Article ID 240202001

RNAseq: A Widely Used Technique for Genome-Wide Expression Analysis

Nayna Tirkey, Sudhir Kumar, Avinash pandey and Kishor U Tribhuvan*
ICAR- Indian Institute of Agricultural Biotechnology, Ranchi – 834003, (Jharkhand), India
*Corresponding Author: kishor.tribhuvan@icar.gov.in

Transcriptomics is the high-throughput study of the transcript profile of the genome produced under specific circumstances or in a particular cell. Two contemporary techniques, microarrays and RNA-seq, are widely used for transcriptome studies. Microarrays techniques were developed in the 1990s, where thousands of DNA fragments (probs) were immobilized on a solid support and hybridized with complementary target sequences in the organism of interest. It is a powerful tool for quantifying the expression level of many genes within particular mRNA samples. It is widely used to study disease diagnosis, drug discovery and development, toxicological research, immunological studies, microbial detection and identification, comparative genomics, determination of virulence factors affected by various conditions, etc. However, it has many applications in research and diagnosis and has certain limitations. Microarray only identifies sequences whose probes are present on the array chip. Therefore, availability of genome sequence information to prepare the probes is a prerequisite for microarray study. Apart from that, quantifying the expression of low-abundance genes remains challenging in microarrays. These limitations led to inventions of new simple and cost-effective techniques for transcriptome studies.

The advancement of sequencing technologies allows high-efficiency genome-wide RNA sequencing (RNAseq). With its accuracy, high throughputs, and cost-effectiveness, RNA-seq has become a common tool for studying transcriptomes to understand the expression behavior of biological samples under different circumstances. The data analysis is critical in the RNAseq experiment to extract meaningful output from datasets generated during the RNAseq experiment. Along with NGS sequencing technologies, advanced bioinformatics tools and algorithms were also developed to handle massive datasets and extract meaningful results. RNAseq data are analyzed using various tools and pipelines. The selection of tools and pipelines depends upon the availability of reference genome sequences and the purpose of the RNAseq experiment. Routine transcriptome workflow may consist of the following steps, as shown in figure 1: (1) isolation of total RNAs from samples, (2) preparation of sequencing library, (3) RNA-seq sequencing using NGS, (4) preprocessing of raw data, (5) read alignment/transcriptome reconstruction (if reference genome not available), (6) expression quantification, (7) differential expression analysis, (8) annotation of genes. The detailed workflow of the RNAseq experiment is depicted in Figure 1.

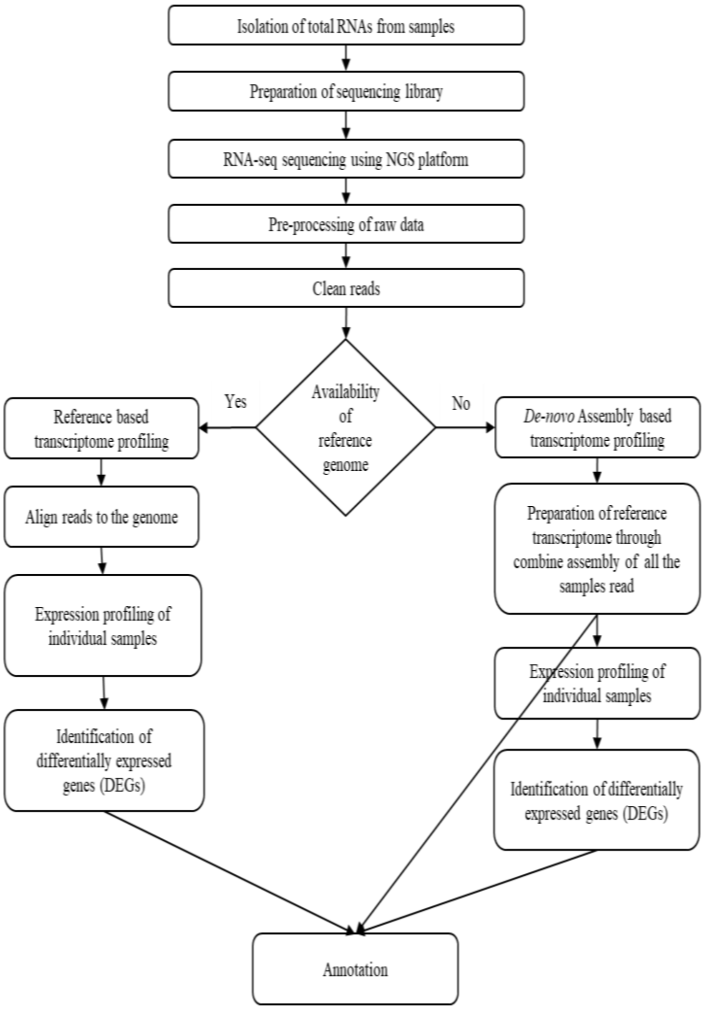
Isolation of total RNAs from samples
Preparation of sequencing library
RNA-seq sequencing using NGS platform
Pre-processing of raw data
Clean reads
Availability of reference genome
Yes
No
Reference based transcriptome profiling
Align reads to the genome
Expression profiling of individual samples
Identification of differentially expressed genes (DEGs)
De-novo Assembly based transcriptome profiling
Preparation of reference transcriptome through combine assembly of all the samples read
Expression profiling of individual samples
Identification of differentially expressed genes (DEGs)
Annotation

Fig. 1: An overview of the transcriptome assembly pipeline

Experimental setup, tissue harvesting and analysis strategies

Experimental set up and sample collection are the most critical stages of any transcriptome study. A very small error in these stages can affect the quality of sequence data and its further inference. The selection of appropriate tissue/stage for RNA isolation is equally important to capture the transcriptome profile of desired traits The use of three biological replicates is always advisable to minimize the variability present among the biological samples. One can refer the publication by Roble et al., (2012) for details of experimental designs and analysis strategies depending on the purpose of the experiment.

RNA Sequencing (RNAseq)

Micrograms of high-quality RNA are required for most library construction kits. It is always preferable to use fresh tissues for RNA isolation. RNA samples with RNA integrity number (RIN) value of >7 are of good quality and suitable for processing. Sequencing library preparation kits varies platform to platform and are commercially available in the market. Commercially available kits for library preparation are compatible with one or more platforms, and one can choose the kit based on compatibility with the targeted NGS platform. Several sequencing platforms, like Illumina, Helicos, PacBio, Ion Torrent, Nanopore, etc., have diverse data formats, throughputs, and qualities. Illumina is the most popular platform for RNA-seq experiments due to its high throughput and low error rates. Single molecule real time sequencing (SMRT) based PacBio has recently gained attention due to its increased read length capable of capturing full-length transcripts. However, it has a high error rate of ~5% and therefore, a hybrid approach combining Illumina and PacBio is widely adapted these days. Sequence depth is a crucial factor for identifying rare splicing events and low-abundance transcripts. Therefore, a deeper sequencing is required to detect low abundance transcripts and rare splicing events.

Data Analysis

The raw reads obtained from the sequencer can be subjected to data analysis using different bioinformatic analysis software (Table 1). The quality of raw reads can be assessed using various tools like fastQC, multiQC, NGS QC Toolkit, etc. Adaptor contamination and low-quality reads are usually removed using cleaning software like Trimmomatic, cutadapt, FASTX-Toolkit, etc. The preprocessed cleaned reads are then mapped, or *de-novo* assembled based on a reference genome sequence availability. The availability of a good quality reference genome provides ease in data analysis. Reference-guided assembly uses gene information of the reference genome for expression quantification. If the studied species lacks a reference genome sequence, the reads can be *de novo* assembled, and assembled contigs can be used as a reference for expression quantification. Both de novo and reference-based assembly provide transcript sequences expressed at particular conditions.

Uses of transcriptome data

Generation of genomic resources

The availability of genome sequence information is a crucial resource to carry out molecular dissection of various phenomena in the species. With the advancement in sequencing technologies, many organisms' genome sequences are available in public databases. Even though most crops and animal species' genome information is yet to be available. Transcriptome sequencing is a commonly used technique to generate genomic resources where species lack genomic information because it deals with expressed parts of the genome, carried out quickly with low input cost. Thus, transcriptome data is a valuable resource in crops where genome sequence information is not available.

Development of molecular markers

Transcriptome sequencing is widely used for the development of molecular markers like simple sequence repeats (SSRs), Single Nucleotide Polymorphism (SNPs), Inter Simple Sequence Repeats (ISSR), etc. These molecular markers can be used for fingerprinting, diversity analysis, and gene mapping. Identification of SSRs from transcriptome data can be carried out using the Microsatellite identification tool (MISA) (https://academic.oup.com/bioinformatics/article/33/16/2583/3111841) and Krait - Microsatellites

Investigation and Primer Design (http://krait.biosv.com/en/latest/) tools. The added advantage of using markers designed from the transcriptome data is they are directly linked to the genes.

Identification of Differentially expressed genes

The main objective of most RNA-seq experiments is to measure and compare differential gene expression under various conditions to infer biological function. Intra-sample abundance comparisons were commonly performed with metrics of Reads Per Kilobase per Million (RPKM mapped reads) or Fragments Per Kilobase per Million (FPKM mapped reads). Transcript level quantification can be analyzed using RSEM or Cufflinks, and differential gene expression (DEGs) can be calculated using edgeR (https://bioconductor.org/packages/release/bioc/html/edgeR.html) and DESeq2 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html) tools.

Identification of non-coding RNAs

The non-coding RNAs (lncRNA, circular RNA, miRNA, siRNAs) play an essential regulatory role in developmental stages and various biotic and abiotic stress responses. Transcriptome data is an essential resource for the identification of these non-coding RNAs. Long non-coding RNAs can be identified using CPAT (https://cpat.readthedocs.io/en/latest/), PLEK (https://sourceforge.net/projects/plek/files/), CNCI (https://github.com/www-bioinfo-org/CNCI), and CPC2 (http://cpc2.gao-lab.org/) tools. Similarly, circular non-coding RNAs were carried out using CIRI (https://sourceforge.net/projects/ciri/) tools.

Gene prediction and annotation of the sequenced genome

Pac-Bio-based transcriptome sequencing generates full-length transcript sequences. The full-length transcripts obtained from various developmental tissues were used as a reference for gene prediction and annotation of the newly sequenced genome.

The details of the software/ tools used in the RNAseq experiment are provided in Table 1.

Table 1. Software/ tools for RNAseq analysis

Category		**Tools**
Raw reads quality control		fastQC, multiQC, NGS QC Toolkit
Reads preprocessing		Cutadapt, Trimmomatic, FASTX-Toolkit
Assembly	Genome guided	Cufflinks, Scripture, StringTie
	De novo	Trinity, Trans-ABySS, Oases
Assembly quality assessment		BUSCO
Transcript level quantification		RSEM, Cufflinks
Differential gene expression (DEGs)		edgR, DESeq2, CuffDiff2
Annotation		Blast$_2$Go
Identification of SSRs		MISA, Krait
Prediction of lncRNAs		CPAT, PLEK, CPC2, CNCI
Prediction of circular RNAs		CIRI

References

Stark, R., Grzelak, M. & Hadfield, J. (2019) RNA sequencing: the teenage years. Nat Rev Genet 20, 631–656. https://doi.org/10.1038/s41576-019-0150-2

Bumgarner, R. (2013) Overview of DNA microarrays: types, applications, and their future. Curr Protoc Mol Biol.

Conesa, A., Madrigal, P., Tarazona, S. *et al.* (2016) A survey of best practices for RNA-seq data analysis. Genome Biol 17, 13.

Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., & Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC genomics, 13(1), 1-14.

* * * * * * * *