ

RNA-seq, also known as whole transcriptome shotgun sequencing, is another high throughput technique, which is utilized for transcriptomics related studies. RNA-seq is based on next-generation sequencing concepts which have the potential to provide qualitative as well as quantitative analysis of RNA in a given biological sample [11]. RNA-seq was arising in the last decade as a powerful method for transcriptome analyses that will eventually make microarrays obsolete for gene expression analysis.

3.1. RNA-Seq Database

RNA-seq generates a huge amount of short-read sequences. To avoid any error, normally the scientists perform deep sequencing, which exponentially added the data size that creates the needs of a robust data management system to handle this data. At the same time to match up the required space for data processing, few of RNA-seq databases are also linked with cloud computing facilities. As per the requirement, the scientists developed various RNA-seq data resources from where user can access the data (Table 5).

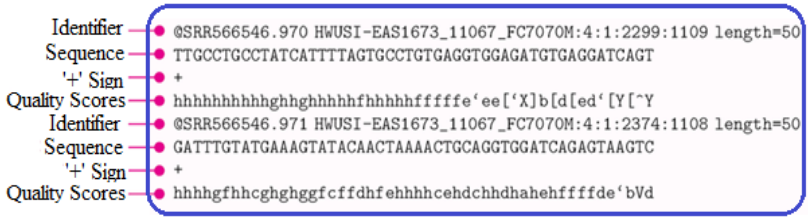
Table 5. Commonly used RNA-Seq database

S. No.	Database	Description
1	SRA	The Sequence Read Archive (SRA) stores raw sequence data from "next-generation" sequencing technologies [12].
2	ENA	The European Nucleotide Archive (ENA) provides a comprehensive record of the nucleotide sequencing information, i.e., raw sequencing data, sequence assembly information and functional annotation [13].
3	ANTE	Oncobox Atlas of Normal Tissue Expression (ANTE) provides the collection of gene expression database of normal human tissues; it has been designed by analyzing 67 original and 396 published experimental datasets [14].
4	CIRCpedia v2	It's a compressive database of circular-RNA, which allows users to search, browse and download annotated circRNAs, with cell/tissue-specific expression characteristics. It also provides conservation analysis of circRNAs between humans and mice [15].
Human related RNA-Seq database		
5	Brain RNA-Seq	RNA-Seq transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex, http://brainrnaseq.org/.
6	FusionCancer	A database of cancer fusion genes derived from RNA-seq data [16].
7	Hipposeq	Hipposeq a comprehensive RNA-Seq database of gene expression in hippocampal principal neurons [17].
8	Mitranscriptome	Mitranscriptome provides a systematic list of long poly-adenylated Human RNA transcripts based on RNA-seq data from more than 6,500 samples associated with a variety of cancer and tissue types [18].
9	RNA-Seq Atlas	A reference database for gene expression profiling in normal tissue by next-generation sequencing [19].
10	DASHR	A database of human small RNA genes and mature products derived from small RNA-seq data [20].

3.1.1. Sequence Retrieval Archives (SRA)

GEO is the main transcriptome database which provides all types of transcriptome data. One can choose an RNA-seq experiment of their interest from the GEO and then redirect to SRA, which provides a collection of raw sequencing data (genome, exome, transcriptome), and alignment information from high throughput sequencing platform. SRA provides a range of information related to NGS experiments in the form of different accession (https://www.ncbi.nlm.nih.gov/sra).

Box 2. Sample of Fastq format which stores sequence from NGS along with a quality score



- **SRA (submission accession):** It holds the objects represented by the other five accessions and is used to track the submission in the archive.
- **SRP (study accession):** It contains the project metadata describing a sequencing study or project. Since the SRP accession ultimately references (links to) all other 5 data types in a study, it can be used as a starting point to access any of the data in that study.
- **SRX (experiment accession):** It contains the metadata describing the library, platform selection, and processing parameters involved in a particular sequencing experiment.
- **SRR (run accession):** It contains actual sequencing data for a particular sequencing experiment.
- **SRS (sample accession):** It contains the metadata describing the physical sample upon which a sequencing experiment was performed.
- **SRZ (analysis accession):** It contains a sequence data analysis BAM file and the metadata describing the sequence analysis.

All of these data resources provide data in some standard formats, most important of them is Fastq (Box 2). Fastq is different from FASTA format in terms of the presence of additional information, i.e., base quality score, which indicates the quality of each base of the sequence in the form of ASCII values (corresponding to PHRED quality score). On the basis of this quality

score, scientists decide whether to use or mask the corresponding base for further analysis.

3.2. RNA-Seq Pipeline

Like cDNA microarray, RNA-seq also comprises the two parts-experimental and computational biology. Experimental biology includes four main steps *i*) RNA extraction, *ii*) RNA fragmentation and reverse transcription, *iii*) Library construction and *iv*) Sequencing. RNA-seq sequencing utilizes the NGS approach, thus resulting in a huge amount of reads which requires computational facilities to store them and for further analysis (Figure 1). Computational biology also in broad comprised of four steps, including *i*). quality control, *ii*) transcriptome profiling *iii*) differential expression and *iv*) functional profiling. Due to the enormous size of the RNA-seq data, computational biology steps need an intensive computational facility to analyze data without compromising the biological information.

3.3. RNA-Seq Data Analysis

As discussed, RNA-seq is based on the NGS technologies, thus provides short deep reads from the sequencing of total RNA of the cell (transcriptome). Raw RNA-seq data is infected with various noises and unwanted data. Thus before performing any biological analysis, RNA-seq data must pass through various statistical steps, where each step attempts to improve the quality of the reads. For each step, various software's are available (Table 6, 8), but till now there is no standard data analysis pipeline, is available for the RNA-seq data, various pipelines are reported which show variation mainly based on the objective of data analysis. Though, as we discussed, there is no standard pipeline for RNA-seq data analysis, and all of the pipelines must follow the following data analysis steps to conclude the experiment.

Table 6. Purpose specific software for RNA-Seq data analysis

S. No.	Tool/ Software	Description
A)	Quality check and pre-processing tools	
1	FASTQC	A quality control tool for high throughput sequence data and import data in BAM/SAM/FASTQ/FASTQC file format [21].
2	FASTX Toolkit	It is a collection of command-line tools for Short-Reads FASTA/FASTQ files preprocessing [22].
3	RNA-SeqC	RNA-SeqC is a java program which computes a series of quality control metrics for RNA-seq data [23].
4	AfterQC	Automatic Filtering, Trimming, Error Removing and Quality Control for fastq data [24].
5	NGS QC Toolkit	A toolkit for the quality control (QC) of NGS data [25].
B)	Trimmer/Adapter/Error Removal	
1	Trimmomatic	A flexible read trimming tool for Illumina NGS data [26].
2	Cutadapt	Cutadapt removes adapter sequences from high-throughput sequencing data, https://cutadapt.readthedocs.io/en/stable/.
3	AdapterRemoval	It provides facilities to remove residual adapter sequences from NGS reads (from both, single and paired-end data), [27].
4	SEECER	SEECER removes mismatch and indel errors from the raw reads and significantly improves downstream analysis of the data, http://sb.cs.cmu.edu/seecer/.
5	Flexbar	Flexible barcode and adapter removal for NGS platforms, https://github.com/seqan/flexbar.
C)	Mapping/Alignment/Assembly	
1	Trinity	Trinity assembles transcript sequences from Illumina RNA-seq data [28].
2	Bowtie 2	Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences [29].
3	HISAT2	It is a fast and sensitive alignment program for mapping NGS reads (both DNA and RNA) to human genomes (as well as to a single reference genome), [30].
4	STAR	Spliced Transcripts Alignment to a Reference (STAR) is a fast NGS read aligner for RNA-seq data [31].
5	TopHat	TopHat is a fast splice junction mapper for RNA-seq reads. It aligns RNA-seq reads to large-sized genomes using Bowtie, and then analyzes the mapping results to identify splice junctions between exons [32].
D)	Reads/Transcripts Quantification	
1	RSEM	Accurate transcript quantification from RNA-seq data with or without a reference genome [33].
2	Salmon	Salmon is a tool for quantifying the expression of transcripts using RNA-seq data [34].
3	Kallisto	Kallisto is a program for quantifying abundances of transcripts using high-throughput sequencing reads, i.e., RNA-seq data [35].

Table 6. (Continued)

S. No.	Tool/ Software	Description
4	Sailfish	Sailfish is a tool for transcript quantification from RNA-seq data. It follows a supervised approach for quantification, requires a set of target transcripts (either from a reference or de-novo assembly) to quantify [36].
5	Htseq-count	It is a tool for RNA-seq data analysis, given a SAM/BAM file and a GTF or GFF file with gene models; it counts for each gene how many aligned reads overlap its exons [37].
E)	Differential Expression Tools and Packages	
1	DESeq2	It is a method for differential analysis of count data, which is based on shrinkage estimation for dispersions and fold changes. It provides quantitative analysis that focuses on the strength rather than the mere presence of differential expression [38].
2	edgeR	It is designed for the analysis of replicated count-based expression data [39].
3	Cufflinks/ Cuffdiff	Cufflinks include a program, "Cuffdiff," which use to find significant changes in transcript expression, splicing, and promoter use [40].
4	NOISeq	NOISeq is a comprehensive resource that meets the current needs for robust data-aware analysis of RNA-Seq differential expression [41].
5	DESeq	DESeq is an R package to examine count data from high-throughput sequencing assays such as RNA-seq and test for differential expression (differential expression analysis for sequence count data), [42].

3.3.1. Data Preprocessing

Quality assessment and enhancement: quality assessment is the first step of the bioinformatics pipeline of RNA-seq, often, it is necessary to filter data, removing (trimming) low-quality sequences or bases adaptors, contaminations, or overrepresented sequences to ensure a coherent final result. Arrays of tools are available for this purpose with reads quality visualized graphically such as Fastqc.

Trimmomatic was developed to remove adapters and scan every read with a 4-base sliding window and trim the lower-scored bases along with low-quality N bases to enhance the quality of reads before alignment to the reference genome. It is also a good practice to assess the RNA-seq data quality after the preprocessing procedure.

3.3.2. Read Mapping

Once high-quality data are obtained from preprocessing, the next step is to map the short reads to the reference genome or to assemble them into contigs and align them to the reference genome. There are many popular bioinformatics programs that can be used for this purpose (Table 6, 8).

3.3.2.1. Challenges and Possible Solution

Presence of Poly (A) tails or exon-intron splicing junctions: Most of the available programs are typically suitable for reads that are not located at the poly (A) tails or exon-intron splicing junctions. Poly (A) tails can be easily identified by the presence of the multiple (As) or (Ts) and a partial junction library that contains the known junction sequence has been compiled to allow the alignment of difficult mapping reads.

Polymorphism: Another problem in reads mapping is that of polymorphisms, which occur when a sequence read aligns to multiple locations of the genome. Polymorphisms are especially common for the large and complex transcriptomes. For lower repetitive reads, one can employ the solution of assigning the reads to multiple locations proportionally based on the neighboring unique reads (Figure 2).

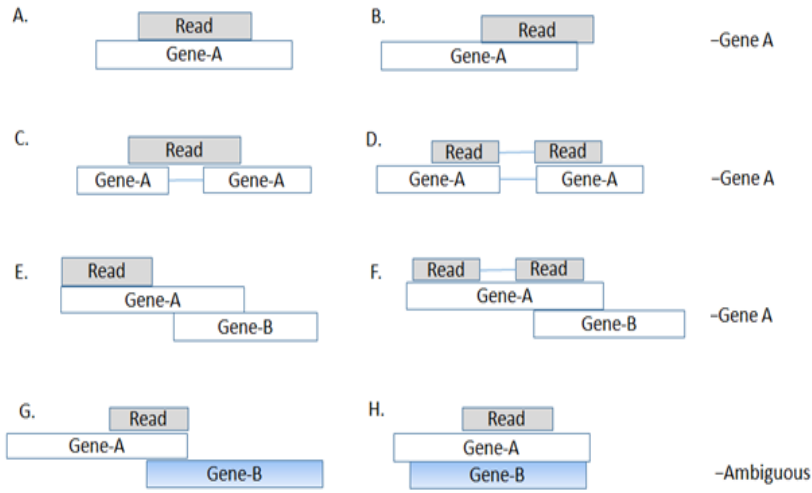


Figure 2. Assignment of reads which shows similarity with multiple genes.

However, for the short reads that have a very high copy number and repetitive sequences, polymorphism is still a great challenge. A longer read sequencer such as the Roche 454 or PacBio sequence analyzer might be required. Alternatively, there are bioinformatics solutions to extend the short pair-end reads into 200–500 bp fragments before deciding upon the multiple-aligned reads.

3.3.3. Quantification

The simplest approach to quantifying gene expression by RNA-seq is to count the number of reads that map (i.e., align) to each gene (read count) using programs such as HTSeq-count. This gene-level quantification approach utilizes a gene transfer format (GTF) file containing gene models, with each model representing the structure of transcripts produced by a given gene (Box 3).

Raw read counts are affected by factors such as transcript length (longer transcripts have higher read counts, at the same expression level) and the total number of reads. Thus, if we want to compare expression levels between samples, we need to normalize the raw read counts. The measure RPKM (reads per kilobase of exon model per million reads) and its derivative FPKM (fragments per kilobase of exon model per million reads mapped) account for both gene length and library size effects [43].

Box 3. Standard GTF file format which used to describe genes and other features of DNA, RNA and protein sequences

Seqid	source	type	start	end	score	strand	phase	attributes
Chr1	Snap	exon	234	1543	.	+	.	gene_id "gene1"; transcript_id "transcript1";
Chr1	Snap	CDS	577	1543	.	+	0	gene_id "gene1"; transcript_id "transcript1";
Chr1	Snap	exon	1822	2674	.	+	.	gene_id "gene1"; transcript_id "transcript1";
Chr1	Snap	CDS	1822	2674	.	+	2	gene_id "gene1"; transcript_id "transcript1";

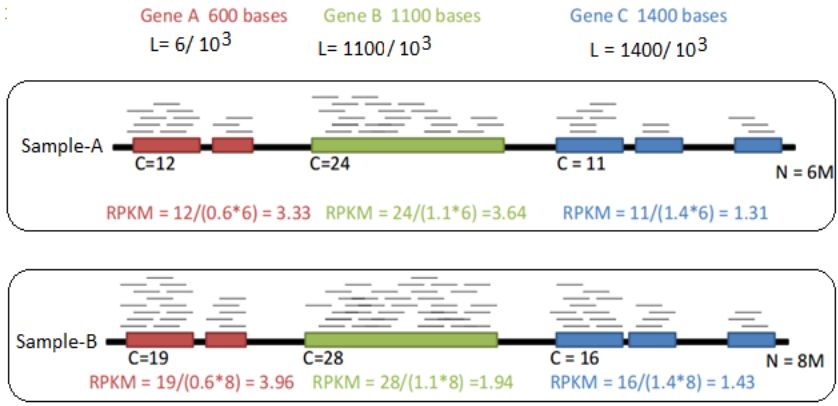


Figure 3. Diagrammatic representation of the use of RPKM for normalization of the effect of gene length and library depth.

Correcting for gene length is not necessary when comparing changes in gene expression within the same gene across samples. However, it is necessary for correctly ranking gene expression levels within the sample to account for the fact that longer genes accumulate more reads (at the same expression level).

$$\text{RPKM} = \frac{\text{The number of reads of the region}}{\text{Length of region} / 10^3 \times \text{total number of mapped read} / 10^9}$$

$$\text{RPKM (X)} = \frac{10^9 \times C}{N \times L}$$

Here:

N - Library depth (in millions)

C - Number of mapped reads (transcript, exons)

L - Length of Reference (transcript/exons in kb).

3.3.3.1. Challenges and Possible Solution

Selection of Tools: Different tools and their different related parameters generate different read's numbers and thus affect downstream analysis because they use different strategies to assign reads to features.

Gene Model: The gene model that hypothesizes the structure of transcripts produced by a gene also affects the analysis. In general, a different gene definition of the gene models frequently results in inconsistency in gene quantification. Among multiple genome annotation databases, RefGene, Ensembl, and the UCSC annotation databases are the most popular ones. The choice of genome annotation directly affects gene expression estimation.

3.3.4. Normalization

Table 7. Available methods for RNA-Seq data normalization

S. No.	Methods	Description
Scaling based methods	Lowess Normalization	Lowess normalization calculates local scaling factors within a certain window size [45].
	Trimmed Mean Method (TMM)	It assumes the majority of the mRNAs in NGS output are similar, except the data points that lie within the extreme M-value and A-value ranges. It derives a simple scaling factor after trimming the data points located in extreme M-value and A-value ranges [46].
	Global Normalization	Global normalization scales all the data of the experimental condition against the control condition by a factor of the difference in the means of two data [45].
	Scaling Normalization	Scaling normalization assumes the ranges of data are the same and that the noise and the stochastic variations of microRNAs are proportional to the signal intensity [45].
Scaling free methods	Quantile normalization:	Quantile normalization is non-scaling and assumes that the overall distribution of signal intensity does not change [47].
	Variance stabilization (VSN)	VSN assumes that most miRNAs do not change and transform the data such that the transformed variance is constant among different expression levels [48].
	Invariant method (INV)	INV assumes that a subpopulation of expressed microRNAs does not change, and it learns a set of "invariants" through algorithms, instead of assigning "housekeeping genes" subjectively [49, 50].

After getting the read counts, data normalization is one of the most crucial and essential steps of data processing, and it creates a considerable impact on high-throughput RNA-seq data analysis.

Normalization process must be carefully considered, as it is essential to ensure accurate inference of gene expression and subsequent analyses thereof.

Although there are numerous methods for read-count normalization, it remains a challenge to choose an optimal method due to multiple factors contributing to read-count variability that affects the overall sensitivity and specificity, though lowness and Quantile normalization methods are reported to be more suitable for RNA-seq normalization [44]. Available normalization methods can be classified into two groups on the basis of the application of linear scaling or not (Table 7).

3.3.5. Differential Gene Expression

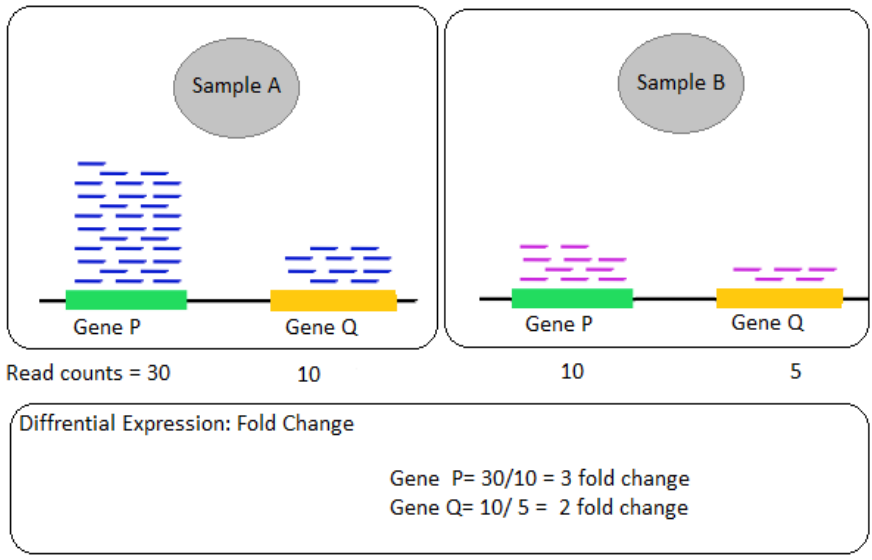


Figure 4. Fold change measurement for differential expression of genes from two different RNA-seq sample sources. In the RNA-seq experiment fold change measurement depends on read counts.

In general, statistical testing takes place to decide whether for a given gene or set of genes an observed difference in read counts is significant or not.

A number of methods for assessing differential gene expression from RNA-seq counts are available which are mostly depending on either direct read counts or RPKM [51]. Most of these methods are parametric in nature which mostly utilizes a negative binomial distribution to make probabilistic statements about the differences in gene expression seen in an experiment. Parametric methods are also there, but because of the low number of replicates, typically available in RNA-seq experiments; they do not offer enough detection power. Despite the availability of a range of methods and software, to the best of our knowledge, there is no one-size-fits-all method is there. Like microarray, RNA-seq also utilizes fold-change to explore fold-change variations in gene expression (Figure 4).

4. CLUSTERING

Differential expression of the gene analysis is followed by various explorations towards the gene enrichment, but before that, they need to classify on the basis of their differential gene expression patterns. The most commonly used *in silico* approach to classifying genes and experiments on the basis of gene expression data is clustering (Figure 5). The goal of clustering is to reduce the amount of data by categorizing or grouping similar data items (genes) together. The term cluster analysis first used by Tryon (1939), encompasses a number of different algorithms and methods for grouping objects of a similar kind into respective categories (4.1). These methods now used in various fields where grouping is required [52].

Clustering technique has proven to be helpful to understand gene function, gene regulation and cellular processes. Genes with similar expression patterns (co-expressed genes) are assumed to be involved in similar cellular functions. This approach may further help to understand the functions of many genes for which information has not been previously available [52, 53]. Furthermore, co-expressed genes in the same cluster are likely to be involved in the same cellular processes, and a strong correlation of expression patterns between those genes indicates co-regulation.

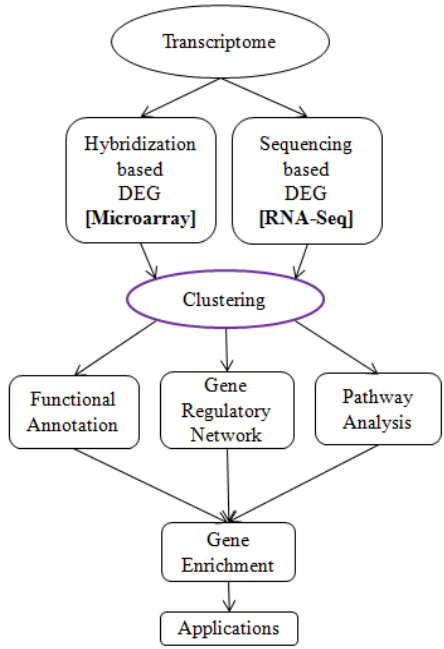


Figure 5. Schematic representation of the steps followed by DEG analysis, as the figure indicates, most of the biological interpretations follows clustering.

4.1. Clustering for Transcriptome Analysis

- Hierarchical Clustering (unsupervised): It builds a hierarchy of clusters, in a greedy manner and represents them by dendrogram (Figure 6), which show relationships of objects and clusters as hierarchies [52].
- Self-Organizing Maps (unsupervised): SOM facilitates the presentation of high dimensional datasets into lower dimensional ones, usually 1-D, 2-D and 3-D. It learns to classify data without supervision [54].
- K-Mean (unsupervised): K-mean is a randomized algorithm which generates cluster centers randomly and assigns objects to the nearest cluster center. The algorithm modifies the location of the centers to

minimize the sum of squared distances between objects and their closest cluster centers [55, 56].

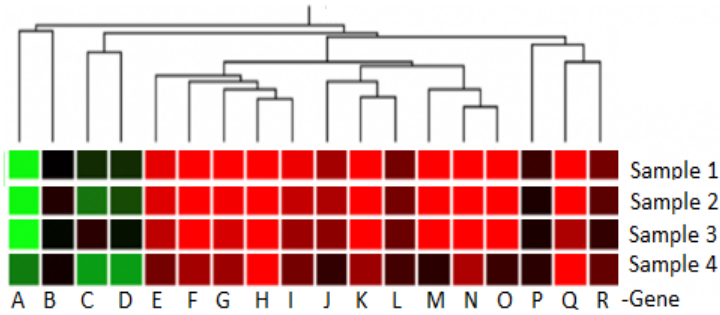


Figure 6. An example of a heatmap from hierarchical clustering in which genes have been grouped based on their pattern of gene expression.

- PCA (supervised): It is a statistical method that can be used for clustering. It is based on multivariate analysis and dimensionality reduction [57].

Bioinformatics provides a range of open access as well as paid resources to perform clustering with different algorithms and their further visualization, in some cases analysis too (e.g., Cluster, SAM, Tree view, Gene cluster, J-express, Genesis).

5. COMPUTATIONAL TRANSCRIPTOMICS THROUGH BIOCONDUCTOR

Bioconductor is the specialized repository for bioinformatics software in the form of packages, developed and maintained by the R community. It offers advanced facilities for analysis of various microarrays (e.g., Affymetrix, Illumina, Nimblegen, Agilent; and one- and two-color technologies), as well as RNA-seq platform [58, 59].

Major workflows for microarray include pre-processing, Normalization, quality assessment, data filtering, differential expression, clustering and classification and gene set enrichment analysis (Figure 5). Bioconductor offers extensive interfaces to community resources, including GEO, ArrayExpress, Biomart, genome browsers, GO, KEGG, and diverse annotation sources.

Table 8. List of frequently used Bioconductor packages for microarray and RNA-Seq data analysis and annotation purpose

S. No.	Purpose	Packages
A).	Packages for microarray data analysis	
1	Pre-processing	a4Preproc, yaqcaffy, limma, affy
2	DEG	TTCA, diffGeneAnalysis, Limma
3	Clustering	Mfuzz, GOexpress
4	Annotation	adSplit, GSEABase
5	Visualization	Heatmaps, GOexpress, maCorrPlot
B).	Packages for RNA-Seq data analysis	
1	Pre-processing	limma, edgeR, gplots, org.Mm.eg.db, RColorBrewer, Glimma,
2	Alignment and Counting	Rsubread, easyRNA-Seq
3	DEG and normalization	DESEQ2, edgeR, compcodeR
4	Clustering	CountClust
5	Annotation and Visualisation	org.Mm.eg.db, TRAPR, derfinder, Goexpress, goseq
6	Gene Set Testing	goseq, SeqGSEA
7	Alternative splicing	IsoformSwitch, AnalyzeR

6. SCOPE OF TRANSCRIPTOMICS

cDNA microarray and RNA-seq are high throughput techniques with enormous potential for transcriptomics studies [60-62]. After completion of the human genome project in the early twenty's scientists shifted their research orientation form structural genomics to functional genomics and for such purpose, they mainly relied on transcriptomics studies.

Though in early days various techniques are utilized for transcriptome analysis, i.e., SAGE, ESTs, cDNA sequencing, but because of high throughput nature of cDNA-microarray and RNA-seq, they become the technique of interest for transcriptome based various analyses (Figure 7, 8).

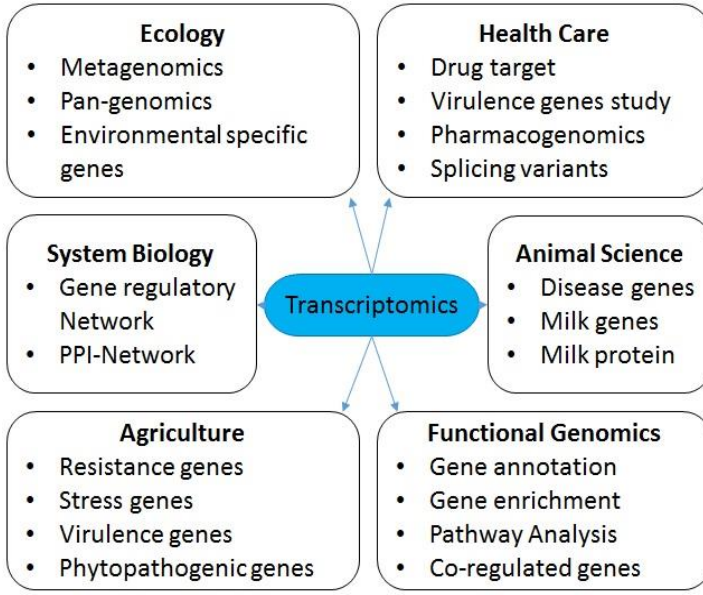


Figure 7. Major scope of transcriptomics in various fields.

In the last one and half decades, almost each and every biological field utilized the potential of transcriptomics for various purposes and solves a large number of biological questions (Figure 7, 8). One of the most utilized features of transcriptomics, which were utilized mainly is DEG, followed by variant analysis and so on.

7. cDNA Microarray versus RNA-Seq

Both cDNA microarray and RNA-seq technologies are in practice to generate transcriptome profiling, both of them follows similar steps to answering a biological question (Figure 8). This includes - experimental

design, data acquisition, and finally analysis and interpretation. However, there are a few considerable key differences between the technologies are there [63].

Table 9. Comparative features of cDNA microarray and RNA-Seq

S. No.	Feature	Microarray	RNA-Seq
1	Principle	Hybridization	High-throughput sequencing
2	Reference genome	Required for the design of probes	None required, If available it may useful
3	Throughput	High	High
4	Background Noise	High	Low
5	The Required amount of transcript	High	Low
6	Dynamic range to quantify gene expression level	>100-fold	>8,000-fold
7	Sequence resolution	Targeted arrays can detect mRNA splice variants	Detect SNPs and splice variants.
8	Sensitivity	One transcript per thousand	One transcript per million
9	Pipeline/ workflow	Well developed	Various pipelines are available, but none of them is ideal.

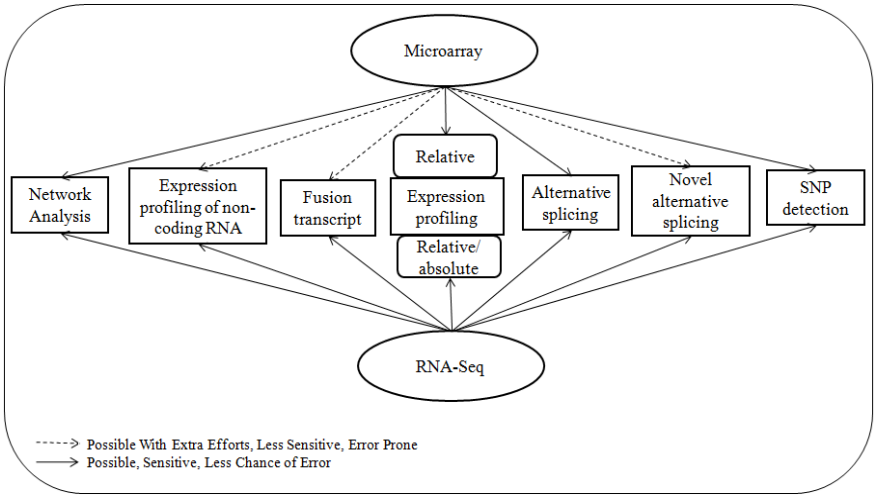


Figure 8. Promising analysis from cDNA microarray and RNA-Seq technology.

8. Selection of Technique for Transcriptome Analysis

Above mentioned features clearly indicate that both techniques have their pros and cons, but overall RNA-seq have the upper hand. Selection of technique mainly depends on the biological question and research goals. If the objective is to analyze relative gene expression in a range of experiments with a well-developed pipeline, at low cost, cDNA microarray will be the blind choice. However, if the objectives of the experiment are centric towards the sensitivity, variation discovery, range of absolute expression, even in the organism lacking a reference genome, RNA-seq is going to be your best choice [64]. Though RNA-seq is comparatively expensive, it will end up being cheaper and more time-efficient than starting with microarrays and having to end up using RNA-Seq later anyway [63, 65].

Conclusion

Nowadays, computational transcriptomics has become an indispensable tool for functional genomics and related aspects. Computational transcriptomics is now a mature field, and its success is greatly influenced by available high throughput techniques, i.e., cDNA microarray and RNA-seq technology.

Bioinformatics provides various databases, software and automated pipelines to perform various processing and statistical testing on high throughput data. This chapter provides an overview of various computational aspects utilized for cDNA microarray and RNA-seq data analysis to interpret the biological significance from experimental data. The Chapter also provides a detailed account of the commonly used databases and software for transcriptomics studies.

Conflict of Interest

The author (s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

[1] Carulli JP, Artinger M, Swain PM, Root CD, Chee L, Tulig C, Guerin J, Osborne M, Stein G, Lian J, Lomedico PT. (1998). High throughput analysis of differential gene expression. *J Cell Biochem Suppl*. 30-31:286-96.

[2] Hoeijmakers WA, Bártfai R, Stunnenberg HG. (2013). Transcriptome analysis using RNA-Seq. *Methods Mol Biol*. 923:221-39.

[3] Gomase VS, Tagore S. (2008). Transcriptomics. *Curr Drug Metab*. 9(3):245-9.

[4] Dong Z, Chen Y. (2013). Transcriptomics: advances and approaches. *Sci China Life Sci.* 56(10):960-7.

[5] Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. (2017). Transcriptomics technologies. *PLoS Comput Biol*. 13(5):e1005457.

[6] Schena M, Shalon D, Davis RW, Brown PO. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science.* 270(5235):467-70.

[7] Miller MB, Tang YW. (2009). Basic concepts of microarrays and potential applications in clinical microbiology. *Clin Microbiol Rev*. 22(4):611-33.

[8] Ball CA, Brazma A. (2006). MGED standards: work in progress. *OMICS. Summer* 10(2):138-44.

[9] Clough E, Barrett T. (2016). The Gene Expression Omnibus Database. *Methods Mol Biol*. 1418:93-110.

[10] Quackenbush J. (2002). Microarray data normalization and transformation. *Nat Genet*. 32 Suppl:496-501.

[11] Kukurba KR, Montgomery SB. (2015). RNA Sequencing and Analysis. *Cold Spring Harb Protoc*. 2015(11):951–969. doi:10.1101/pdb.top084970.

[12] Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic Acids Res.* 39:D19–D21. doi:10.1093/nar/gkq1019.

[13] Leinonen R, Akhtar R, Birney E. et al. (2011). The European Nucleotide Archive. *Nucleic Acids Res*. 39:D28–D31. doi:10.1093/nar/gkq967.

[14] Suntsova M, Gaifullin N, Allina D. et al. (2019). Atlas of RNA sequencing profiles for normal human tissues. *Sci Data.* 6:36. doi:10.1038/s41597-019-0043-4

[15] Dong R, Ma XK, Li GW, Yang L. (2018). CIRCpedia v2: An Updated Database for Comprehensive Circular RNA Annotation and Expression Comparison. *Genomics Proteomics Bioinformatics.* 16(4):226–233. doi:10.1016/j.gpb.2018.08.001.

[16] Wang Y, Wu N, Liu J. et al. (2015). FusionCancer: a database of cancer fusion genes derived from RNA-seq data. *Diagn Pathol.* 10: 131. doi:10.1186/s13000-015-0310-4

[17] Cembrowski MS, Wang L, Sugino K, Shields BC, Spruston N. (2016). Hipposeq: a comprehensive RNA-seq database of gene expression in hippocampal principal neurons. *Elife*. 5:e14997. doi:10.7554/eLife.14997.

[18] Iyer M, Niknafs Y, Malik R. et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet.* 47:199–208. doi:10.1038/ng.3192

[19] Krupp M, Marquardt JU, Sahin U, Galle PR, Castle J, Teufel A. (2012). RNA-Seq Atlas--a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics.* 28(8):1184–1185. doi:10.1093/bioinformatics/bts084.

[20] Kuksa PP, Amlie-Wolf A, Katanić Ž, Valladares O, Wang LS, Leung YY. (2019). DASHR 2.0: integrated database of human small non-

coding RNA genes and mature products. *Bioinformatics*. 35(6):1033–1039. doi:10.1093/bioinformatics/bty709.

[21] Andrews S. (2010). *FastQC: a quality control tool for high throughput sequence data.* Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[22] Gordon A, Hannon GJ. (2010). "*FASTX-Toolkit*," FASTQ/A short-reads pre-processing tools (unpublished) http://hannonlab.cshl.edu/fastx_toolkit/.

[23] DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics.* 28(11):1530-2.

[24] Chen S, Huang T, Zhou Y, Han Y, Xu M, Gu J. (2017). AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics.* 18(Suppl 3):80.

[25] Patel RK, Jain M. (2012). NGSQCToolkit: a toolkit for quality control of next generation sequencing data. *PLoS One.* 7(2):e30619.

[26] Bolger AM, Lohse M, Usadel B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*. 30(15):2114-20.

[27] Schubert M, Lindgreen S, Orlando L. (2016). Adapter Removal v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes*. 9:88.

[28] Grabherr MG, Haas BJ, Yassour M. et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 29(7):644–652. doi:10.1038/nbt.1883

[29] Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 9(4):357-9.

[30] Ahmed M, Nguyen HQ, Hwang JS, Zada S, Lai TH, Kang SS, Kim DR. (2018). Systematic characterization of autophagy-related genes during the adipocyte differentiation using public-access data. *Oncotarget.* 9(21):15526-15541.

[31] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. (2012). STAR: ultrafast universalRNA-seqaligner. *Bioinformatics.* 29(1):15-21.

[32] Trapnell C, Pachter L, Salzberg SL. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 25(9):1105-11.

[33] Li B, Dewey CN. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 12:323.

[34] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 14(4):417-419.

[35] Bray NL, Pimentel H, Melsted P, Pachter L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 34(5):525-7.

[36] Patro R, Mount SM, Kingsford C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol.* 32(5):462-4.

[37] Anders S, Pyl PT, Huber W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics.* 31(2):166-9.

[38] Love MI, Huber W, Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15(12):550.

[39] Robinson MD, McCarthy DJ, Smyth GK. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 26(1):139-40.

[40] Trapnell C, Williams BA, Pertea G. et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 28(5):511–515. doi:10.1038/nbt.1621

[41] Tarazona S, Furió-Tarí P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, Conesa A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.* 43(21):e140.

[42] Anders S, Huber W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11(10):R106.

[43] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X,

Mortazavi A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:13.

[44] Garmire LX, Subramaniam S. (2012). Evaluation of normalization methods in mammalian microRNA-Seq data. *RNA*. 18(6):1279-88.

[45] Smyth GK, Yang YH, Speed TP. (2003). Statistical issues in microarray data analysis. *Methods Mol Biol.* 224:111–136.

[46] Robinson MD, Oshlack A. (2010). A scaling normalization method for differential expression analysis of RNA-Seq data. *Genome Biol.* 11:R25.

[47] Bolstad BM, Irizarry RA, Astrand M, Speed TP. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* 19:185–193.

[48] Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M. (2002). Variance stabilization applied to microarray data calibratio and to the quantification of differential expression. *Bioinformatics (Suppl1)* 18:S96–S104.

[49] Perkins DO, Jeffries CD, Jarskog LF, Thomson JM, Woods K, Newman MA, Parker JS, Jin J, Hammond SM. (2007). MicroRNA expression in the prefrontal cortex of individuals with schizophrenia and schizoaffective disorder. *Genome Biol.* 8:R27.

[50] Pradervand S, Weber J, Thomas J, Bueno M, Wirapati P, Lefort K, Dotto GP, Harshman K. (2009). Impact of normalization on miRNA microarray expression profiling. *RNA.* 15:493–501.

[51] Oshlack A, Robinson MD, Young MD. (2010). From RNA-seq reads to differential expression results. *Genome Biol.* 11(12):220.

[52] Eisen MB, Spellman PT, Brown PO, Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*. 95(25):14863-8.

[53] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. (1999). Systematic determination of genetic network architecture. *Nat Genet.* 22(3):281-5.

[54] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. (1999). Interpreting patterns of gene expression

with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*. 96(6):2907-12.

[55] Pirim H, Ekşioğlu B, Perkins A, Yüceer C. (2012). Clustering of High Throughput Gene Expression Data. *Comput Oper Res.* 39(12):3046-3061.

[56] Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LDF, Rodrigues FA. (2019). Clustering algorithms: A comparative approach. *PLoS One.* 14(1):e0210236.

[57] Raychaudhuri S, Stuart JM, Altman RB. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput*. 455-66.

[58] Gillespie CS, Lei G, Boys RJ, Greenall A, Wilkinson DJ. (2010). Analysing time course microarray data using Bioconductor: a case study using yeast2 Affymetrix arrays. *BMC Res Notes*. 3:81. doi:10.1186/1756-0500-3-81

[59] Love MI, Anders S, Kim V, Huber W. (2015). RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Res*. 4:1070.

[60] Katara P, Sharma N, Sharma S. et al. (2010). Comparative microarray data analysis for the expression of genes in the pathway of glioma. *Bioinformation.* 5(1):31–34. doi:10.6026/973206300 05031.

[61] Katara P, Grover A, Kuntal H, Sharma V. (2011). In silico prediction of drug targets in Vibrio cholerae. *Protoplasma.* 248(4):799–804. doi:10.1007/s00709-010-0255-0.

[62] Katara P, Grover A, Sharma V. (2012). In silico prediction of drug targets in phytopathogenic Pseudomonas syringae pv. phaseolicola: charting a course for agrigenomics translation research. *OMICS*. 16(12):700–706. doi:10.1089/omi.2011.0141.

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

[64] Verma Y, Yadav A, Katara P. (2020). Mining of cancer core-genes and their protein interactome using expression profiling based PPI network approach. *Gene Reports.* 18:100583.

[65] Nagalakshmi U, Waern K, Snyder M. (2010). RNA-Seq: a method for comprehensive transcriptome analysis. *Curr Protoc Mol Biol*. Chapter 4:Unit 4.11.1-13.