

# Applications and challenges of microarray and RNA-sequencing

Ankita Negi<sup>1</sup>, Abhimati Shukla<sup>2,\*</sup>, Akanksha Jaiswar<sup>3,\*</sup>, Jatin Shrinet<sup>4</sup> and Rahul Singh Jasrotia<sup>4</sup>

<sup>1</sup>Centre for Agricultural Bioinformatics (CABin), ICAR-Indian Agricultural Statistics Research Institute (IASRI), New Delhi, India, <sup>2</sup>Department of Biochemical Engineering, Harcourt Butler Technical University, Kanpur, India, <sup>3</sup>Mossakowski Medical Research Centre, Polish Academy of Sciences, Warsaw, Poland, <sup>4</sup>Department of Biological Sciences, Florida State University, FL, United States

## 6.1 Introduction

Transcriptomics including microarray and RNA-sequencing (RNA-seq) are the cheap, fast, and easy methods to detect the differential expressed genes. It plays a very crucial part in analyzing large-scale data of animal and plant transcriptomes. There are several lacunas in microarray that has been dramatically overcome by RNA-seq. Generation of a huge amount of sequencing reads and data reproducibility through flow cells and lanes in RNA-seq reduced the usage of technical replicate (Costa-Silva, Domingues, & Lopes, 2017). After the invention of next-generation sequencing (NGS) method in 2005, it has found drastic changes in the way of RNA-seq data generation and analysis. Using the RNA-seq approach, it is easy to quantify the expression of known isoform or transcripts and for unknown transcripts. Due to its huge popularity, the sequencing cost of RNA-seq is dropped considerably as compared to complementary deoxyribonucleic acid (cDNA) microarray (Agarwal et al., 2010; Kratz & Carninci, 2014).

To understand the complex functional procedures, we require the parallel and global analysis of various cellular processes. This has been achieved by DNA microarrays by making significant contributions, especially to gene expression studies, transcript profiling and genotyping, on-chip synthesis, and epigenetic analysis. Moreover, it possesses the ability for the molecular detection and diagnosis of various infectious disease pathogens (Hoheisel, 2006). Several microarray platforms are available including in situ synthesized arrays, printed oligonucleotide and complementary double-stranded DNA arrays, suspension bead, high-density bead, and electronic microarrays (Miller & Tang, 2009).

Microarray is a technique to detect and calculate the expression level of hundreds of genes or transcripts of a genome. Gene expression microarray and tissue microarray (TMA) are the two types of microarrays. Gene microarray technology can make use of thousands of various DNA transcripts on a tiny solid surface, generally a microscopic slide (generally called as “chip”) of glass upon which DNA binds in a prearranged and organized grid manner. The microscope slide size of the solid glass surface is generally (approximately 75 mm × 26 mm). Each DNA spot in the array contains about  $10^{-12}$  moles (picomoles) of a single gene or known DNA sequence (uniquely cloned cDNA or oligonucleotide), called an oligo or probe or reporter. Probes are a small fragment of transcripts or genes that used to hybridize a targeted cDNA sample under high-stringency conditions. DNA microarrays are called DNA arrays, DNA chips, gene arrays, biochips, and gene chips. All the DNA fragments are placed in columns and rows in such a way that it defines each fragment on the basis of their places on the array chip (Bumgarner, 2013; Govindarajan, Duraiyan, Kaliyappan, & Palanisamy, 2012). Microarray not only shows the order of magnitude of many genes or transcripts but also gives the benefit of genes studied that are not influenced, whereas, on the other hand, RT-PCR (reverse transcriptase polymerase chain reaction) and northern technique blot limit the validation of few genes per experiment (Allanach et al., 2008). TMAs are high-throughput technology that provides the comprehensive assessment of the expression of a single antigen throughout many cells and tissues in a single experiment. It is used to analyze the protein expression at the same time in multiple individual cells and tissue samples on a single slide (Govindarajan et al., 2012; Jawhar, 2009).

---

\*These author contributed equally to this work.

NGS also called as HTS (high-throughput sequencing) is a sequencing technology using DNA, RNA, or methylation sequencing that has brought a revolution in genomic research. Parallely, NGS has the capability of sequencing millions of DNA fragments. It has the ability of sequencing at a high speed and content of sequencing at a fraction of cost. The principle behind NGS is capillary electrophoresis, which is similar to the Sanger sequencing method. First, a DNA library is made by fragmentation, purification, and amplification of the DNA sample. The fragments are then attached to beads or solid surfaces. The sequences of these fragments are then resolved by using different techniques. The time required for the sequencing of high gigabase sequences was reduced by NGS from many weeks to a day or few hours, which includes massive cost reduction. The human genome project was sequenced by J.C. Venter genome using the Sanger approach took around 15 years and it costs approximately 1 million dollars, whereas on the other side sequence using 454 Genome Sequencer FLX (Jarvie & Harkins, 2008) by NGS approach took 2 months with less manpower and very cheap price (Behjati & Tarpey, 2013; Franklin et al., 2020; Mardis, 2011; Sanger, Nicklen, & Coulson, 1977). In this chapter, we have covered both microarray and RNA-seq in detail from biological background to its analysis and applications.

## 6.2 Evolution of microarray

In 1975 Grunstein and Hogness firstly used the colony hybridization method to create an early DNA array. In the colony hybridization method, randomly cloned the DNA of interest from *Escherichia coli* and plated it onto an agar petri dish that was covered with nitrocellulose filters. Colonies were observed lysed on the nitrocellulose filters and DNA was denatured (Grunstein & Hogness, 1975). These colonies have been fixed to generate random collection of DNA spots which correspond to the cloned fragments. The radiolabeled probe of interest was hybridized to quickly screen the colonies to identify the clones that contain cDNA. This is an early example of utilizing a labeled probe to identify complementary base pairing (Grunstein & Hogness, 1975).

In 1979 this approach was used by Gergen et al. to produce ordered arrays by using 144 pin mechanical device well microplates, thus replicating multiple microtiter plates on agar that produced approximately 1700 various colonies in a 26 cm × 38 cm portion of arrays. Colonies were then further transferred to Whatman filter paper to allowing the production of DNA arrays for thousands of different bacterial colonies that could be re-used for several times (Gergen, Stern, & Wensink, 1979).

### 6.2.1 Automated arrays and cDNA cloning to microarray technology

In the next decade, similar protocols and arrays based on filters were used in different applications, such as cloning-based differential gene expression of two different conditions, cloning of specific genes of interest, physical mapping, and identifying single-nucleotide polymorphisms (SNPs) (Crampton, Humphries, Woods, & Williamson, 1980; Miller & Barnes, 1986). In the early 1990s, multiple hybridization targets were analyzed with automation that increased speed, efficiency, and accuracy in the development of microarrays. Hans Lehrach's group made use of robotic systems to speedy clones the array from microtiter plates onto the filters (Lennon & Lehrach, 1991). The advancement of cDNA cloning involves completing the human genome as well as transcriptome sequence both of which led to the invention of human reference of cDNA arrays and for other available genomes. The rapid improvement in DNA array technology helped in extracting the basic information from the DNA sequence of various genomes and assured that arrays represent the genes in a genome, which increases our knowledge about all the sequences in a genome (Auffray & Rougeon, 1980; Auffray, Nageotte, Chambraud, & Rougeon, 1980; Barnhart, 1989; Humphries et al., 1977).

### 6.2.2 Principle of microarray

The basic principle of microarrays lies in a single-stranded DNA molecule that binds to another single-stranded DNA of its complementary sequence to make them a double-stranded DNA helix. The binding or base pairing can be joined or separated. DNA Chips are of two types:

#### 6.2.2.1 cDNA-based microarray

##### 6.2.2.1.1 Collection of samples

Samples can be collected from an organism under study (cells or tissue). These samples (either treated cells or healthy cells or infected cells) are collected and compared with each other on the basis of the study objective (Mello-Coelho & Hess, 2005).

### 6.2.2.1.2 Extraction and isolation of the sample

RNA is isolated by using solvent or column like phenol-chloroform technique. mRNA has a characteristic of poly-A tail, so column beads with poly-T tails can be used to bind mRNA. After extracting mRNA, a buffer is used to rinse the column to separate and isolate mRNA from the beads.

### 6.2.2.1.3 Creation of labeled target DNA/cDNA

mRNA degrades easily, so it is converted to a more stable form of cDNA strand using reverse transcription; furthermore, by using restriction endonucleases, the unknown DNA molecules are cut into fragments labeled with fluorochrome dyes green (Cy3) and red (Cy5). PCR can be used to amplify the fragments and that immobilized on a nylon filter of a glass slide (1 × 3 inches).

### 6.2.2.1.4 Hybridization

When single-stranded labeled target DNA is added to the array, the probe DNA on the array is chemically attracted to the target DNA with the opposite (complementary) sequence. Any two pieces of exactly cDNA hybridize together to form double-stranded DNA. The hybridized probe and abundance of nucleic acid sequences of targets are generally identified by detecting fluorophore, chemiluminescence, or silver-labeled targets. If the probe and target sequences do not match completely (particularly for short pieces of probe DNA), then there will be no hybridization and no signal to analyze. Nonspecific or incomplete bonding between probe and target will be removed in the washing step. The hybridized probe produces an indication that is based on the several hybridization conditions, such as salt in hybridization buffer, temperature, and washing process after hybridization. Also, the target DNA that binds to the probes in a slot determines the total strength of the signal.

### 6.2.2.1.5 Collection and analysis of microarray data

As DNA microarrays need attachment of nucleotide probes to known corresponding DNA therefore the abundances of these DNA or cDNA can only be checked and monitored. For the collection of data, a microarray scanner is used comprising a computer, a camera, and a laser. Microarray technique uses relative quantitation where a laser beam is passed and excites fluorescence, thus emitting the intensity from a particular spot under various conditions, which is matched to the intensity of the same spot. The excited fluorescence generates the signals when the laser beam scans the microarray, the camera captures and records the images (the pattern of fluorescence emission) produced, and this is stored as data in computer and finally analyzed. Based on the various intensities of the colors of each spot determines the gene character of that particular spot as shown in [Fig. 6.1](#).

Now many commercial companies make their own microarrays with their proprietary designs for sale purposes. DNA is “printed” onto the glass in each microscopic spot of approximately 1 nm. The DNA probes are firmly attached to the solid support and are difficult to remove under experimental conditions. For array purposes, short fragments of DNA (oligonucleotides) are being synthesized in the laboratory and longer fragments of DNA are prepared from existing pieces of cloned DNA. A variation on a glass slide support is miniature beads, onto which specific oligonucleotides are synthesized. The beads are then mixed with other beads sporting differing sequences and are poured onto small tray support. The position of each type of bead, held on the tray by electrostatic forces, is defined before it leaves the factory. Microarrays have a varying number of DNA spots, depending on what they are being used for. However, some commercially produced microarray sets are containing up to 5 million separate spots. Each individual spot can contain a unique DNA sequence. However, practically, there is some redundancy (repetition) to serve as an internal check. If two or more spots with the same DNA probe sequence give similar results in an experiment or test and are on different areas of an array, then one can be more confident of that result ([Mah et al., 2004](#)).

### 6.2.2.2 Oligonucleotide-based microarray

Oligonucleotide-based arrays are probably the most versatile type of array. Short DNA oligonucleotides (approximately 25-nucleotide long) represent a gene or group of gene splice variants. Instead of putting intact sequences onto the array, the synthesis of nucleotide sequences directly takes place onto the array surface and can be developed to hybridize to any of the DNA sequences of interest. The important characteristic of microarray is that each gene is represented by more than one probe. Oligo-microarrays method can be used to scan a chromosome or genome either by using known mutations as a design basis or by using tag SNPs from the HapMap project. It can be used for resequencing of a gene or genomic region. This is done by preparing oligonucleotides that match all possible options at every nucleotide

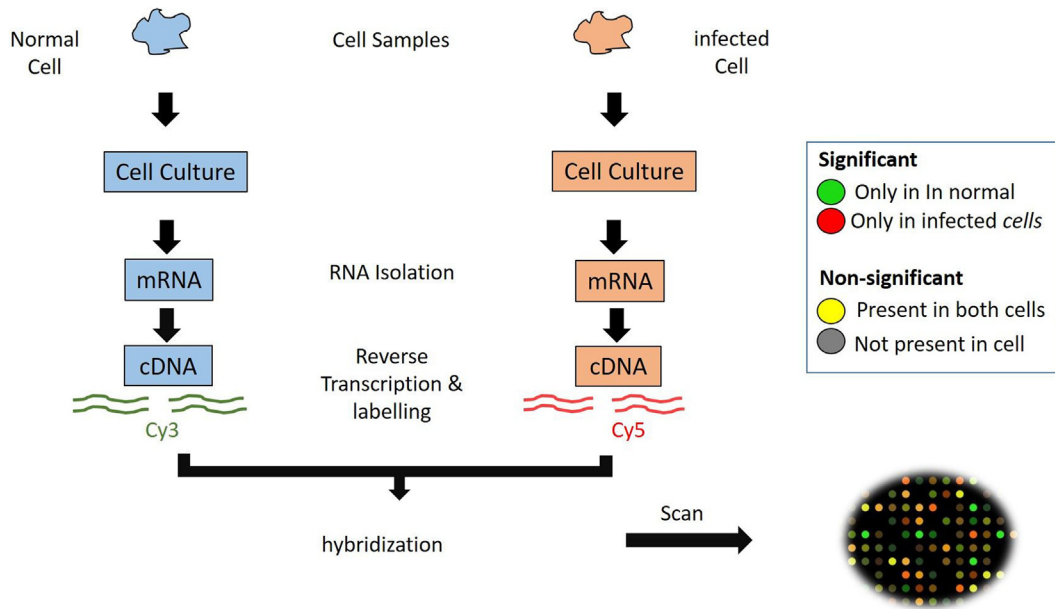


FIGURE 6.1 General overview of microarray.

position. In a genomic area of interest, at each nucleotide position, all four nucleotide options (A, C, G, or T) can be included in the array. This gives massive redundancy, as each nucleotide can be interrogated by up to two hundred different probes. Detection of insertions and deletions of nucleotides can be done by this method when no genomic DNA binds to any of the nucleotides. Furthermore, this may be used in the application for screening mutation in monogenic disorders (Mah et al., 2004).

## 6.2.3 List of microarray tools and their utility

### 6.2.3.1 Pre-processing of microarray data

Filtering, background correction, and normalization of each chip are done for the removal of noise. Principal components analysis is used for sample variations.

### 6.2.3.2 Data analysis

In this step, scaling and normalization of experimental data on the basis of clustering and statistical estimation methods, that is, Robust Multi-chip Average (RMA) for Affymetrix experiment, Model-Based Expression Index (MBEI), Locally weighted regression methods. For identification of significant genes, statistical tests and multiple testing Corrections such as Z-score, *T*-test, ANOVA, and Bonferroni correction can be used. Table 6.1 lists the number of tools available for microarray data analysis.

## 6.3 DNA sequencing

### 6.3.1 First generation

Sanger and Maxam–Gilbert sequencing are considered as the first generation of DNA sequencing (Maxam & Gilbert, 1977; Sanger et al., 1977). In 1977 Sanger sequencing was developed by Frederick Sanger and his colleagues and it is also called as chain-termination method. In this method, the synthesis of the double strand of DNA is initiated in the presence of DNA polymerase from a single strand of DNA; the synthesis terminates if the base is in the form of dideoxynucleotide. Many improvements have been made in the traditional Sanger sequencing method, such as replacing phosphor or tritium radiolabelling with fluorometric-based detection. Detection has been improved by using capillary-based electrophoresis. These improvements contribute largely to automatic DNA sequencing machines (Chidgeavadze et al., 1984; Liu et al., 2012; Sanger et al., 1977).

**TABLE 6.1** Details of tools available for microarray data analysis.

	List of tools
Clustering tools	Cluster Analysis of Gene Expression Dynamics (Ramoni, Sebastiani, & Kohane, 2002) and Cluster/Tree-View (Eisen, Spellman, Brown, & Botstein, 1998)
Analysis suites	TM4 Suite contains TIGR Spotfinder, Multi experiment Viewer, Microarray Data Analysis System, and Microarray Data Manager (Saeed et al., 2003); D-Chip (DNA-Chip Analyzer) (Li & Wong, 2001); SNOMAD (Colantuoni, Henry, Zeger, & Pevsner, 2002); Significance Analysis of Microarrays: Excel-based add on utility for differential expressed genes (Tusher, Tibshirani, & Chu, 2001); Genesis: a comprehensive tool for large microarray datasets (Sturn, Quackenbush, & Trajanoski, 2002); gene expression model selector (Statnikov, Aliferis, Tsamardinos, Hardin, & Levy, 2005); GenMAPP (Dahlquist, Salomonis, Vranizan, Lawlor, & Conklin, 2002); and GeneSpring GX
R packages	AMDA (automated microarray data analysis) (Pelizzola, Pavelka, Foti, & Ricciardi-Castagnoli, 2006), EMA (Easy Microarray data analysis) (Servant et al., 2010), Limma (Linear Models for Microarray Data) (Smyth, 2005), and GCRMA (Wu, Irizarry, & Gentry, 2020)

Maxam–Gilbert method, also known as the chemical degradation method, also belongs to the first-generation sequencing method. In 1976–77 this method was developed by Allan Maxam and Walter Gilbert. It is based on the chemical modification and cleavage of DNA backbone at the sites close to the modified nucleotides. This reaction gives the number of fragments of various lengths that can be divided on the basis of size by electrophoresis (Maxam & Gilbert, 1977).

### 6.3.2 Second generation

Sanger sequencing was dominant for many years but due to high cost and time consumed leads to the emergence of the second generation of DNA sequencing. The methods in this generation were developed with the following basic characteristics: (1) parallelly generate millions of reads, (2) increase in speed in comparison to the first generation, (3) reducing the cost of sequencing, and (4) detection of the sequencing output without using electrophoresis (Heather & Chain, 2016; Sharma et al., 2017). This generation includes the following methods.

#### 6.3.2.1 Roche/454 sequencing

This is the first NGS-based sequencing technology that was developed in 2005 by Roche/454 life sciences. It is based on the “sequencing by synthesis” method and allows bulk parallelization of sequencing reactions that improves in increasing the quantity of DNA that can be sequenced in one experiment. It produces average read length in between 100 and 700 bp and generates ~1 million reads per run, and 0.02–0.7 (Gb) data are generated per run.

#### 6.3.2.2 Ion torrent sequencing

This method is somewhat similar to the 454 pyrosequencing method except that it does not utilize the fluorescently labeled nucleotides like other methods from the same generation. Ion torrent sequencing method is based on the hydrogen ion released during the sequencing. The average read length of ion torrent is 200–400 bp and it generates ~0.06 to ~10 Gb data per run.

#### 6.3.2.3 Illumina/Solexa sequencing

It is based on the sequencing by synthesis method and involves bridge amplification of one of the strands. It is automated in nature and can help in sequencing multiple strands. It produces the average read of ~150–300 bp and generates ~25 to ~6 billion, and data generated per run are ~7.5 Gb to ~1.8 TB.

#### 6.3.2.4 ABi/SOLiD sequencing

Sequencing by Oligonucleotide Ligation and Detection (SOLiD) is based on Polonysequencing (Polymerase + Colony). The average read length is 75 bp, and it generates ~3 to 6 billion and ~160 to 320 Gb data per run (Liu et al., 2012; Sharma et al., 2017).

### 6.3.3 Third generation

The third-generation sequencing technologies have been developed for producing long reads without the use of PCR amplification, less expensive, easy to use, and straight forward sample preparation protocol as compared to previous generations (Goodwin, McPherson, & McCombie, 2016). Synthetic approach and single-molecule real-time sequencing approach (SMRT) are the methods used in this generation. Synthetic approach is developed by Illumina (Moleculo) and 10xGenomics, whereas SMRT is developed by Quake Laboratory (Bentley et al., 2008; Braslavsky, Hebert, Kartalov, & Quake, 2003; Harris et al., 2008).

#### 6.3.3.1 PacBio SMRT

Pacific Biosciences developed the first sequencer using the SMRT approach. This method generates longer read lengths as compared to previous generations but contains a higher error rate and higher cost per base. It can be used to study larger genomes (Pettersson, Lundeberg, & Ahmadian, 2009). It has an average read length of  $\sim 1300$ – $13,500$  bp, produces 0.5–7-GB data in a single run, and generates  $\sim 350$ – $600$  reads per run (Sharma et al., 2017).

#### 6.3.3.2 Oxford Nanopore

This method was established for determining the order of nucleotides. Oxford Nanopore Technologies released MinION, which can generate longer reads. It produces an average read length of  $\sim 9545$  bp and generates 1.5 Mb to 4 Tb data per run (Sharma et al., 2017)

## 6.4 RNA-sequencing

RNA-seq is one of the important applications of NGS technologies to study RNA molecules, such as tRNA (transfer RNA), mRNA (messenger mRNA), rRNA (ribosomal RNA), and ncRNA (noncoding RNA), which present in the cells. RNA-seq provides the information of tens to hundreds of millions of transcripts and data on billions of individual bases of an organism. It is the study of a whole set of RNAs that are transcribed in a cell or tissue and their quantity for certain physiological and developmental stages (Gedil et al., 2016). Using the NGS approach, it is easy to identify the quantity and presence of RNA in a biological sample at a specific condition and found a more sensitive and accurate way to study the differential expression analysis which drastically overcomes the lacuna of microarray (Voelckel, Gruenheit, & Lockhart, 2017). RNA-seq played the important role in the field of human science to plant to animals transcriptomic to identify the specific genes. Moreover, it can be carried out without the prior information of the reference genome and that is very informative for a nonmodel organism whose genome is not yet sequenced (Voelckel et al., 2017).

RNA is polymeric in nature involving many biological processes of an organism, such as gene expression, regulation of genes, coding, and decoding (Yang & Kim, 2015). Traditional methods based on cloning-based techniques, such as Suppression Subtractive Hybridization or cDNA libraries are more laborious and complex, which need lots of time to do, whereas on the other side RNA-seq is very easy and straight forward. RNA-seq plays an important role in gene expression, coexpression analysis, genic identification, genic variants, alternative splicing, and miRNA target identifications (Yang & Kim, 2015). Fig. 6.2 provides an overview of RNA-seq and its applications.

### 6.4.1 Library preparation and sequencing

In RNA-seq library preparation, cellular RNA is converted into molecules that can be sequenced. The rRNAs comprise up to 80% of total RNA. These rRNAs' sequencing reads are noninformative resources and lead to less detection so these should be removed by enzymatic digestion or hybridization-based depletion methods, to maintain the depth of sequence coverage (Heather & Chain, 2016). Library preparation is done depending on the size of the desired library. Fragmentation of RNA can be done either by enzymatic digestion or by chemical hydrolysis to the size required. In the case of miRNAs, fragmentation is not required as the size of bases is small (under 200 bases), whereas for long RNAs fragmentation is done to a size of approximately 200–250 bp for sequencing by various sequencing platforms, such as Illumina or SOLiD. RNAs after fragmentation were then converted to cDNA by a reverse transcriptase enzyme using random primers. The adapter is ligated to cDNA for amplification and sequencing. An oligo-dT primer and polyadenylated RNA is reverse transcribed with a universal primer at 5' end. At the 3' end of the cDNA poly(A) tail is added by using polynucleotide tailing. Amplification of this cDNA is done with an oligo-dT sequence at 3' end and universal PCR primers. Then, the amplified cDNA is further used for standard DNA library construction. Moreover, the cDNA

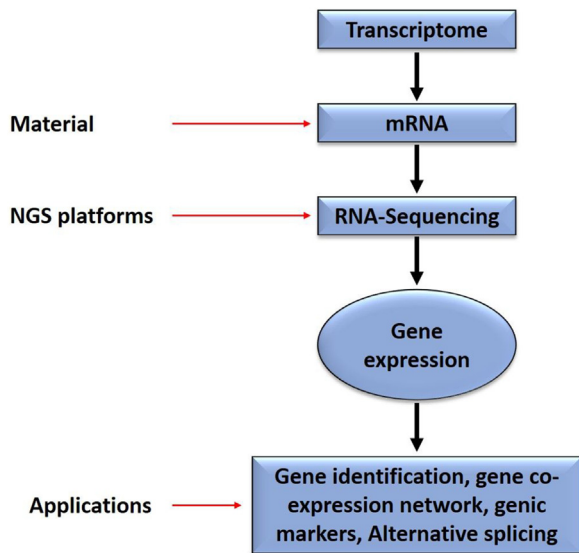


FIGURE 6.2 Overview of RNA-sequencing and its application.

synthesis is initiated by an anchored oligo(dT) primer, which adds a universal primer sequence. The reverse transcription polynucleotide tails the cDNA by producing a 3' overhanging tail. Then, the second universal primer sequence hybridized and initiates template switching at the 3' end of the cDNA. Therefore by using PCR the cDNA is amplified and taken into a standard library protocol. The anchored oligo(dT) primer with adapter sequence and T7 promoter sequence at the 5' end begins the cDNA synthesis. The T7 promoter initiates in vitro transcription and generates the complementary RNA (cRNA) copies of the cDNA with a 5' barcode and adapter. Finally, at the 3' end of the cRNA, a second adapter is attached. The RT-PCR amplification is then done, which thus completes the library construction (Chu & Corey, 2012).

## 6.4.2 Pipeline and usage of RNA-sequencing

There are two different ways to do the RNA-seq analysis, such as reference-based analysis and de novo assembly-based analysis. Reference-based RNA-seq analysis is that when the reference genome is already known whereas in the de novo assembly-based approach is that when the reference genome of an organism is not sequenced. In the de novo transcriptome assembly can be done by pooling all the data and generate assembly. Fig. 6.3 shows the workflow of RNA-seq analysis.

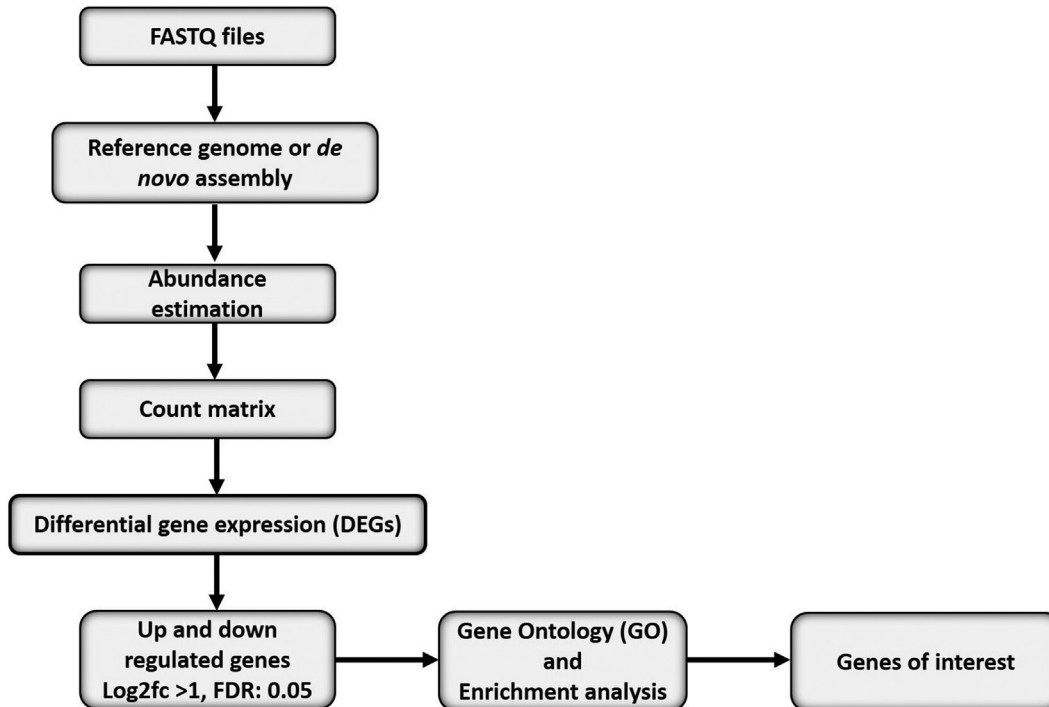
### 6.4.2.1 Steps involved in RNA-sequencing

A better understanding of the transcriptome of various organisms using RNA-seq data has grown rapidly over the past few years. One of the important steps in experiment designing is to decide whether to do single-end or paired-end sequencing. Single-end sequencing sequences one end of each cDNA fragment, whereas paired-end sequencing sequences both the ends of the fragment (Hu et al., 2010).

Paired-end sequencing produces the number of reads in the same amount of time and labors as single-end sequencing and also offers additional and more accurate information about read alignment and mapping across splice junctions, high-homology regions, and de novo transcript assembly, helpful in the identification of novel transcripts, variants like insertion-deletion (InDels), and estimation of alternative transcripts (Williams, Thomas, Wyman, & Holloway, 2014). However, paired-end sequencing is costlier than the single end (Ching, Huang, & Garmire, 2014).

#### 6.4.2.1.1 Quality check

FASTQ format is a text-based format that stores biological sequence (generally nucleotide sequence) and its corresponding Phred quality scores in a single file. It is a widely used format in the sequence analysis as it contains much more information than FastA; it is required by many analysis tools. The FASTQ sequence identifier contains information related to the sequencer and its position on the flow cell. The sequence description follows a particular format and contains sample information (<https://learn.gencore.bio.nyu.edu/ngs-file-formats/fastq-format/>). There are tools available



**FIGURE 6.3** Pipeline of RNA-sequencing data analysis. Raw data are generated in the form of FASTQ format and, then, the pipeline contains various steps to analyze the data.

like FASTQC that can be used to analyze sequences obtained in the FASTQ format. The FASTQ file contains the reads as well as its quality score. FASTQC measures average quality score, read length, GC content, adaptors, and overrepresented sequences.

#### 6.4.2.1.2 Adaptor and quality trimming

The adaptors are attached to the sequence for proper attachment of the DNA fragment, which can be sequenced to the flow cell, and these adaptors are usually sequenced during the sequencing process and then removed. If the adaptors have not been removed from the reads, they need to be removed. Also, the bases with a quality score (phred quality score) of  $< 20$  are considered poor and need to be removed from the reads. The reads with length  $< 20$  bases are considered short reads, and these should also be removed. Cutadapt, PrinSeq, and Trimmomatic (Bolger, Lohse, & Usadel, 2014; Martin, 2011; Schmieder & Edwards, 2011) tools can be used for quality control and trimming of reads.

#### 6.4.2.1.3 De novo assembly

For nonmodel organisms whose genomes are not yet available, in that case, de novo transcriptome can be done by using several open-source tools, such as Trinity (Grabherr et al., 2011), rnaSPAdes (Bushmanova, Antipov, Lapidus, & Prjibelski, 2019), and Trans-ABYSS (Robertson et al., 2010). The quality of transcriptome assemblies can be assessed by basic statistical measures that include assembly size (base pairs), reads' percentage that is assembled into contigs, the total number of contigs, singletons, scaffolds, GC percentage, and N25, N50, N75, NG50, L90, and L99 scaffold lengths. N50 is the shortest contig length that has a size equal or more than 50% of the total sum of all the contigs length in the assembly, that is, the contig length is calculated by adding up all the contigs length and the length that takes the sum length 50% and more of the total assembly length is N50. Half of the assembly sequence is covered by contigs larger than or equal to the N50 contig size. NG50 is the length of the contig at which the sum length accounts for 50% of the estimated genome size (Bradnam et al., 2013). The GC content affects the fold-change estimation, which can mislead differential expression analysis (Hansen, Irizarry, & Wu, 2011). The GC content provides information about coverage variation, an evaluation measure used by assemblers to reconstruct the genomes, that further affects the k-mer selection to sequence assembly based on NGS reads (Miranda et al., 2018).

#### 6.4.2.1.4 Mapping/alignment of reads onto the assembly

In this step, all the cleaned reads are mapped onto reference genome assembly or de novo transcriptome assembly by using short read aligner, such as Burrow-Wheeler Transform, Bowtie2, Tophat2, STAR, and HISAT2 (Dobin et al., 2013; Kim et al., 2013; Kim, Langmead, & Salzberg, 2015; Langmead & Salzberg, 2012; Li & Durbin, 2009).

#### 6.4.2.1.5 Abundance estimation

This step is used for the calculation of read count per transcripts or per gene. Several tools, such as RSEM, FeatureCounts, and HTSeq, can be used to get the count matrix and abundance estimation of all the transcripts of each sample separately (Anders, Pyl, & Huber, 2015; Li & Dewey, 2011; Liao, Smyth, & Shi, 2013).

#### 6.4.2.1.6 Differentially expressed genes

To identify the differentially expressed gene (DEG), all the count matrices of two different conditions can be used. EdgeR, EBseq, DESeq2, and NOISeq (Leng et al., 2013; Love, Huber, & Anders, 2014; Robinson, McCarthy, & Smyth, 2010; Tarazona et al., 2015) are the several R packages that provide the list of expressed genes based on the FDR and log<sub>2</sub>fc. FDR is also known as false discovery rate (adjusted *P*-value) for multiple tests. Log<sub>2</sub>fc is the criteria for selecting significant genes.

#### 6.4.2.1.7 Gene of interest

Finally, the homology search, functional characterization, and pathways of DEGs can be done by using Blast tool (Altschul, Gish, Miller, Myers, & Lipman, 1990) and gene ontology tools, such as DAVID ontology (<https://david.ncifcrf.gov/>), KEGG (<https://www.genome.jp/kegg/>), WebGestalt (<http://www.webgestalt.org/>), GeneCodis 4.0 (<https://genecodis.genyo.es/>), KOBAS (<http://kobas.cbi.pku.edu.cn/kobas3/?t=1>), AmiGO 2 (<http://amigo.geneontology.org/amigo>), and PANTHER (<http://www.pantherdb.org/>). Gene ontology and functional characterization of the gene of interest can be done using the Blast2GO tool, which is further categorized into three subcategories, such as biological processes, molecular function, and cellular component, whereas coexpression analysis of DEGs can be performed using Cytoscape (Shannon et al., 2003) and ClueGO (Bindea et al., 2009) tools.

## 6.5 Biological databases for data submission

Storage of microarray and RNA-seq data is also an important task that has been overcome by National Center for Biotechnology Information (NCBI) (<https://www.ncbi.nlm.nih.gov/>), DNA Data Bank of Japan (DDBJ) (<https://www.ddbj.nig.ac.jp/index-e.html>), and EMBL (<https://www.embl.org/>). All these databases share the data regularly through the International Nucleotide Sequence Database Collaboration (<http://www.insdc.org/>).

Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) database is used for a repository of microarrays data and high-throughput gene expression data. Transcriptome Shotgun Assembly database is used for the transcriptome assembly (<https://www.ncbi.nlm.nih.gov/genank/tsa/>). Sequence Read Archive (<https://www.ncbi.nlm.nih.gov/sra>), DDBJ Sequence Read Archive (<https://www.ddbj.nig.ac.jp/dra/index-e.html>), and European Nucleotide Archive (<https://www.ebi.ac.uk/ena/browser/home>) are the database portals for NGS reads.

## 6.6 Applications of microarray

1. It is a fast, specific, and sensitive technique that enables assessment of whole genome in an one run (Oostlander, Meijer, & Ylstra, 2004), a genomic region in different cells or closely related organisms and aberrations of DNA copy number that results in human genetic disorders (Pollack et al., 1999).
2. Microarray technique can be used to assess how particular conditions of an organism can affect the expression pattern of the gene in various cells or tissues by studying the gene expression profile to identify the gene of interest (Adomas et al., 2008).
3. Microarray technique has huge application in pharmaceutical and healthcare industries that includes the field of drug discovery, such as conducting basic research, pharmacology, target discovery, target selectivity, biomarker determination, toxicogenomics, and developing prognostic tests. Information from comparative genomic hybridization, microRNAs, ChIP-on-chip, and gene splice variants along with gene expression analysis can be applied to the process of drug discovery.

## 6.7 Applications of RNA-sequencing

1. RNA-seq approach is used to sequence the RNA using sequencing technologies to know about its quantitative and qualitative aspects. It involves analyzing the expression pattern of genes of the transcriptome (total cellular content of RNA, tRNA, mRNA, and rRNA) in the cell (Ozsolak & Milos, 2011).
2. It can also be identifying gene-associated SNPs and disease variants that would help in determining allele-specific expression in the whole genome (Liu et al., 2020).
3. With the rapid decrease in the cost of sequencing techniques and high-throughput assessment of gene expression levels, RNA-seq has now become a common method for expression quantitative loci analysis in various diseases (Liu et al., 2020).
4. It can involve determining small RNA profiling, single-cell RNA, differential gene expression, variants' detection, splicing patterns, allele-specific expression, characterization of alternative, and system biology.
5. It plays a major role in studying pathogen–host interactions in eukaryotic cells including the immune response (Sirbu, Kerr, Crane, & Ruskin, 2012).
6. The RNA-seq analysis can be used for the identification of novel gene identification, expression, splicing, and mutation analysis (Han, Gao, Muegge, Zhang, & Zhou, 2015).

There are several studies based on RNA-seq that have been performed to identify the candidate genes. Jaiswal et al. (2018) identified several candidate genes, such as ATP-citrate synthase family, aspartate kinase, homoserine dehydrogenase, Dhn, and LEA genes, in leaf and root drought-responsive genes. In another study, 297 DEGs were obtained from high and low growth animals of muscle transcriptome analysis of *Bos frontalis* (Mukherjee et al., 2020). Jasrotia et al. (2017) identified a pathogenesis-related protein, salicylic acid-binding protein 2-like, MAPK, and NBS-LRR in the resistant and susceptible culture of *Vigna mungo*.

## 6.8 Advantages of transcriptome sequencing over microarray technology

1. RNA-seq can detect low abundance transcripts.
2. Capable of differentiating critical isoforms in the biological sample.
3. Capable of mining more significant differentially expressed genes.
4. It allows the detection of genetic variants and genic simple sequence repeats (SSR).
5. RNA-seq does not have technical issues that are commonly found in microarray cDNA–probe hybridization, such as cross-hybridization, nonspecific hybridization, detection of limited numbers of individual probes, problems associated with probe redundancy, and annotation (Zhao, Fung-Leung, Bittner, Ngo, & Liu, 2014)

## 6.9 Limitations and future perspective of RNA-sequencing

1. Sequencing, functional annotation, and data interpretation are very challenging when a reference genome is not available.
2. Quantitation of transcripts is affected by cDNA library preparation and alignment.
3. Different sequencing platforms and reads depth are tough for hybrid analysis.
4. Although the cost of RNA-seq decreases day by day, still it is not affordable as it required high computation resources.
5. Recent advancements in RNA-seq help in the quantification and identification of the transcriptome. It can be helpful in building a complete catalog of genes or transcripts obtained from genomes from unicellular to complex mammals under various stages, such as healthy or infected samples (Ozsolak & Milos, 2011). These technologies provide genomic information and solve the basic biological problems in crop improvement, personalization of health care, and evolutionary studies.

## 6.10 Conclusion

High-throughput screening of genes and recent advancements in the field of molecular genetics has revealed gene-based mechanisms in several science fields. Microarray and RNA-seq techniques find its applications in many areas of the science field, such as expression analysis, sequencing by hybridization, alternative splicing, genic SSR, detecting a mutation in genomic regions, assessing DNA copy number, resequencing, drug discovery, coexpression analysis, comparative genomics, and immunoassay (using protein microarrays). Scientific studies related to the progression of

diseases, such as cancer and pathogenesis, have been revolutionized with the increased use of new molecular biology techniques. Studying new diagnostic and prognostic markers of clinical specimens in large number is an important step in discovering the new findings in basic science and medical science.

## Conflict of interest

The authors declare that there is no conflict of interest.

## References

- Adomas, A., Heller, G., Olson, Å., Osborne, J., Karlsson, M., & Nahalkova, J. (2008). Comparative analysis of transcript abundance in *Pinus sylvestris* after challenge with a saprotrophic, pathogenic or mutualistic fungus. *Tree Physiology*, 28(6), 885–897.
- Agarwal, A., Koppstein, D., Rozowsky, J., Sboner, A., Habegger, L., Hillier, L. W., . . . Gerstein, M. (2010). Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics*, 11(1), 383.
- Allanach, K., Mengel, M., Einecke, G., Sis, B., Hidalgo, L. G., Mueller, T., & Halloran, P. F. (2008). Comparing microarray vs RT-PCR assessment of renal allograft biopsies: Similar performance despite different dynamic ranges. *American Journal of Transplantation*, 8(5), 1006–1015.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403–410.
- Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 31(2), 166–169.
- Auffray, C., Nageotte, R., Chambraud, B., & Rougeon, F. (1980). Mouse immunoglobulin genes: A bacterial plasmid containing the entire coding sequence for a pre- $\gamma$  2a heavy chain. *Nucleic Acids Research*, 8(6), 1231–1242.
- Auffray, C., & Rougeon, F. (1980). Nucleotide sequence of a cloned cDNA corresponding to secreted  $\mu$  chain of mouse immunoglobulin. *Gene*, 12(1–2), 77–86.
- Barnhart, B. J. (1989). The Department of Energy (DOE) human genome initiative. *Genomics*, 5(3), 657–660.
- Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood-Education and Practice*, 98(6), 236–238.
- Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., . . . Boutell, J. M. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), 53–59.
- Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., . . . Galon, J. (2009). ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics (Oxford, England)*, 25(8), 1091–1093.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15), 2114–2120.
- Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., . . . Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience*, 2, 10. Available from <https://doi.org/10.1186/2047-217X-2-10>.
- Braslavsky, I., Hebert, B., Kartalov, E., & Quake, S. R. (2003). Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7), 3960–3964.
- Bumgarner, R. (2013). Overview of DNA microarrays: Types, applications and their future. *Current Protocols in Molecular Biology*, Chapter 22, Unit 22.1.
- Bushmanova, E., Antipov, D., Lapidus, A., & Prjibelski, A. D. (2019). rnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, 8(9), pgiz100.
- Chidgeavdze, Z. G., Beabealashvili, R. S., Atrazhev, A. M., Kukhanova, M. K., Azhayeve, A. V., & Krayevsky, A. A. (1984). 2',3'-Dideoxy-3' amino-nucleoside 5'-triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases. *Nucleic Acids Research*, 12(3), 1671.
- Ching, T., Huang, S., & Garmire, L. X. (2014). Power analysis and sample size estimation for RNA-Seq differential expression. *RNA*, 20(11), 1684–1696.
- Chu, Y., & Corey, D. R. (2012). RNA sequencing: Platform selection, experimental design, and data interpretation. *Nucleic Acid Therapeutics*, 22(4), 271–274.
- Colantuoni, C., Henry, G., Zeger, S., & Pevsner, J. (2002). SNOMAD (Standardization and Normalization of MicroArray Data): Web-accessible gene expression data analysis. *Bioinformatics (Oxford, England)*, 18(11), 1540–1541.
- Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One*, 12(12), e0190152.
- Crampton, J., Humphries, S., Woods, D., & Williamson, R. (1980). The isolation of cloned cDNA sequin which are differentially expressed in human lymphocytes and fibroblasts. *Nucleic Acids Research*, 8(24), 6007–6017.
- Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C., & Conklin, B. R. (2002). GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genetics*, 31(1), 19–20.
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1), 15–21.
- Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25), 14863–14868.

- Franklin, W. A., Aisner, D. L., Davies, K. D., Crooks, K., Post, M. D., Kleinschmidt-DeMasters, B. K., . . . Varella-Garcia, M. (2020). Pathology, biomarkers, and molecular diagnostics. *Abeloff's clinical oncology*, 225–253.
- Gedil, M., Ferguson, M., Girma, G., Gisel, A., Stalvolone, L., & Rabbi, I. (2016). Perspectives on the application of next-generation sequencing to the improvement of Africa's staple food crops. *Next generation sequencing—Advances, applications and challenges*, 287–321.
- Gergen, J. P., Stern, R. H., & Wensink, P. C. (1979). Filter replicas and permanent collections of recombinant DNA plasmids. *Nucleic Acids Research*, 7(8), 2115–2136.
- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews. Genetics*, 17(6), 333.
- Govindarajan, R., Duraiyan, J., Kaliyappan, K., & Palanisamy, M. (2012). Microarray and its applications. *Journal of Pharmacy and Bioallied Sciences*, 4(Suppl 2), S310.
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., . . . Chen, Z. (2011). Trinity: Reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, 29(7), 644.
- Grunstein, M., & Hogness, D. S. (1975). Colony hybridization: A method for the isolation of cloned DNAs that contain a specific gene. *Proceedings of the National Academy of Sciences of the United States of America*, 72(10), 3961–3965.
- Han, Y., Gao, S., Muegge, K., Zhang, W., & Zhou, B. (2015). Advanced applications of RNA sequencing and challenges. *Bioinformatics and Biology Insights*, 9, BBI-S28991.
- Hansen, K., Irizarry, R., & Wu, Z. (2011). Removing technical variability in RNA-Seq data using conditional quantile normalization. *Tech Rep 227* Department of Biostatistics, Johns Hopkins University.
- Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., . . . Giladi, E. (2008). Single-molecule DNA sequencing of a viral genome. *Science (New York, N.Y.)*, 320(5872), 106–109.
- Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing. *DNA Genomics*, 107(1), 1–8.
- Hoheisel, J. D. (2006). Microarray technology: Beyond transcript profiling and genotype analysis. *Nature Reviews. Genetics*, 7(3), 200–210.
- Hu, Y., Wang, K., He, X., Chiang, D. Y., Prins, J. F., & Liu, J. (2010). A probabilistic framework for aligning paired-end RNA-seq data. *Bioinformatics (Oxford, England)*, 26(16), 1950–1957.
- Humphries, P., Cochet, M., AndréeKrust, M., Gerlinger, P., Kourilsky, P., & Chambon, P. (1977). Molecular cloning of extensive sequences of the in vitro synthesized chicken ovalbumin structural gene. *Nucleic Acids Research*, 4(7), 2389–2406.
- Jaiswal, S., Antala, T. J., Mandavia, M. K., Chopra, M., Jasrotia, R. S., Tomar, R. S., . . . Rai, A. (2018). Transcriptomic signature of drought response in pearl millet (*Pennisetum glaucum* (L.) and development of web-genomic resources. *Scientific Reports*, 8(1), 1–16.
- Jarvie, T., & Harkins, T. (2008). Transcriptome sequencing with the Genome Sequencer FLX system. *Nature Methods*, 5(9), vi–viii.
- Jasrotia, R. S., Iquebal, M. A., Yadav, P. K., Kumar, N., Jaiswal, S., Angadi, U. B., . . . Kumar, D. (2017). Development of transcriptome-based web genomic resources of yellow mosaic disease in *Vigna mungo*. *Physiology and Molecular Biology of Plants*, 23(4), 767–777.
- Jawhar, N. M. (2009). Tissue microarray: A rapidly evolving diagnostic and research tool. *Annals of Saudi Medicine*, 29(2), 123–127.
- Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, 12(4), 357.
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4), R36.
- Kratz, A., & Carninci, P. (2014). The devil in the details of RNA-seq. *Nature Biotechnology*, 32(9), 882.
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357.
- Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., & Kendziorski, C. (2013). EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics (Oxford, England)*, 29(8), 1035–1043.
- Lennon, G. G., & Lehrach, H. (1991). Hybridization analyses of arrayed cDNA libraries. *Trends in Genetics*, 7(10), 314–317.
- Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1), 323.
- Li, C., & Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biology*, 2(8), 1–11.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), 1754–1760.
- Liao, Y., Smyth, G. K., & Shi, W. (2013). featureCounts: An efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, 30(7), 923–930.
- Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., . . . Law, M. (2012). Comparison of next-generation sequencing systems. *BioMed Research International*, 2012.
- Liu, Y., Liu, X., Zheng, Z., Ma, T., Liu, Y., Long, H., . . . Zhao, S. (2020). Genome-wide analysis of expression QTL (eQTL) and allele-specific expression (ASE) in pig muscle identifies candidate genes for meat quality traits. *Genetics Selection Evolution*, 52(1), 1–11.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.
- Mah, N., Thelin, A., Lu, T., Nikolaus, S., Kühbacher, T., Gurbuz, Y., . . . Costello, C. M. (2004). A comparison of oligonucleotide and cDNA-based microarray systems. *Physiological Genomics*, 16(3), 361–370.
- Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, 470(7333), 198–203.
- Martin, M. (2011). Cutadaptremoves adapter sequences from high-throughput sequencing reads. *EMBNet.journal*, 17(1), 10–12.

- Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2), 560–564.
- Mello-Coelho, V. D., & Hess, K. L. (2005). A conceptual and practical overview of cDNA microarray technology: Implications for basic and clinical sciences. *Brazilian Journal of Medical and Biological Research*, 38(10), 1543–1552.
- Miller, J. K., & Barnes, W. M. (1986). Colony probing as an alternative to standard sequencing as a means of direct analysis of chromosomal DNA to determine the spectrum of single-base changes in regions of known sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 83(4), 1026–1030.
- Miller, M. B., & Tang, Y. W. (2009). Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical Microbiology Reviews*, 22(4), 611–633.
- Miranda, F., Batista, C., Silva, A., Morais, J., Neto, N., & Ramos, R. (2018). Improving metagenomic assemblies through data partitioning: a GC content approach. In *International Conference on Bioinformatics and Biomedical Engineering* (pp. 415–425). Springer, Cham.
- Mukherjee, S., Mukherjee, A., Jasrotia, R. S., Jaiswal, S., Iquebal, M. A., Longkumer, I., ... Rai, A. (2020). Muscle transcriptome signature and gene regulatory network analysis in two divergent lines of a hilly bovine species Mithun (*Bos frontalis*). *Genomics*, 112(1), 252–262.
- Oostlander, A. E., Meijer, G. A., & Ylstra, B. (2004). Microarray-based comparative genomic hybridization and its applications in human genetics. *Clinical Genetics*, 66(6), 488–495.
- Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities. *Nature Reviews. Genetics*, 12(2), 87–98.
- Pelizzola, M., Pavelka, N., Foti, M., & Ricciardi-Castagnoli, P. (2006). AMDA: An R package for the automated microarray data analysis. *BMC Bioinformatics*, 7(1), 1–9.
- Pettersson, E., Lundeberg, J., & Ahmadian, A. (2009). Generations of sequencing technologies. *Genomics*, 93(2), 105–111.
- Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., ... Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, 23(1), 41–46.
- Ramoni, M. F., Sebastiani, P., & Kohane, I. S. (2002). Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 99(14), 9121–9126.
- Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., ... Griffith, M. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, 7(11), 909–912.
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1), 139–140.
- Saeed, A., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., ... Sturn, A. (2003). TM4: A free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2), 374–378.
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), 5463–5467.
- Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)*, 27(6), 863–864.
- Servant, N., Gravier, E., Gestraud, P., Laurent, C., Paccard, C., Biton, A., ... Hupé, P. (2010). EMA-AR package for Easy Microarray data analysis. *BMC Research Notes*, 3(1), 1–4.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504.
- Sharma, T. R., Devanna, B. N., Kiran, K., Singh, P. K., Arora, K., Jain, P., ... Singh, J. (2017). Status and Prospects of Next Generation Sequencing Technologies in Crop Plants. *Current Issues in Molecular Biology*, 27, 1–36.
- Smyth, G. K. (2005). *Limma: Linear models for microarray data. Bioinformatics and computational biology solutions using R and bioconductor* (pp. 397–420). New York: Springer.
- Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., & Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics (Oxford, England)*, 21(5), 631–643.
- Sturn, A., Quackenbush, J., & Trajanoski, Z. (2002). Genesis: Cluster analysis of microarray data. *Bioinformatics (Oxford, England)*, 18(1), 207–208.
- Sîrbu, A., Kerr, G., Crane, M., & Ruskin, H. J. (2012). RNA-Seq vs dual-and single-channel microarray data: Sensitivity analysis for differential expression and clustering. *PLoS One*, 7(12), e50986.
- Tarazona, S., Furió-Tarí, P., Turra, D., Pietro, A. D., Nueda, M. J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, 43(21), e140.
- Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9), 5116–5121.
- Voelckel, C., Gruenheit, N., & Lockhart, P. (2017). Evolutionary transcriptomics and proteomics: Insight into plant adaptation. *Trends in Plant Science*, 22(6), 462–471.
- Williams, A. G., Thomas, S., Wyman, S. K., & Holloway, A. K. (2014). RNA-seq data: Challenges in and recommendations for experimental design and analysis. *Current Protocols in Human Genetics*, 83(1), 11–13.
- Wu, J., Irizarry, R., & Gentry, J. (2020). *gcRma: Background adjustment using sequence information R package version 2.62.0*.
- Yang, I. S., & Kim, S. (2015). Analysis of whole transcriptome sequencing data: Workflow and software. *Genomics & Informatics*, 13(4), 119.
- Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One*, 9(1), e78644.