Chapter 4

# Transcriptomic Profiling Using Next Generation Sequencing - Advances, Advantages, and Challenges

Krishanpal Anamika, Srikant Verma, Abhay Jere and Aarti Desai

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/61789

**Abstract**

Transcriptome, the functional element of the genome, is comprised of different kinds of RNA molecules such as mRNA, miRNA, ncRNA, rRNA, and tRNA to name a few. Each of these RNA molecules plays a vital role in the physiological response, and understanding the regulation of these molecules is extremely critical for the better understanding of the functional genome. RNA Sequencing (RNASeq) is one of the latest techniques applied to study genome-wide transcriptome characterization and profiling using high-throughput sequenced data. As compared to array-based methods, RNASeq provides in-depth and more precise information on transcriptome characterization and quantification. Based upon availability of reference genome, transcriptome assembly can be reference-guided or *de novo*. Once transcripts are assembled, downstream analysis such as expression profiling, gene ontology, and pathway enrichment analyses can give more insight into gene regulation. This chapter describes the significance of RNASeq study over array-based traditional methods, approach to analyze RNASeq data, available methods and tools, challenges associated with the data analysis, application areas, some of the recent advancement made in the area of transcriptome study and its application.

**Keywords:** RNASeq, *de novo* and reference-based transcriptome assembly, Differential gene expression, Next Generation Sequencing

## 1. Introduction

Completion of the Human Genome Project in 2001 brought with it the realization that while understanding the genome is of great value, our understanding of biology is woefully incomplete without the knowledge of the functional elements of the genome. The functional element of the genome is the transcriptome, which is the set of RNA molecules such as mRNA, rRNA, tRNA, and various small RNAs. A large number of research projects are now focused

on the transcriptome rather than on genome and proteome as only 1-2% of genes are coding and 80-90% of the transcribed genes are not translated to proteins. However, these are known to be involved in epigenetic regulation and gene expression regulation [1-4]. Gene expression is a complex process regulated at multiple levels such as gene transcription, post-transcriptional modifications, and translation. Briefly, complexity at the transcription regulation arises from the presence of multiple Transcription Start Sites (TSSs), which can result in production of multiple transcripts from a single gene [5] and alternate splicing as well as alternate polyadenylation of the primary RNA to produce several different forms of transcripts originating from the same gene [6, 7]. Because of different TSSs, eventually each mature transcript will code for different protein [8]. Additionally, noncoding RNAs, which are not translated to proteins, play catalytic and structurally important roles. For example, tRNAs and rRNAs play a critical role in translation, small nuclear RNAs (snRNAs) participate in mRNA splicing, small nucleolar RNAs (snoRNAs) regulate rRNA splicing, guide RNAs (gRNAs) regulate RNA editing, and miRNA are involved in translational repression [9]. Study of the transcriptome provides an understanding of the regulation of gene expression pattern [10], alternative splicing and transcript structure [11], dynamic regulation of transcripts in different tissues [12], and detailed information about the gene regulation in normal and diseased conditions [13].

Transcriptome profiling is typically performed using hybridization or sequencing-based methodologies. Hybridization-based methods involve binding of fluorescently labeled fragments to complementary probe sequences either in solution or on a solid surface, e.g., microarray [14, 15]. These approaches, however, suffer from limitations such as low resolution, low specificity, and low sensitivity [16]. Later, Sanger sequencing-based approaches such as SAGE (Serial Analysis of Gene Expression) [17], CAGE (Cap Analysis of Gene Expression) [18], and MPSS (Massively Parallel Signature Sequencing) [19] were developed, but these approaches have serious limitations such as consideration of partial transcripts structure for gene expression and inability to distinguish between isoforms [20]. With the advent of Next Generation Sequencing (NGS), a technology that enables sequencing of millions of nucleotide fragments in parallel, RNA Sequencing (RNASeq) has emerged as a powerful method for studying the transcriptome. Though microarrays are high-throughput and economical, RNASeq offers numerous advantages over microarrays [15]. Some of the key benefits of using RNASeq over microarrays are:

**a.** Genome-wide coverage of transcripts is offered by RNASeq.

**b.** No prior knowledge of genome sequence is required in the case of RNASeq as opposed to microarray and hence RNASeq experiment can be performed in the absence of the reference genome.

**c.** Improved sensitivity and specificity: RNASeq offers enhanced detection of transcripts and differentially expressed genes and isoforms. Moreover, RNASeq is known to be more accurate in terms of fold change detection for both high- and low-abundance genes.

**d.** Detection of novel transcripts: Unlike microarray, RNASeq enables genome-wide unbiased study and is not dependent on transcript or region-specific probes and hence it investigates both known and novel transcripts.

http://dx.doi.org/10.5772/61789

**e.** Detection of low-abundance transcripts if sequencing is done at high depth.

**f.** No or minimal background signal: While mapping the reads to the genome, one can consider reads mapping unambiguously, which results in noise reduction. On the other hand, cross-hybridization increases noise-to-signal ratio in microarrays.

**g.** SNP detection: RNASeq data can be used for SNP detection especially for highly and medium expressed genes.

Because of its wider detection range, more sensitivity, genome-wide capture of expression profile, and rapidly decreasing cost, RNASeq technology is being preferred over array-based methods for transcriptome profiling. RNASeq has been widely used in the detection of differentially expressed genes between cancerous and normal tissue samples [21], identification of novel gene fusion events in melanoma [22], discovery of several novel miRNAs in cancerous cells [23], identification of differential gene expression and splicing events in Alzheimer's disease [24], identification of differential promoter usage, and higher expression of noncoding RNA in diabetes [25, 26]. RNASeq is now being used extensively for transcriptome assembly, thus enabling better characterization of economically important plants such as Garlic [27], Pea [28], Chickpea [29], Rice [30], Olive [31], Wheat [32], and many other plants [33]. Further, combination of molecular biology and biochemical techniques with sequencing has led to the study of different aspects of the transcriptome, such as mRNASeq, miRNASeq, GROSeq, CLIPSeq, NETSeq, PARESeq, and ChIRPSeq (additional information in Table 1). Projects such as ENCODE (ENCyclopedia of the DNA Elements) and TCGA (The Cancer Genome Atlas) have characterized transcriptome of several different human cell lines and tumor samples, respectively, using NGS-based transcriptome profiling. Goal of ENCODE (https://www.**encode**project.org/) is to identify genome-wide transcriptome profile to understand the downstream effects of gene regulation in the human genome. TCGA (www.cancergenome.nih.gov/), which contains information on cancer patient data, aims to understand the mechanism of tumor transformation and progression.

| RNASeq methods | Description | Reference |
|---|---|---|
| mRNASeq | To identify messenger RNAs (mRNAs) | [12] |
| miRNASeq | To identify micro RNAs (miRNAs) | [167] |
| GROSeq (Global Run On Sequencing), PROSeq | To identify nascent RNAs that are actively transcribed by RNA Pol II | [168] |
| ChIRPSeq (Chromatin Isolation by RNA Purification) | To discover regions of the genome bound by a specific RNA | [169] |
| RiboSeq (Ribosome profile Sequencing) | To identify RNAs that are being processed by the ribosome and hence this method helps to monitor the translation process | [170] |
| CLIPSeq (Cross-Linking and Immunoprecipitation Sequencing) | To identify the binding sites of cellular RNA-binding proteins (RBPs) using UV light to cross-link RNA to RBPs without the incorporation of photoactivatable groups into RNA | [171] |

| RNASeq methods | Description | Reference |
|---|---|---|
| PAR-CLIP Seq(Photoactivatable-Ribonucleoside-Enhanced Cross-Linking and Immunoprecipitation Sequencing) | To identify and sequence the binding sites of cellular RNA-binding proteins (RBPs) and microRNA-containing ribonucleoprotein complexes (miRNPs) | [172] |
| NETSeq (Native Elongation Transcript Sequencing) | It sequences and captures nascent RNA transcripts after immunoprecipitation of RNA Pol II elongation complex | [173] |
| TRAPSeq (Targeted Purification of Polysomal mRNA Sequencing) | To detect and identify translating mRNAs | [174] |
| PARESeq (Parallel Analysis of RNA Ends Sequencing) and GMUCT (Genome-wide Mapping of Uncapped Transcripts) | To detect and identify miRNA cleavage sites and uncapped transcripts that undergo degradation | [175] |
| TIFSeq (Transcript Isoform Sequencing) or Paired-End Analysis of Transcription start site (PEAT) | RNA isoforms are identified after 5' and 3' paired-end sequencing | [176] |
| CELSeq (Cell Expression by Linear amplification and Sequencing), SMARTSeq (Switching Mechanism At the 5′ end of the RNA Template Sequencing), STRT (Single-cell Tagged Reverse Transcription) | Single-cell transcriptomics methods | [177] |

**Table 1.** Various RNASeq-based methods to study transcriptome

One of the first steps while designing the RNASeq experiment is choosing an appropriate sequencing platform. Several sequencing platforms such as Illumina, Roche, PacBio, and Ion Torrent, which are based on different sequencing chemistry and technology, are available [reviewed in 34, 35]. Current leading platform for RNASeq (and other NGS-based analyses) is the HiSeq series of sequencers from Illumina (https://www.illumina.com/systems.html) because it provides high throughput, deep sequencing, low sequence error, and long enough read data to be useful in multiple applications. Recently, the PacBio RS II (http://www.pacificbiosciences.com/) is gaining popularity for better transcriptome construction, because of its ability to generate long reads. Once the millions of reads are generated from an RNASeq experiment, the bioinformatics data analysis begins. In the following section, we briefly present the bioinformatics data analysis steps, tools, and methods.

http://dx.doi.org/10.5772/61789

## 2. Bioinformatics analysis of RNASeq data

Analysis of the RNASeq data is a multistep process that typically includes quality check, data preprocessing, transcriptome assembly (reference-guided and *de novo* transcriptome assembly), quantification, statistical analysis, and functional annotation (Figure 1). These steps are described in details in the following.

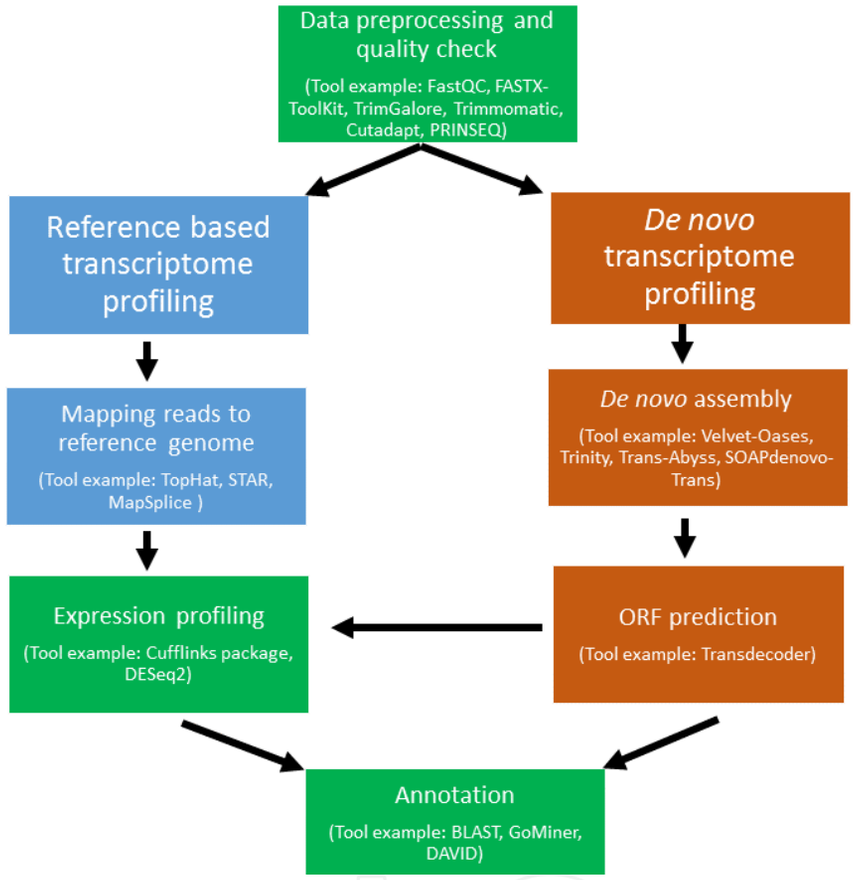


**Figure 1.** Basic RNASeq data analysis workflow. Firstly, raw sequenced data are checked for the quality and, if required, low-quality reads and artifacts are removed. In the case of reference-based assembly, the reads are mapped to the reference genome in order to know their location. All the mapped reads are then analyzed for expression profiling. Further, differentially expressed genes and isoforms can be annotated using Gene Ontology (GO) and Pathway enrichment analyses. In *de novo* assembly approach, after preprocessing of the raw reads, transcriptome can be assembled using different *de novo* transcriptome assemblers. Once transcripts are constructed and abundance estimate is obtained, the complete Open Reading Frames (ORFs) transcripts are predicted. The predicted ORFs can be annotated or analyzed for expression profiling and then annotated using remote homology search method, GO, and pathway enrichment analyses.

### 2.1. Quality check and data preprocessing

Next generation sequencers assign a Phred quality score, which is the probability of the base call being inaccurate, to the called bases. Low Phred scores ($Q < 30$) indicate read data of poor quality. Poor-quality read data can arise from problems in the library preparation or from

sequencing itself. Additionally, PCR artifacts, sequence-specific biasness, untrimmed adapter sequences, and other possible contaminants can lead to poor data quality. These factors can affect the downstream analysis and data interpretation, and can give inaccurate results. In order to assess quality of raw sequenced data several tools such as FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and PRINSEQ [36] are available. Once the data are checked for quality, they should be processed to remove reads with low-quality bases, adapter sequences, and other contaminating sequences. Tools such as Cutadapt [37], Trimmomatic [38], TrimGalore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), which trim adapter or other contaminants based upon user-provided parameters, can be used for performing these operations. A brief description of some of these quality and data preprocessing tools is provided below:

**FastQC:** FastQC is a simple, easy-to-use tool that evaluates the quality of read data generated from the next generation sequencers. The input file/s for FastQC can be in Fastq, SAM, or BAM format either in the compressed or uncompressed form. FastQC reports basic statistics for the read data such as overrepresented sequences, k-mer content, base quality and content, adapter content, read duplication level, etc. FastQC is available as a stand-alone Java-based program with a graphical user interface and can be run from both Linux (using command line) and Windows systems.

**PRINSEQ:** PRINSEQ reports base quality, GC content, duplicates, adapters, presence of ambiguous sequences represented as "N," poly A tails, etc. Unlike FastQC, PRINSEQ also has the option of trimming and filtering reads. PRINSEQ is available as stand-alone as well as web application (http://prinseq.sourceforge.net/). It accepts uncompressed files in Fasta, Qual, and Fastq formats.

**Trimmomatic:** Trimmomatic is a Java-based program for the preprocessing of NGS read data (http://www.usadellab.org/cms/?page=trimmomatic). It can trim contaminant sequences, adapters, and filter reads based upon the quality. It supports compressed files in Fastq format and generates output in Fastq format. Because of its multithreading option, its data processing speed is higher than other tools available to perform the same function. Unlike some of the other tools, Trimmomatic can analyze both single-end as well as paired-end read data.

**Cutadapt:** Cutadapt is a python-based tool for read preprocessing and can be run as a command line application (https://cutadapt.readthedocs.org/en/stable). It accepts compressed files in Fasta, Qual, and Fastq formats, and supports both paired-end and single-end files. It trims low-quality bases, multiple adapter sequences from either 3', 5', or from both ends. In addition, Cutadapt can remove fixed number of bases from either ends of the sequences and supports demultiplexing, i.e., reads can be written to different output files depending upon the adapter sequence found in the reads. The demultiplexing feature is particularly useful since pooling multiple samples in a single run is an increasingly common practice as a result of increased sequencer throughput.

**TrimGalore:** TrimGalore is a wrapper tool written around FastQC and Cutadapt for quality check and adapter trimming for regular as well as MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. It accepts compressed Fastq files and supports paired-end and single-end data.

http://dx.doi.org/10.5772/61789

**FASTX-Toolkit:** FASTX-Toolkit is a collection of tools that accepts read data in Fasta and Fastq file formats and trim the data based on base quality and adapter sequence contamination. Additionally, the toolkit has tools that can perform file format conversion, split sequences based upon barcodes, and generate reverse complement of sequences.

Once the read data are filtered and trimmed to remove low-quality bases, adapter sequence, and contaminants, they are ready for transcriptome assembly and profiling analysis. There are two different approaches for constructing full-length transcripts: reference-based assembly (when a reference genome is available) and *de novo* assembly (when the reference genome is not available), a computationally intensive and complex process (Table 2). Reference-based or genome-guided assembly refers to mapping sequenced reads to the reference genome followed by assembling the transcripts. In contrast, in *de novo* transcriptome assembly, transcripts are constructed directly from the overlapping sequenced reads. For the transcriptome assembly of organisms without reference genome, only *de novo* transcriptome assembly approach is available for transcriptome construction. However, for organisms with known reference genome, both reference-based and *de novo* transcriptome assembly can be employed for transcriptome construction. In fact, in this case, *de novo* transcriptome assembly will be more effective in filling in the gaps (observed due to variation in reference genome sequence and poor-quality annotation) and hence would complement the reference-based transcriptome assembly. More details on these two transcriptome assembly approaches are discussed in the following sections.

| **Reference-based assembly** | ***de novo assembly*** |
|---|---|
| Reference genome is required to assemble the transcriptome | Transcriptome is assembled *de novo* |
| Relatively less computation- intensive | Computation- intensive |
| Contaminants and sequencing artifacts are not of major concern | Contaminants and sequencing artifacts can lead to poor quality of assembled transcriptome |
| Mapping quality of transcripts is dependent on splice aligners | Mapping is not required |
| Can assemble transcripts of low abundance | Difficult to assemble the transcripts of low abundance unless sequencing depth is high |
| Can work well with low sequencing depth data (~10X) | Work well with high sequencing depth data (~30X) |
| Less efficient in identifying novel isoforms and SNPs | Efficient in identifying novel isoforms and SNPs |
| Completeness and contiguity of transcriptome is relatively higher | Completeness and contiguity of transcriptome is relatively lower especially for low sequencing depth data |

**Table 2.** Difference between reference-based and *de novo* assembly approaches

## 2.2. Transcriptome assembly

### *2.2.1. Reference-based transcriptome assembly and profiling*

Typically, in a reference-based transcriptome profiling study, the computational workflow starts with aligning the quality-checked reads to the reference genome or transcriptome using a suitable read aligner. The aligned reads are then used to quantitate the genomic features (genes/isoforms). The quantity of the features needs to be normalized before comparison between different experimental conditions. The normalized feature counts are then used for drawing statistical inference on their difference in expression between samples under study. Finally, the differentially expressed set of genes is processed to derive biological insights relevant to the experimental setup. The success of this analysis depends very much on decisions that the user takes while choosing reference genome, annotation, tools, and associated parameter values at every step of the analysis. Steps involved in reference-based transcriptome assembly and analysis are described below.

#### *2.2.1.1. Choice of reference build and annotation file*

Reference genome and annotation files of a large number of species are available from a number of publicly available resources. Three of the most widely used resources are Ensembl (http://www.ensembl.org), the National Center of Biotechnology Information (NCBI; ftp://ftp.ncbi.nih.gov/genomes), and UCSC genome browser (http://genome.ucsc.edu). Ensembl is jointly headed by the European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI) and the Wellcome Trust Sanger Institute (WTSI). Ensembl generates genome annotation for vertebrates and other eukaryotic species, and the information is made freely available to the research community [39]. According to the latest Ensembl release 81, a total 23,636 genomes from 4,991 species are available. The NCBI also hosts genome sequence annotation data of over 1000 organisms including bacteria, archaea, eukaryote, viruses, phages, viroids, plasmids, and organelles. The UCSC genome browser is maintained by the UCSC Genome Bioinformatics group and provides data for over 90 organisms that belong to vertebrates, deuterostomes, insects, nematodes, yeast, viruses, and others [40]. In addition to the aforementioned data resources, Genome Reference Consortium (GRC), comprising of WTSI, the Genome Institute of Washington University (TGI), EBI, and NCBI ensures that the human, mouse, and zebrafish, and the genome assemblies of other model organisms are continuously updated and properly maintained.

Irrespective of the source, it is always recommended to use the latest genome sequence and its annotation. Zhao et al. [41] demonstrated that the choice of a gene model (annotation information/annotation catalog) has a dramatic effect on both gene quantification and differential analysis. We would recommend using Ensembl as it provides more detailed annotation of the genomic features.

#### *2.2.1.2. Choice of read aligner*

One of the most challenging parts of RNASeq analysis is mapping the sequencing reads to the genome correctly, especially for eukaryotes where presence of splicing events adds to the

http://dx.doi.org/10.5772/61789

complexity. Multiple aligners, which can be divided into two categories, are available for aligning short-reads to the genome:

1. Non-spliced aligners: These aligners do not handle splicing events and are therefore suitable for prokaryotic RNASeq analysis only.

2. Spliced aligners: These aligners can place spliced reads across introns and determine exon–intron boundaries. Therefore, these are preferred for eukaryotic RNASeq analysis.

The non-spliced aligners can be further classified, on the basis of the algorithm used, into two categories:

- Hash table-based aligners: This set of aligners uses a seed sequence to identify alignment candidates, which are then either extended or discarded using more precise dynamic programming alignment algorithms. These aligners can be further divided, based upon the approach of finding a seed, into two groups:

  a. Reference indexing: Aligners create index using reference genome. Examples include BFAST [42], Novoalign (http://www.novocraft.com), GNUMAP [43], SHRiMP2 [44], Mosaik [45].

  b. Read indexing: Aligners use read-based index. Examples include MAQ [46], RMAP [47], and RazerS [48].

- FM-index-based aligners: This set of aligners creates FM-index of the genome using Burrows Wheeler Transform data compression algorithm. FM-index's compressed, yet searchable suffix array-like structure makes these aligners both memory-efficient and ultrafast. Examples include Bowtie1 [49], Bowtie2 [50], BWA [51], and SOAP2 [52].

Example of spliced aligners include GSNAP [53], MapSplice [54], SpliceMap [55], STAR [56], and TopHat2 [57]. GSNAP can identify a splice site in two ways: first, by evaluating the surrounding genomic sequence using probabilistic models of donor and acceptor splice site; second, by utilizing user-provided database of known exon–intron boundaries, which improves the sensitivity and specificity of the tool. Both MapSplice and TopHat2 use a two-step algorithm where in the first step potential splice sites are detected, which are then used in the second step to find correct map of reads. MapSplice is a *de novo* spliced aligner, whereas TopHat2 can perform both *de novo* and gene-annotation-based splice alignment. TopHat2 incorporates Bowtie1 or Bowtie2, in the back-end, for initial alignments. SpliceMap is also a *de novo* splice aligner, which is highly sensitive and specific in finding novel splice junctions without using any existing gene model information in arbitrary RNASeq read lengths. Another splice-aware aligner, STAR, utilizes sequential maximum mappable seed search in uncompressed suffix arrays followed by seed clustering and stitching procedure. It has been evaluated to be the fastest aligner among the above-listed spliced aligners with lowest false-positive rate at high sensitivity [56]. However, its RAM requirement is higher as compared to its counterparts.

The latest addition to the list of spliced aligners is HISAT (Hierarchical Indexing for Spliced Alignments of Transcripts) [58], which is claimed to be the fastest aligner currently available.

The reason for this highly efficient system is believed to be the indexing scheme it utilizes. As compared to its counterparts, HISAT uses two different types of indexes instead of a single index: (i) a whole-genome FM index to anchor each alignment, and (ii) numerous local FM indexes for very rapid extension of these alignments. HISAT is 50 times faster than TopHat2, 12 times faster than GSNAP, and slightly faster than STAR [56]. In addition, HISAT requires comparable amount of RAM as TopHat2 but maximum 20% of RAM as GSNAP or STAR needs. Similar to TopHat2, HISAT also uses Bowtie2 in the back-end. Furthermore, it is the only aligner that can work directly on an SRA file, which eliminates the sra to fastq file conversion requirement.

Considering the options available, selecting the right aligner is a nontrivial task and there are several publications comparing the read aligners. Fonseca et al. [59] published a feature-level comparison of 60 mappers and highlighted the difficulties in determining the best aligner (in terms of accuracy and speed). Other comparative studies include one by Lindner and Friedel [60] on non-spliced aligners and another by Engstrom et al. [61] on spliced aligners.

Answers to the following questions may help to choose a suitable aligner:

**1.** Does the genome sequence belong to a prokaryote (where a gene lacks intron) or eukaryote (where a gene has introns)?

If the genome is bacterial (example of a prokaryote), then computationally intensive splice aligners such as TopHat2 or STAR are not required. In this case, non-splice aligners such as Bowtie1, Bowtie2, or BWA are more appropriate because of the contiguous read mapping to the reference genome. On the contrary, for eukaryotic genomes such as human/mouse, where the reads will span an exon boundary and therefore a part of it will not map contiguously on the reference genome; it is better to use a splice aligner that can identify splice sites.

**2.** Are the sequence data available in base space or color space format?

If the data are generated from a SOLiD sequencing platform, they will be in color space format and almost all recently developed tools do not support color space data. In this case, the only available options are aligners such as BWA (older than 0.6.x), Bowtie1, and TopHat2.

**3.** Does the aim of RNASeq experiment include calling variants in transcripts?

In experiments where the aim is to find variants in transcripts, mapping quality plays a crucial role, and hence it is advisable to use only aligners that provide accurate mapping quality. BWA and STAR aligners are suitable for this purpose; however, Bowtie 1 is not because it does not assign appropriate quality score to the mapped reads.

Additionally, one should also consider the comparative precision and recall statistics, CPU, and RAM requirements of the aligners. In addition to the aligners used, the data type itself plays a critical role in the quality of mapped data. For example, paired-end information improves mapping accuracy and, therefore, paired-end data are favored over single-end data for RNASeq experiment.

The aligned read data generated from aligners mentioned in the previous section are stored in Sequence Alignment/Map (SAM) file format, which is a gold standard to store alignment

data. The SAM format has been created by the SAM/BAM format specification working group (https://samtools.github.io/hts-specs/SAMv1.pdf) for standardizing the format in which aligned data are stored. A SAM file contains information about the reference sequence name, query sequence name, alignment position and direction of the read on the genome, mapping quality, etc. However, SAM files are typically very large; hence, these files are converted into binary counterpart known as BAM (Binary of SAM) files. This is done typically using SAMtools [62], which provides a set of programs to manipulate the alignment files. Alignment files can be further manipulated with utilities such as SAMtools and Picard (http://broadinstitute.github.io/picard/) to efficiently retrieve reads and regions of interest.

*2.2.1.3. Choice of annotation source*

Depending upon the biological question of interest, one may wish to perform expression study either on known transcripts only, as per a given annotation catalog, or on reconstructed transcriptome built using a known reference annotation. This enables the quantification of novel genes/isoforms in addition to the known ones. In the first case, the mapped reads and the annotation catalog can be used to assign read counts to each feature (genes/transcripts) using a tool like htseq-count [63], and then perform statistical analysis to identify the differentially expressed genes/isoforms. In the second case, transcriptome reconstruction is required prior to differential expression analysis. It requires assembly of reads into transcription units using either the reference-based or de novo assembly approach. Given a reference genome and an annotation catalog, there are tools such as Cufflinks [64, 65] that first map all the reads to the genome and then use spliced reads directly to reconstruct the transcriptome. It generates a GTF file that contains the assembled isoforms along with isoform-level relative abundance in Fragments Per Kilobase of exon model per Million mapped fragments (FPKM) units [65].

*2.2.2. De novo transcriptome assembly*

Building a transcriptome using *de novo* methods is a powerful way to create the transcriptome of a divergent or novel species. Mainly three features affect the quality of assembled transcripts: a) type of transcript: presence of repeats, polymorphisms, splicing event, complexity of organism, e.g., ploidy level, GC content; b) sequencing technology: library preparation, sequencing accuracy; c) bioinformatics workflow: assembly algorithms and annotation. Currently available *de novo* assemblers have different sensitivity, and specificity in terms of transcript identification are error-prone, and lead to fused transcripts, splicing errors, and gaps [66]. In order to enhance the sensitivity and specificity one can take the combined approach, which employs *de novo* assembly method with reference-guided approach.

*2.2.2.1. De novo assembly approaches*

There are several algorithms available for *de novo* transcriptome assembly (Table 3). In *de novo* transcriptome assembly, contigs or transfragments are created from overlapping reads. Process of assembly involves either de Bruijn graphs construction using k-mers or overlap-layout-consensus (OLC) approach for short and long reads, respectively [67].

| Tool name | Algorithm | Read type | Reference |
|---|---|---|---|
| Trinity | de Bruijn graph | Single and Paired end | [78] |
| Velvet-Oases | de Bruijn graph | Single and Paired end | [74, 77] |
| SOAPdenovo-Trans | de Bruijn graph | Single and Paired end | [80] |
| IDBA-tran | de Bruijn graph | Paired end | [178] |
| Trans-ABySS | de Bruijn graph | Single and Paired end | [79] |
| EBARDe novo | Extension, Bridging, and Repeat-sensing *de novo* | Paired end | [179] |
| Bayesembler | Bayesian model | Paired end | [180] |
| Mira | Overlap graph | Single and Paired end | [68] |

**Table 3.** A list containing different *de novo* transcriptome assemblers

Overlap-Layout-Consensus (OLC) approach:

OLC approach was initially developed for reconstruction of the genome from Sanger sequence and EST (Expressed sequenced tag) data. As the name suggests, in the OLC approach, the read data are searched for overlapping sequences and merged to create longer reads. Depending on the volume of data and complexity of genome (e.g., repeats), the OLC approach is computation- intensive. Some of the OLC-based assemblers are MIRA [68], Newbler (from Roche/454 Life Sciences), and CAP3 [69]. The assemblers using the OLC approach are more suitable for small volume of data, not sensitive to repeat region detection and resolution, and cannot handle the high-depth short read data generated from sequencers such as Illumina. The Eulerian path assemblers, which are based on *de Bruijn* graph algorithms [70], are more suitable for the high-depth short read data and are discussed in detail below.

*De Bruijn*-graph-based approach:

*De Bruijn* graph is a mathematical graph that uses a substring of letters (here nucleotides) of length k to represent nodes. Nodes are connected if shifting a substring by one nucleotide creates an exact k-1 overlap between the nodes [70]. *De Bruijn* graph can be created for both small as well as large sequences. Based upon the defined k-mer (a nucleotide substring of length k) length, reads are broken in k-length to generate substrings. Using these substrings, *de Bruijn* graph is generated in which each unique substring represents a node (or vertex) connected with overlaps between the last k-1 nucleotides of the previous sequence with the first k-1 nucleotides of the subsequent sequence [71]. Identical overlaps of k-mers are merged and counted while creating the graph. If the assembler finds differences in the nodes, the graph is branched. Upon subsequent identity and overlap in the nodes, the graph will join the ends. Presence of single nucleotide difference between the sequence data gives rise to bubbles in the graph. In the case of RNASeq data, occurrence of large bubbles and open-ended branches in the graph suggests presence of alternative splicing and alternative transcription start and end. Occurrence of small bubbles can be due to single nucleotide variation or sequencing errors [72]. In most of the *de Bruijn*-graph-based assemblers, the preferred value of k-mer is usually an odd number in order to avoid reverse complement of k-mers. The chosen size of k-mer has

great impact on the assembly process as using a large k-mer can result in a unique *de Bruijn* graph, but this approach is computationally intensive. On the other hand, using small k-mers can result in a fragmented assembly. According to some of the previous studies it has been observed that smaller k-mers can be useful in more accurate transcriptome assembly of lowly expressed genes whereas larger k-mers perform better for abundant transcripts [73-75]. It is therefore essential to identify the optimal k-mer for the sequence being assembled and it depends to a large extent on the read length, sequencing depth, sequencing error rate, and the complexity of the genome. Additionally, using directionality of the read from paired-end data, assemblers can generate more accurate assembly as compared to single-end data [76]. Some of the most commonly used *de Bruijn*-graph-based assemblers are: Velvet/Oases [74, 77], Trinity [78], Trans-Abyss [79], SOAPdenovo-Trans [80].

Oases: Oases has a set of algorithms that post-processes the assembly generated by Velvet at different k-mers such as dynamic filtering of the noise, resolution of alternative splicing transcripts, and merging of the multiple assemblies generated using different k-mers (www.ebi.ac.uk/~zerbino/oases/). Data generated from different k-mers are merged to generate a complete assembly. Oases works well for the correction of errors and resolution of repeats in the case of paired-end data.

Trinity: Trinity uses three steps to produce transcriptome assembly: inchworm, chrysalis, and butterfly. Inchworm builds initial sets of contigs using k-mer graphs. Chrysalis groups these contigs and builds *de Bruijn* graphs from them. Butterfly simplifies and resolves the graphs to generate the final set of transcripts containing spliced variants and isoforms.

Trans-Abyss: Trans-Abyss considers multiple assemblies generated from Abyss to optimize the assembly and can tackle varying coverage of the transcripts very well.

SOAPdenovo-Trans: SOAPdenovo-Trans is derived from the genome assembler, SOAPdenovo2 [81] and is known to construct transcriptome faster than the above-mentioned assemblers.

#### *2.2.2.2. Choosing the transcriptome assembler*

Choosing an assembly algorithm is difficult as it depends on a number of factors such as read type, length, and complexity of the genome. Some instrument vendors such as Roche provide assembly algorithms, e.g., Newbler, which can handle the long read data and the homopolymer issue frequently observed in the data generated from 454. A recent study using peanut plant RNASeq data suggests that performance of Trinity is better than TransAByss and SOAPdenovo-Trans when raw reads are mapped to the reconstructed assembly of the polyploidy transcriptome [82]. Another study suggested use of multiple k-mers and clustering of k-mer assemblies and at the same time identifying unique contigs from each assembly for effective extraction of biological information from transcriptome assembly [83].

#### *2.2.2.3. Assessing quality and accuracy of de novo assembled transcriptome*

Because of sequencing errors and presence of repeats in the genome, it is hard to achieve a perfect assembly. Moreover, different assemblers use different heuristic approaches to assemble the transcriptome, which results in different number of identified transcripts.

Quality and accuracy of assembled transcriptome are assessed in several different ways [84, 85]:

1. Assembly statistics: Most algorithms generate an assembly statistic that includes the number of contigs/transfragments generated, total contigs/transfragments length and singletons, size of the assembly (in number of nucleotides), percentage of reads assembled to transfragments, percent GC content, etc. Assembly statistics provide overview of the organisms' transcriptome.
2. Transfragments/contigs statistics: This statistics includes lengths of the largest and shortest transfragments, average and median length of transfragments, and N50 of assembled transcriptome. N50 of the assembly is calculated by sorting the contigs in descending order and the size of the contig that makes the total greater than or equal to 50% of the genome size is regarded as the N50 value. A large N50 is indicative of a more contiguous assembly.
3. Mis-assembly and variations: Some of the major reasons for mis-assembly of the transcriptome are presence of ambiguous bases, repeat regions, insertions, deletions, SNPs, and chromosomal rearrangements in the transcriptome. Percentage of mis-assembled contigs can be calculated by mapping the contigs back to the reference genome. QUAST, a tool, generates consolidated report on mis-assembly statistics [84].
4. Number of transfragments matching with the closest reference genome: Once transcripts are assembled, it can be compared against a closely related species/genome. Assembly is considered to be of high quality if the number of reference transcripts matching with the transfragments is high. However, the genes that are not expressed, or lowly expressed, might not be captured.
5. Hybrid or fused transcripts: Hybrid transcripts result from joining of two or more different transcripts and hence matching to different locations of the genome. Reasons for hybrid transcript generation are sequencing error, improper trimming of the adapter/contaminant from the raw read, similarity of the transcripts, assembly algorithm's parameters, etc. Low number of hybrid transcripts reflects better assembly.

### 2.3. Quantification

#### Choice of expression unit: CPM, RPKM, FPKM, TPM, or read count

Once the read data is aligned to the reference genome, the gene expression can be quantitated by read counting at exon, transcript, or gene-level. Here are few possible expression units:

a. Read Count: read counts are number of reads overlapping a genomic feature such as a gene or transcript.
b. CPM (Counts Per Million mapped reads): CPMs are read counts scaled by the number of fragments sequenced times one million. This unit is used in a differential expression analysis R package edgeR [86].
c. RPKM (Reads Per Kilobase of transcript per Million): RPKM for a feature is computed by dividing the number of read counts by it length and total number of reads sequenced, followed by multiplication with one billion [12]. Applicable only for single-end data.

http://dx.doi.org/10.5772/61789

**d.** FPKM (Fragments Per Kilobase of transcript per Million): similar as RPKM. But takes into account a fragment (not reads) [65]. For pair-end data, there will be two reads for a single fragment of genome while for single-end data, there will be one read for a single fragment. Both the situations will add only one count.

**e.** TPM (Transcripts Per Million): TPM for a transcript is calculated by dividing the ratio of its read counts over its length by the summation of ratios for all the transcripts, and multiplying with one million [87]. Especially for transcript abundance.

### 2.4. Normalization

**Why should one normalize the expression data?**

RNASeq experiments have multiple sources of systematic variations introduced through inter-sample differences such as difference in library size (sequencing depth) or unwanted variations due to batch effects such as sampling time or different sequencing technology [12] or through intra-sample differences such as difference in read length [88] or GC content between genes [89, 90]. These variations, if ignored, can dramatically reduce the accuracy of statistical inference and hence should be removed or controlled during statistical analysis. Therefore, read count and FPKM of a feature, as calculated for example by htseq-count and Cufflinks, respectively, may not be appropriate to compare across features and samples without normalization.

Normalization is a process that aims to ensure that expression estimates are comparable. There are a number of normalization methods, such as:

**a.** Total Count: each read count of a feature expression is divided by total number of mapped reads in that sample and multiplied by the average total count across all the samples.

**b.** Upper Quartile: each feature expression is divided by the upper quartile of expression values, other than 0, in that sample and multiplied by the average upper quartile across all the samples [91]. Upper quartile for FPKMs or fragment counts has been implemented in Cuffdiff2 tool from Cufflinks suite [92].

**c.** Median: each feature expression is divided by the median of these expression values (other than 0) in that sample and multiplied by the average median expression across all the samples.

**d.** Quantile: the distribution of expression values for each sample is made identical [93]. Quantile method is available in R package limma [94].

**e.** Trimmed Mean of M-values (TMM): TMM normalization factor for each sample is computed as the weighted mean of log ratios between a test sample and a reference sample after excluding the features with highest expressions and features with largest log ratios. These factors are rescaled by the mean of normalized library sizes. Finally, each feature expression value is divided by these rescaled normalization factors to get the normalized expression [86, 95]. TMM method has been implemented in R package edgeR [86].

**f.** Median of ratio: the normalization factor for each sample is computed as the median of ratios of expressions of features over their geometric means across all samples. Finally, each feature expression is divided by this factor to get the normalized expression [96]. Median of ratio has been implemented in R packages DESeq [96], DESeq2 [97], and in Cuffdiff2 [92].

Several publications [98, 99] comparing normalization methods suggest that median of ratio is the best method for normalization in differential expression study for mRNASeq experiment.

In addition to normalization methods, several packages have been developed to control batch effects, for example, svaseq [100]. svaseq can work on both, count-based data (e.g., htseq-count generated data) as well as FPKMs (Cufflinks generated data).

### 2.5. Differential expression analysis

Differential expression analysis helps identify genes that are important in the experimental conditions being tested and hence is the most routine analysis performed using the RNASeq data. In RNASeq data, a linear relationship has been observed between the number of reads that map to a transcript and the abundance of the transcript [12]. The goal of differential expression analysis is to compare these read counts for a feature between distinct sample groups and perform a statistical test to determine whether the difference is significant. For this purpose, a distribution is required to be fitted to the count data using generalized linear model (GLM). Based upon the assumption that reads are independently sampled from a population with a given, fixed fractions of genes, it can be said that the read counts will follow a multinomial distribution. This multinomial distribution can be approximated by the Poisson distribution and therefore Poisson distribution has been used to test differential expression in several studies [101-103]. But it has been found that this distribution predicts smaller variations than what is seen in the data. To overcome this issue, negative binomial (NB) distribution and beta negative binomial distribution were proposed. NB has been used in several differential expression tools such as edgeR [86], DESeq [96], DESeq2 (an enhanced version of DESeq) [97], and BaySeq [104]. Though these tools use a common distribution, the method of variance (dispersion) estimation differs, which affects the final outcome of the analysis. Cuffdiff2 uses beta negative binomial distribution to fit fragment counts [92].

Recent advances in this area of research suggest that a combination of Poisson distribution and NB distribution may yield better results. Chen et al. [105] derived a novel algorithm XBSeq from DESeq, where they used Poisson distribution to fit read counts that map to nonexonic regions (considered as sequencing noise) and used NB distribution to fit read counts that map to exonic regions (considered as true signals).

Recently, limma [94], a well-known R package for performing differential expression analysis of microarray data, has been empowered with RNASeq data analysis ability. It does not use the above-mentioned distribution, rather converts count data (or normalized count data) to log-counts per million using voom transformation, then fits a linear model to this data and performs differential expression analysis using an empirical Bayes method.

http://dx.doi.org/10.5772/61789

There is no clear evidence as such about the best tool for differential expression analysis; however, multiple studies comparing available methods have been performed. Soneson and Delorenzi [106] evaluated and compared eleven methods for differential expression analysis on simulated and real RNASeq data, whereas Seyednasrollah et al. [107] compared eight widely used tools on real data sets. Both the studies concluded that no single method is optimal under all circumstances. Soneson and Delorenzi [106] observed that limma performed well under many different conditions and Seyednasrollah et al. [107] found limma and DESeq as the preferred choice. Additionally, these studies have suggested that the method of choice should depend on the experimental conditions that include the number of samples per condition.

## 2.6. Annotation and pathway analysis

### *2.6.1. Annotation of de novo assembled transcriptome*

In addition to transcriptome abundance calculation after mapping the assembled contigs/transfragments to the assembled transcriptome or reference genome and differential expression data analysis, coding regions within *de novo* assembled transcripts can be searched using ORF predictor tools such as Transdecoder (http://transdecoder.github.io/). Further, homologous gene/protein identification of assembled transcripts can be done using tools such as BLAT and BLAST [108].

### *2.6.2. Making sense of the differentially expressed gene list*

List of differentially expressed genes is just the first tangible outcome of an RNASeq experiment. In order to derive biological insight from this list of genes, it is important to identify functional categories of the genes that are differentially expressed and the biological pathways that are enriched as a result of these differentially expressed genes. In order to do so, enrichment analysis is typically performed using publicly available resources such as GO (Biological Processes and Molecular Functions) databases [109], KEGG pathways [110], BioCarta (www.biocarta.com), and Reactome [111].

In a review article, Khatri et al. [112] elaborated the current approaches of pathway analysis and their challenges and divided the existing approaches into three generations:

**a.** First Generation: Overrepresentation Analysis (ORA) approach

This approach statistically evaluates the fraction of genes, among the set of differentially expressed genes, in a particular pathway. There are many tools that follow this approach, for example, Onto-Express [113], GenMAPP [114], GoMiner [115], and DAVID [116, 117]. However, this approach has certain limitations. For example, it does not consider the fold change values associated with the genes, thereby ignoring the extent of regulation. Moreover, it does not consider the gene product interactions that are found in a pathway. This approach also ignores the dependency between the pathways.

**b.** Second Generation: Functional Class Scoring (FCS) approach

This approach addresses few limitations of ORA. It considers all the genes and their expression for pathway enrichment, so as to take into consideration the coordinated changes (irrespective of the magnitude) unlike ORA where only differentially expressed genes were considered and that too without considering their expression levels. Example of such tools include global test [118], GSEA [119].

But this approach too has some limitations. Similar to ORA, this approach ignores the dependency between the pathways and the interaction between gene products in a given pathway.

**c.** Third Generation: Pathway Topology (PT)-based approach

To overcome the limitations of ORA and FCS, the Pathway-Topology-based approach has been devised. It uses pathway knowledgebase to include pathway topology information for enrichment analysis [112]. This information includes genes that are interacting, their mode of interaction (e.g, activation, inhibition), and their location of interaction (e.g, cytoplasm, nucleus). SPIA [120], an R package, is an example of this category of pathway analysis approach, which combines evidence of pathway overrepresentation and unusual signaling perturbations. NetGSA [121] is another method in this category that takes into consideration the change in correlation as well as the change in network structure as experimental condition changes. However, in the absence of high-resolution knowledge databases that can provide knowledge for all conditions, tissue- and cell-specific functions of a gene product; the true pathway topology is rarely inferred. And hence this restricts a researcher to investigate the dynamic states of a system [112].

### 2.7. Visualization

Analyzed RNASeq data can be visualized in many different ways. Several tools such as Cummerbund (an R package), RNAseqViewer for single and multiple sample visualization [122], HeatmapGenerator for heatmap visualization, GOexpress for GO term enrichment visualization (http://www.bioconductor.org/packages/devel/bioc/html/GOexpress.html), RNASeq-specific genome viewers such as RNASeqExpressionBrowse [123], and RNASeq-Browser [124] are available for RNASeq data visualization.

We have recently developed SanGeniX (www.sangenix.com), an easy-to-use client-server-based NGS data analysis application with a highly intuitive user interface (manuscript under preparation). SanGeniX supports primary, secondary, and tertiary analysis of sequence data from Illumina, Ion Torrent, SOliD, and PacBio RS. SanGeniX integrates multiple robust and validated algorithms in the form of predefined workflows and offers flexibility to construct custom workflows for RNASeq (reference-based as well as *de novo*), genome assembly, ChIPSeq and DNASeq (for SNP and CNV calling). For example, in the case of RNASeq workflow, the analysis starts with quality check (using tool FastQC), contaminant/adapter trimming and removal (using Cutadapt and in-house scripts), read mapping using splice aware aligners (using STAR, TopHat2), transcript quantification, differential expression analysis (using Cufflinks packages and DESeq2), and gene ontology, as well as pathway enrichment analysis (using GoMiner) (Figure 2). Further, graphically enriched visuals such as heatmap based on clustering, scatter plot, and volcano plot for differentially expressed genes,

http://dx.doi.org/10.5772/61789

pie chart on gene-ontology-based annotation, visualization of read data in the genome viewer, etc., are generated for easy interpretation of the data (Figure 3). These figures and underlying data can be downloaded in svg, png, and tsv formats. Moreover, the raw output files such as output of mapping in SAM and BAM formats can also be downloaded. The executed workflows can be shared with peers, rerun after changing parameters or tools. SanGeniX is available as cloud-hosted as well as on premise solution and supported on multiple Linux platforms such as Ubuntu, CentOS, and RedHat.


**Figure 2.** Snapshots of RNASeq data analysis workflow canvas in SanGeniX using (A) Cufflinks package and (B) HTSeq and DESeq2 are shown.

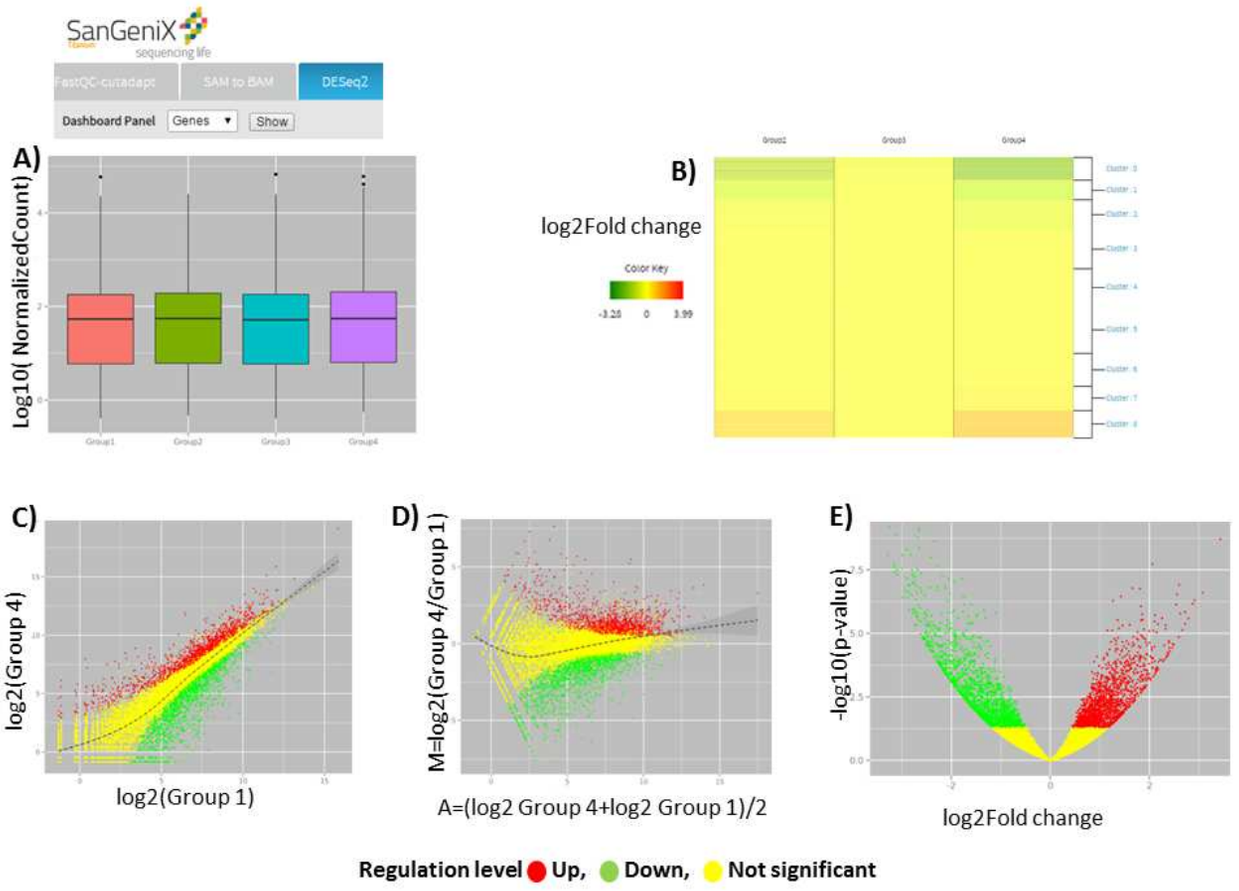


**Figure 3.** Snapshots from RNASeq results dashboard from SanGeniX for an experiment consisting of four groups (or samples). (A) Boxplot: It displays distribution of normalized expression values among different groups. Similar distribution of normalized expression values among the different groups of interest indicates that any technical biases due to difference in sequencing depth have been taken care of. (B) Heatmap is a convenient way to visualize cluster of genes based upon their expression. Here, log2 fold change of genes in three groups with respect to a reference group, Group1 has been plotted. The color-code helps to infer gene expression level. Scatter plot (C), MA plot (D), and Volcano plot (E) present visual investigation of differentially expressed genes between two conditions, for example, here Group 4 and Group 1. Scatter plot helps to quickly compare the expression of a gene between the two conditions, while MA plot depicts trends of difference in expression over the average expression, and Volcano plot helps to spot genes by considering both fold change and test statistic.

## 3. Challenges in RNASeq data generation and analysis

As described above, NGS-based transcriptomic data generation and analysis is a complex and multistep process. Every step has some key challenges that hinder the data analysis.

### 3.1. Library preparation

The process of library preparation is generating cDNA from the large RNA fragments, adding the adapters, and amplifying the cDNA for sequencing. Due to a series of experimental reactions, several biases can be introduced in the library preparation step. In majority of the

http://dx.doi.org/10.5772/61789

cases, fragmentation of RNA or DNA, which plays an important role in the preparation of high-quality sequencing library, is done using physical or enzymatic methods or chemical shearing. The fragmentation of RNA has even coverage in the gene body and hence it is biased toward the gene body as compared to the 5′ and 3′ ends where the coverage is relatively depleted [20]. The library preparation step is further complicated by the presence of several identical short reads and hence duplicate sequences in the library could arise from abundance of RNA molecules. Another source of duplicate sequences in a library could be due to PCR artifacts. These two different scenarios can be assessed by considering biological replicates in the study. In the case of total RNAseq, abundance of ribosomal RNA (rRNA) dominates sequenced reads and hence creates bias if not removed.

### 3.2. Sequencing platform

Sequencing platforms are available from multiple vendors such as Illumina (http://www.illumina.com/), Life Technologies (https://www.lifetechnologies.com/), and Pacific Biosciences (www.pacificbiosciences.com/), and each of the platforms has its set of advantages and disadvantages [35]. In choosing a sequencing platform, some of the factors to be considered are sequencing length, sequencing type (single end or paired end), throughput, error rate, and type of errors in the generated sequence data. The gigabytes of short reads generated from the current platforms are not error-free, which affects the downstream analysis and interpretation. For transcriptome assembly, the larger read length (such as produced from 454, PacBio) is preferred over short read length (as produced by Illumina) as it will result in assembly of the high-quality and reliable transcripts. However, both 454 and PacBio platforms have limited throughput and hence the approach most commonly used is to generate data from multiple platforms and combine the data during analysis.

### 3.3. Mapping

Accurate mapping of RNASeq reads is a challenging issue because of large data volume, slow mapping speed, false-positive splicing events and incorrect estimation of exon–intron boundaries, large genome size, repeat sequences in the genome, and annotation quality of the genome. Usually, aligners search for introns smaller than a fixed length to reduce the computational power, which often leads to missing the splice reads spanning longer introns [66]. Multiple mapping of reads is another major problem that can be due to presence of repeat regions, similar sequences, and number of mismatches allowed in the mapping step. If such reads mapping to multiple regions are discarded, it will lead to gap in the regions that cannot be mapped uniquely, and if it is included, it can lead to false-positive transcription status. Reference-based assembly cannot efficiently detect trans-spliced genes that are formed from splicing and joining of two different precursor mRNAs and found in some disease conditions such as cancer [125, 126]. Additionally, aligners have to cope with sequencing errors, SNP, InDels, other genomics variations and parameters-based, suboptimal mapping outcome. In summary, mapping-based RNASeq analysis can be more effective and complete when reads are long, genome is well-annotated, and it can be combined with *de novo* genome assembly to identify novel transcripts.

### 3.4. Read quantification for the estimation of gene expression

Once the sequenced reads are aligned, gene expression is measured. The most common way of read quantification is counting the number of reads overlapping the exons of a gene and if the exon boundaries are not well-annotated, it may lead to false-positive hits. Another major challenge in read quantification is reads mapping to multiple locations.

### 3.5. Count normalization

There are several methods such as quantile-based normalization, GC-content-based normalization, Poisson model with variable rates for different positions, available to normalize and correct the biasness in the count data for the improved detection of differentially expressed genes [91, 127, 128]. The increasing number of normalization methods requires a state-of-the-art technique for comparing these methods. In the absence of such technique, there is no consensus on the best method for normalization. For example, Zyprych-Walczak et al. [99] found that TMM method worked poorly for them while Dillies et al. [98] found TMM and median of ratio methods to be the best as compared to other methods. The transcript length is another source of bias and leads to detection of more differential expression in longer transcripts compared to shorter transcripts [88].

### 3.6. Differential expression analysis

There are several tools and methods developed for the differential expression analysis comparing differences in gene expression in different conditions (see section 2). Nonparametric methods are not capable of better differential expression detection in the absence of sample replicates and hence parametric methods are preferred for differential expression analysis [129]. A study comparing various differential expression methods suggests that there is no optimized method that can serve well for all the different conditions. As compared to other tools, Cuffdiff performed poorly with large number of false-positives [130]. The accuracy of differentially expressed genes is statistically significant and makes more sense if multiple replicates are used in the analysis.

Similar to the situation as in normalization, picking up the best tool for differential analysis is a tricky job. This is because there is no consensus about the tool best-suited for all experimental setups. Soneson and Dolerenzi [106] found limma performing well under many conditions but it required at least three replicates. Furthermore, they found limma performing worse when dispersion differed between two conditions. They also observed that with large sample sizes DESeq was overly conservative, while edgeR was producing large number of false-positives.

### 3.7. *De novo* assembly

The performance and accuracy of the *de novo* transcriptome assembly is largely dependent on the complexity of the genome (e.g., genome size, number of paralogs, ploidy level), differential read coverage of the sequenced data, and sequencing error. Transcriptome assembly is complex and different from genome assembly in which read coverage is uniform. In contrast, in RNASeq, the abundance of reads vary based upon gene expression, in which case isoforms

http://dx.doi.org/10.5772/61789

originating from same gene can have different expression levels and hence poses significant challenge in estimating the abundance especially for the lowly expressed genes if the sequencing depth is too low. In general, *de novo* transcriptome requires much higher sequencing depth than the reference-based transcriptome assembly.

The *de novo* transcriptome assembly generally consumes more time and is more computation-intensive than reference-based assembly [131]. The number of transfragments produced using the *de novo* approach is quite high, which can be due to multiple similar transcripts/isoforms at the locus from allelic variation, or could be due to artifacts. Additionally, the contiguity and completeness of the *de novo* assembled transcriptome is less than the reference-based assembly especially for the data with less sequencing depth [132].

### 3.8. Deep sequencing versus cost

Another challenge associated with the RNASeq technology is read coverage and cost associated with it. In order to detect lowly expressed genes or rare variants in the coding region, high read coverage is required. According to Nagalakshmi et al. [10], for simple organism such as yeast, which does not undergo alternative splicing, 30 million reads are sufficient to observe genome-wide transcriptome profile [10]. But for larger and complex genomes such as the human genome, higher-depth RNASeq data are required in order to capture the complete transcriptomes. Moreover, in a given organism the number of transcripts expressed in different conditions is different and hence same coverage may not be sufficient to capture all the transcripts expressed under different conditions. Hence, before designing an experiment, one should be aware of both sequencing depth required and the number of samples to be sequenced. If the aim of experiment is to detect rare variants or lowly expressed genes, one should go for high coverage of the transcriptome, whereas, if the aim of the experiment is focused on gene expression differences between different samples (or conditions), one should consider generating replicate data for statistical power [133].

There are other bioinformatics challenges such as data retrieval, storing, unavailability of optimized statistical methods, and high-end compute infrastructure requirement that add to the complexity of transcriptome analysis.

## 4. Applications of RNASeq

RNASeq provides an unprecedented view into the complexity of the transcriptome and hence is a powerful tool to characterize and profile transcriptome on a genome-wide scale. Some of these applications with detailed examples are discussed below.

### 4.1. Transcriptome profiling of economically important plants

Understanding the transcriptome and the functional elements of the economically important plants can provide tremendous insights into biological entities, critical for traits such as disease resistance, productivity, and characteristics such as flavor. Recently, Hu et al. [134] performed

transcriptome assembly and annotation for the spice black pepper. Black pepper is one of the most widely used fruit for adding flavor to food as well for its medicinal properties. The authors were able to identify genes that might participate in piperidine, quinolizidine, indolizidine, and lycopodium alkaloid biosynthesis, of which piperidine alkaloids account for pungent taste and medicinal properties of black pepper. Similarly, Shudeesh et al. [135] performed assembly and annotation of field pea, a legume that is cultivated worldwide for human as well as livestock consumption. Studies have also been undertaken to identify transcriptomes of the pathogens that infect economically important plants and the defense mechanisms deployed by the plants. For example, the transcriptome of coffee leaf rust pathogen Hemileia vastatrix was sequenced by Talhinhas et al. [136] to identify genes/pathways that play a key role in the early stage of the infection, and Yang et al. [137] sequenced the sand pear germplasm with differential resistance to infection by *Alternaria alternata* to identify genes that contribute toward the resistance.

### 4.2. Transcriptome profiling of economically important animals

Similar to the value provided by transcriptome profiling of plants, transcriptome profiling of economically important animals contributes toward better understanding of disease resistance, productivity, breeding, quality of meat, etc., in animals. Ropka-Molik et al. [138] have used the NGS transcriptome profiling approach to identify genes that are differentially expressed between two pig breeds with differences in muscularity that could contribute toward the quality of meat. Gene expression profiles have been generated from different breeds of cows to identify genes that contribute toward milk protein and fat percentage in cow milk [139, 140] and milk yield [141]. Transcriptome profiling has also been used very recently to identify the genes that are differentially expressed in silkworms (*B. mori*) undergoing thermal parthenogenesis [142]. Thermal parthenogenesis is a process that is used in silkworm breeding and selection.

### 4.3. Cancer

Cancer is a complex and heterogeneous genetic disorder that results from either inherited or somatic genetic variations such as single nucleotide variations (SNV), insertions, deletions, copy number variations, dysregulation of gene expression, and epigenetic modifications. As changes in the gene expression pattern play a key role in tumorigenicity [143], metastasis [144], prognosis [145], and relapse [146, 147], gene expression profiling has been used extensively in cancer research and diagnosis. OncotypeDx (http://www.oncotypedx.com/) is a gene-expression-based commercially available test that is used for breast cancer, colon cancer, and prostate cancer diagnosis and prognosis.

Contrary to microarrays and RT-PCR-based approaches used earlier, RNASeq, which can detect coding and noncoding RNA, strand orientation, and genetic variants all in one go, is a very powerful tool in deciphering the complex transcriptome changes usually found in cancer. One of the most comprehensive studies published recently is the transcriptome profiling of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types [148]. In this study, in addition to identifying tissue specific gene signature, the authors were able to identify

http://dx.doi.org/10.5772/61789

a 14-gene signature that accurately distinguished the cancer samples from the normal. Using a whole transcriptome sequencing approach, Koh et al. [149] recently reported 14 candidate genes that are important in rhabdoid glioblastoma (R-GBM) tumor, a rare form of GBM. Similarly, RNASeq approach was used to identify gene signature in flow-sorted viable EpCAM + tumor epithelial cells and CD45+ tumor-infiltrating immune cells that were obtained from cervical cancer samples [150]. The authors identified TCL1A as a novel biomarker, found specifically in the immune cells, for predicting survival in cervical cancer patients.

The aforementioned studies highlight the varied approaches that can be used for identifying biomarkers or gene signatures associated with distinct cancer characteristics.

### 4.4. Reproductive health

With the advancing parental age and a desire to limit the number of pregnancies, many couples opt for assisted reproduction for childbearing. The advanced parental age is a key factor that contributes toward the complications in assisted reproduction, and genomics-based approaches are widely used to ensure a high success rate. Gene expression changes in ovarian granulosa cells in women >35 years of age include downregulation of polo-like kinase pathway, which plays an important role in cell cycle arrest of granulosa cells, and the G2/GM checkpoint pathway [151]. Another very recent study also used the RNASeq approach to identify differential gene expression profiles in women with successful pregnancy and a failed pregnancy through assisted reproduction [152]. The authors found that the genes that were differentially expressed played a role in immune response and inflammation, oocyte meiosis, and rhythmic process.

The application of RNASeq in reproductive health is relatively new and as more knowledge is gleaned through this, it might be possible to develop a signature that can be used for predicting the success of assisted reproductive approach.

### 4.5. Developmental disorders

Developmental disorders are ones in which the child develops slower than peers in areas such as motor function, social skills, and cognitive ability. Developmental disorders include Austism, Asperger's Syndrome, Attention Deficit Hyperactivity Disorder (ADHD), Rett Syndrome, and stereotypic movement disorder, to name a few. Gene expression profiling has been used extensively in Austism and genes involved in neuronal action potential, myelination, axon ensheathment, cellular development, and cellular proliferation have been found to be differentially expressed in autistic children [153]. Another study, using an *in vitro* model of Autism found expression differences in genes involved in cell proliferation, neuronal differentiation, and synaptic assembly [154]. Similarly, a gene expression study in Rett Syndrome [155], which is a rare variant of Austism, has identified genes involved in mitochondrial functions, cellular protein metabolic processes, and RNA processing and DNA organization to be differentially regulated.

In addition to the applications listed here, gene expression profiling can be used in number of other human disorders such as diabetes, hypertension, psychiatric disorders, and infectious diseases.

## 5. Future perspective

RNASeq technology is proving to be a valuable tool to study known and novel transcripts of an organism by providing more insights into the role of gene expression in development, differential expression between different conditions, changes in gene expression in disease progression, alternative splicing events, RNA editing, fusion transcripts, allele-specific expression, etc. This technology is revolutionizing the field of plant and animal transcriptome, where many of the species lack reference genome because of genome size and complexity. Metatranscriptomic-NGS technology employed to study microbial transcriptome is another emerging area of research in which construction of transcriptome assembly has led to simultaneous identification of thousands of transcripts from the microbial community of the human gastrointestinal tract [156], and the marine [157, 158] and soil [159]. Because of the fact that gene expression levels vary significantly from one cell to another, researchers are now moving toward single-cell transcriptomics, in which cell-to-cell variability on a genome-wide scale can be profiled. Hence, transcriptome of single cell can be probed more efficiently as compared to cell population where average transcript abundance of population is seen [160, 161]. A recent study by Sasagawa et al. developed the method Quartz-Seq for individual cell isolation followed by RNA sequencing and distinguished mouse embryonic stem cells from primitive endoderm based upon transcriptome profile as well as cell-to-cell stochastic variation [162]. Another recently developed method, RaceID, is very useful in identifying rare cell types in healthy and diseased tissues using mRNA sequencing [163]. Tissue-specific RNASeq is another emerging area of research that can reveal tissue-specific requirement of RNA expression. A recent study done on 13 different cell types discovered many tissue-specific and novel miRNAs, which suggests that the repertoire of human miRNA is more extensive than our current knowledge [164]. RNASeq is used as a powerful tool for clinical application as well. A recent study developed exome capture RNASeq protocol for degraded clinical formalin-fixed samples, which has shown to work successfully on prostate cancer samples suggesting that capture transcriptome study can be used beyond cell lines and in the clinical setting [165].

Moreover, there are several publicly available RNASeq data repositories such as ENCODE (https://www.**encode**project.org/), TCGA (www.cancergenome.nih.gov), and The Geuvadis Project (http://www.geuvadis.org/), which provide enormous amount of data to researchers to conduct genome-wide analyses beyond traditional gene expression and profiling analysis. Mining data from public repositories will provide new insights into the transcriptome and hence enable researchers to gain more information on gene regulation, which has been previously neglected.

Sequencing method and experimental protocols are also continuously improving to reduce the challenges associated with the technology. Platforms such as PacBio can produce a full-length transcript in a single read, which can eventually eliminate the transcript assembly step of the data analysis.

Additionally, to cater to the high volume of data and the demand for high-end computational resources for the transcriptome assembly, many assemblers have started supporting parallel data processing, which has significantly reduced the time required for the assembly (reviewed in [66]). Cloud computing is another lucrative approach for parallel computing, which is scalable and can be used as per the user requirement [166].

## Author details

Krishanpal Anamika*, Srikant Verma, Abhay Jere and Aarti Desai

*Address all correspondence to: anamika_krishanpal@persistent.co.in

LABS, Persistent Systems Limited, Pingala - Aryabhata, Erandwane, Pune, India

## References

[1] Elgar G, Vavouri T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet*. 2008;24(7):344–352. DOI: 10.1016/j.tig.2008.04.005

[2] Blignaut M. Review of non-coding RNAs and the epigenetic regulation of gene expression. *Epigenetics*. 2012;7(6):664–666. DOI: 10.4161/epi.20170

[3] Mattick JS, Dinger ME. The extent of functionality in the human genome. *HUGO J*. 2013;7:2. DOI: 10.1186/1877-6566-7-2

[4] Shabalina SA, Spiridonov NA. The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biol*. 2004;5(4):105. DOI: 10.1186/gb-2004-5-4-105

[5] Davuluri RV, Suzuki Y, Sugano S, Plass C, Huang TH. The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet*. 2008;24(4):167-177. DOI: 10.1016/j.tig.2008.01.008

[6] Chen M, Manley JL. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol*. 2009;10(11):741–754. DOI: 10.1038/nrm2777

[7] Tian B, Manley JL. Alternative cleavage and polyadenylation: the long and short of it. *Trends Biochem Sci*. 2013;38(6):312–320. DOI: http://dx.doi.org/10.1016/j.tibs.2013.03.005

[8] Kochetov AV. Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. *Bioessays*. 2008;30(7):683–691. DOI: 10.1002/bies.20771

[9] Eddy S R. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet*. 2001;2(12):919–929. DOI: 10.1038/35103511

[10] Nagalakshmi U, Wang Z, Waern K et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*. 2008;320(5881):1344–1349. DOI: 10.1126/science.1158441

[11] Sultan M, Schulz MH, Richard H et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*. 2008;321(5891): 956–960. DOI: 10.1126/science.1160342

[12] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNASeq. *Nat Methods*. 2008;5(7):621–628. DOI: 10.1038/nmeth.1226

[13] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403(6769):503–511. DOI: 10.1038/35000501

[14] Schena M, Shalon D, Davis R W, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995;270(5235):467–470. DOI: 10.1126/science.270.5235.467

[15] Okoniewski MJ, Miller CJ. Hybridization interactions between probesets in short oligo microarrays. *BMC Bioinformatics*. 2006;7:276. DOI: 10.1186/1471-2105-7-276

[16] Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, Stefano GB. Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Med Sci Monit Basic Res*. 2014;20:138–142. DOI: 10.12659/MSMBR.892101

[17] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science*. 1995;270(5235):484–487. DOI: 10.1126/science.270.5235.484

[18] Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, Hayashizaki Y, Carninci P. CAGE: cap analysis of gene expression. *Nat Methods*. 2006;3:211–222. DOI: 10.1038/nmeth0306-211

[19] Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol*. 2000;18(6):630–634. DOI: 10.1038/76469

[20] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10:57–63. DOI: 10.1038/nrg2484

[21] Tuch BB, Laborde RR, Xu X et al. Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS ONE*. 2010;5(2):e9317. DOI: 10.1371/journal.pone.0009317

[22] Berger MF, Levin JZ, Vijayendran K et al. Integrative analysis of the melanoma transcriptome. *Genome Res*. 2010;20(4):413–427. DOI: 10.1101/gr.103697.109

[23] Jima DD, Zhang J, Jacobs C, Richards KL, Dunphy CH et al. Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs. *Blood*. 2010;116(23):e118–e127. DOI: 10.1182/blood-2010-05-285403

[24] Twine NA, Janitz K, Wilkins MR, Janitz M. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. *PLoS ONE*. 2011;6:e16266. DOI: 10.1371/journal.pone.0016266

[25] Ku GM, Kim H, Vaughn IW et al. Research resource: RNASeq reveals unique features of the pancreatic β-cell transcriptome. *Mol Endocrinol*. 2012;26(10):1783–1792. DOI: 10.1210/me.2012-1176

[26] Morán I, Akerman I, van de Bunt M, Xie R, Benazra M, Nammo T, Arnes L, Nakić N, García-Hurtado J, Rodríguez-Seguí S, Pasquali L, Sauty-Colace C, Beucher A, Scharfmann R, van Arensbergen J, Johnson PR, Berry A, Lee C, Harkins T, Gmyr V, Pattou F, Kerr-Conte J, Piemonti L, Berney T, Hanley N, Gloyn AL, Sussel L, Langman L, Brayman KL, Sander M, McCarthy MI, Ravassard P, Ferrer J. Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metab*. 2012;16(4):435–448. DOI: 10.1016/j.cmet.2012.08.010

[27] Sun X, Zhou S, Meng F, Liu S. De novo assembly and characterization of the garlic (Allium sativum) bud transcriptome by Illumina sequencing. *Plant Cell Rep*. 2012;31(10):1823–1828. DOI: 10.1007/s00299-012-1295-z

[28] Franssen SU, Shrestha RP, Bräutigam A, Bornberg-Bauer E, Weber AP. Comprehensive transcriptome analysis of the highly complex Pisum sativum genome using next generation sequencing. *BMC Genomics*. 2011;11:12:227. DOI: 10.1186/1471-2164-12-227

[29] Garg R, Patel RK, Tyagi AK, Jain M. De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res*. 2011;18(1):53–63. DOI: 10.1093/dnares/dsq028

[30] Takehisa H, Sato Y, Igarashi M, Abiko T, Antonio BA, Kamatsuki K, Minami H, Namiki N, Inukai Y, Nakazono M, Nagamura Y. Genome-wide transcriptome dissection of the rice root system: implications for developmental and physiological functions. *Plant J*. 2012;69(1):126–140. DOI: 10.1111/j.1365-313X.2011.04777.x

[31] Alagna F, D'Agostino N, Torchia L, Servili M, Rao R, Pietrella M, Giuliano G, Chiusano ML, Baldoni L, Perrotta G. Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics*. 2009;10:399. DOI: 10.1186/1471-2164-10-399

[32] Pingault L, Choulet F, Alberti A, Glover N, Wincker P, Feuillet C, Paux E. Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome. *Genome Biol*. 2015;16:29. DOI: 10.1186/s13059-015-0601-9

[33] Chapman MA. Transcriptome sequencing and marker development for four underutilized legumes. *Appl Plant Sci*. 2015;3(2):1400111. DOI: 10.3732/apps.1400111.

[34] Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet*. 2010;11(1):31–46. DOI: 10.1038/nrg2626

[35] Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of next-generation sequencing systems. *J Biomed Biotechnol*. 2012;2012:251364. DOI: 10.1155/2012/251364

[36] Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 2011;27(6):863–864. DOI: 10.1093/bioinformatics/btr026

[37] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 2011;17(1):10–12. DOI: 10.14806/ej.17.1.200

[38] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*. 2014;30(15):2114–2120. DOI: 10.1093/bioinformatics/btu170

[39] Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S et al. Ensembl 2015. *Nucl Acids Res*. 2015;43:D662–D669. DOI: 10.1093/nar/gku1010

[40] Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M. et al. 2015. The UCSC Genome Browser database: 2015 update. *Nucl Acids Res*. 2015;43:D670–D681. DOI: 10.1093/nar/gku1177

[41] Zhao S, Zhang B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in context of RNASeq read mapping and gene quantification. *BMC Genomics*. 2015;16:97. DOI: 10.1186/s12864-015-1308-8

[42] Homer N, Merriman B, Nelson SF. BFAST: An alignment tool for large scale genome resequencing. *PLoS ONE*. 2009;4(11):e7767. DOI: 10.1371/journal.pone.0007767

[43] Clement NL, Snell Q, Clement MJ, Hollenhorst PC, Purwar J, Graves BJ, Cairns BR, Johnson WE. The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics*. 2010;26(1):38–45. DOI: 10.1093/bioinformatics/btp614

[44] David M, Dzamba M, Lister D, Ilie L, Brudno M. SHRiMP2: Sensitive yet Practical Short Read Mapping, *Bioinformatics*. 2011;27(7):1011–1012. DOI: 10.1093/bioinformatics/btr046

[45] Lee W, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. MOSAIK: A Hash-based algorithm for accurate Next-Generation Sequencing short-read mapping. *PLoS ONE*. 2014;9(3):e90581. DOI: 10.1371/journal.pone.0090581

[46] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*. 2008;18(11):1851–1858. DOI: 10.1101/gr.078212.108

[47] Smith AD, Chung W, Hodges E, Kendall J, Hannon G, Hicks J, Xuan Z, Zhang MQ. Updates to the RMAP short-read mapping software. *Bioinformatics*. 2009;25(21):2841–2842. DOI: 10.1093/bioinformatics/btp533

[48] Weese D, Emde A, Rausch T, Doring A, Reinert K. RazerS-fast read mapping with sensitivity control. *Genome Res*. 2009;19(9):1646–1654. DOI: 10.1101/gr.088823.108

[49] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10:R25. DOI: 10.1186/gb-2009-10-3-r25

[50] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Natur Meth*. 2012;9:357–359. DOI: 10.1038/nmeth.1923

[51] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009:25(14):1754–1760. DOI: 10.1093/bioinformatics/btp324

[52] Li R, Yu C, Li Y, Lam T, Yiu S, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*. 2009;25(15):1966–1967. DOI: 10.1093/bioinformatics/btp336

[53] Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*. 2010;26(7):873–881. DOI: 10.1093/bioinformatics/btq057

[54] Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucl Acids Res*. 2010;38(18):e178. DOI: 10.1093/nar/gkq622

[55] Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNASeq data by SpliceMap. *Nucl Acids Res*. 2010:38(14): 4570–4578. DOI: 10.1093/nar/gkq211

[56] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNASeq aligner. *Bioinformatics*. 2013;29(1):15–21. DOI: 10.1093/bioinformatics/bts635

[57] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14:R36. DOI: 10.1186/gb-2013-14-4-r36

[58] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12(4):357–360. DOI: 10.1038/nmeth.3317

[59] Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. *Bioinformatics*. 2012;28(24):3169–3177. DOI: 10.1093/bioinformatics/bts605

[60] Lindner R, Friedel CC. A comprehensive evaluation of alignment algorithms in the context of RNASeq. *PLoS ONE*. 2012;7(12):e52403. DOI: 10.1371/journal.pone.0052403

[61] Engstrom PG, Steijger T, Sipos B, Grant GR, Kahles A, Ratsch G, Goldman N, Hubbard TJ, Harrow J, Guigo R, Bertone P, The RGASP Consortium. Systematic evaluation of spliced alignment programs for RNA-seq data. *Natur Meth*. 2013;10(12):1185–1191. DOI: 10.1038/nmeth.2722

[62] Li H, Handshakes B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence Align-

ment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079. DOI: 10.1093/bioinformatics/btp352

[63] Anders S, Pyl PT, Huber W. Htseq-a Python framework to work high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166–169. DOI: 10.1093/bioinformatics/btu638

[64] Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNASeq. *Bioinformatics*. 2011;27(17):2325–2329. DOI: 10.1093/bioinformatics/btr355

[65] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ. Pachter L. Transcript assembly and quantification by RNASeq reveals unannotated transcripts and isoform switching during cell differentiation. *Natur Biotechnol*. 2010;28:511–515. DOI: 10.1038/nbt.1621

[66] Martin JA, Wang Z. Next-generation transcriptome assembly. *Nat Rev Genet*. 2011;12(10):671–682. DOI: 10.1038/nrg3068

[67] Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. *Nat Meth*. 2009;6:S6–S12. DOI: 10.1038/nmeth.1376

[68] Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, Suhai. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res*. 2004;14:1147–1159. DOI: 10.1101/gr.1917404

[69] Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Res*. 1999;9:868–877. DOI: 10.1101/gr.9.9.868

[70] Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A*. 2001;98(17):9748–9753. DOI: 10.1073/pnas.171285098

[71] Compeau PEC, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol*. 2011;29:987–991. DOI: 10.1038/nbt.2023

[72] Schliesky S, Gowik U, Weber AP, Brautigam A. RNA-Seq assembly – Are we there yet? *Front Plant Sci*. 2012;3:220. DOI: 10.3389/fpls.2012.00220

[73] Surget-Groba Y, Montoya-Burgos JI. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res*. 2010;20(10):1432–1440. DOI: 10.1101/gr.103846.109

[74] Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNASeq assembly across the dynamic range of expression levels. *Bioinformatics*. 2012;28(8):1086–1092. DOI: 10.1093/bioinformatics/bts094

[75] Gruenheit N, Deusch O, Esser C, Becker M, Voelckel C, Lockhart P. Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants. *BMC Genomics*. 2012;13:92. DOI: 10.1186/1471-2164-13-92

[76] Gongora-Castillo E, Buell CR. Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence. *Nat Prod Rep*. 2013;30(4):490–500. DOI: 10.1039/c3np20099j

[77] Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 2008;18(5):821–829. DOI: 10.1101/gr.074492.107

[78] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNASeq data without a reference genome. *Nat Biotechnol*. 2011;29:644–652. DOI: 10.1038/nbt.1883

[79] Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJ, Hoodless PA, Birol I. De novo assembly and analysis of RNASeq data. *Nat Methods*. 2010;7(11):909–912. DOI: 10.1038/nmeth.1517

[80] Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, Zhou X, Lam TW, Li Y, Xu X, Wong GK, Wang J. SOAPdenovo-Trans: de novo transcriptome assembly with short RNASeq reads. *Bioinformatics*. 2014;30(12):1660–1666. DOI: 10.1093/bioinformatics/btu077

[81] Luo R, Liu B, Xie Y, Li Z, Huang W et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 2012;1:18. DOI: 10.1186/2047-217X-1-18

[82] Chopra R, Burow G, Farmer A, Mudge J, Simpson CE, Burow MD. Comparisons of de novo transcriptome assemblers in diploid and polyploid species using peanut (Arachis spp.) RNA-Seq data. *PLoS ONE*. 2014;9(12):e115055. DOI: 10.1371/journal.pone.0115055

[83] Haznedaroglu BZ, Reeves D, Rismani-Yazdi H, Peccia J. Optimization of de novo transcriptome assembly from high-throughput short read sequencing data improves functional annotation for non-model organisms. *BMC Bioinformatics*. 2012;13:170. DOI: 10.1186/1471-2105-13-170

[84] Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072–1075. DOI: 10.1093/bioinformatics/btt086

[85] O'Neil ST, Emrich SJ. Assessing de novo transcriptome assembly metrics for consistency and utility. *BMC Genomics*. 2013;14:465. DOI: 10.1186/1471-2164-14-465

[86] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2009;26(1):139–140. DOI: 10.1093/bioinformatics/btp616

[87] Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNASeq gene expression estimation with read mapping uncertainty. *Bioinformatics*. 2009;26(4):493–500. DOI: 10.1093/bioinformatics/btp692

[88] Oshlack A, Wakefield MJ. Transcript length bias in RNASeq data confounds systems biology. *Biol Direct*. 2009;4:14. DOI:10.1186/1745-6150-4-14

[89] Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J, Stephens M, Gilad Y, Pritchard JK. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*. 2010;464(7289):768–772. DOI: 10.1038/nature08872

[90] Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. *Natur Rev Genet*. 2010;11(10):733–739. DOI: 10.1038/nrg2825

[91] Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNASeq experiments. *BMC Bioinformatics*. 2010;11: 94. DOI: 10.1186/1471-2105-11-94

[92] Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNASeq. *Natur Biotechnol*. 2013;31(1):46–53. DOI: 10.1038/nbt.2450

[93] Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003;19(2):185–193. DOI: 10.1093/bioinformatics/19.2.185

[94] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Symth GK. limma powers differential expression analyses for RNASequencing and microarray studies. *Nucl Acids Res*. 2015;43(7):e47. DOI: 10.1093/nar/gkv007

[95] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNASeq data. *Genome Biol*. 2010;11(3):R25. DOI: 10.1186/gb-2010-11-3-r25

[96] Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11:R106. DOI: 10.1186/gb-2010-11-10-r106

[97] Love MI, Huber W, Anders S. Moderated estimation fold change and dispersion for RNASeq data with DESeq2. *Genome Biol*. 2014;15:550. DOI: 10.1186/s13059-014-0550-8

[98] Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloe D, Gall CL, Schaeffer B, Crom SL, Guedj M, Jaffrezic F. A comprehensive evaluation of normali-

zation methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinformatics*. 2013;14(6):671–683. DOI: 10.1093/bib/bbs046

[99] Zyprych-Walczak J, Szabelska Handschuh L, Gorczak K, Klamecka K, Figlerowicz M, Siat-kowski I. The impact of normalization methods on RNASeq data analysis. *Biomed Res Int*. 2015;2015:621690. DOI: 10.1155/2015/621690

[100] Leek JT. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucl Acids Res*. 2014:42(21):e161. DOI: 10.1093/nar/gku864

[101] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNASeq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008;18(9):1509–1517. DOI: 10.1101/gr.079558.108

[102] Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNASeq data. *Bioinformatics*. 2010;26(1):136–138. DOI: 10.1093/bioinformatics/btp612

[103] Auer PL, Doerge RW. A two-stage Poisson model for testing RNASeq data. *Stat Appl Genet Mol Biol*. 2011;10(1):1–26. DOI: 10.2202/1544-6115.1627

[104] Hardcastle TJ, Kelly KA. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*. 2010;11:422. DOI: 10.1186/1471-2105-11-422

[105] Chen HH, Liu Y, Zou Y, Zhao L, Sarkar D, Huang Y, Chen Y. Differential expression analysis of RNA sequencing data by incorporating non-exonic mapped reads. *BMC Genomics*. 2015;16(Suppl No. 7):S14. DOI: 10.1186/1471-2164-16-S7-S14

[106] Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-se data. *BMC Bioinformatics*. 2013;14:91. DOI: 10.1186/1471-2105-14-91

[107] Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNASeq studies. *Brief Bioinform*. 2015;16(1):59–70. DOI: 10.1093/bib/bbt086

[108] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *JMB*. 1990;215(3):403–410. DOI: 10.1016/S0022-2836(05)80360-2

[109] Ashburner M, Ball CA, Blake JA, Bostein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Natur Genet*. 2000;25(1): 25–29. DOI: 10.1038/75556

[110] Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucl Acids Res*. 2004;32(Suppl 1):D277–D280. DOI: 10.1093/nar/gkh063

[111] Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schimdt E, Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. Reactome: a knowledge

base of biological pathways. *Nucl Acids Res*. 2005;33(Suppl. 1):D428–D432. DOI: 10.1093/nar/gki072

[112] Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*. 2012;8(2):e1002375. DOI: 10.1371/journal.pcbi.1002375

[113] Khatri P, Draghici S, Ostemeier GC, Krawetz SA. Profiling gene expression using onto-express. *Genomics*. 2002;79(2):266–270. DOI: 10.1006/geno.2002.6698

[114] Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conkin BR. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Natur Genet*. 2002;31:19–20. DOI: 10.1038/ng0502-19

[115] Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*. 2003;4(4):R28. DOI: 10.1186/gb-2003-4-4-r28

[116] Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*. 2003; 4(9):R60. DOI: 10.1186/gb-2003-4-9-r60

[117] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Natur Protocols*. 2008;4(1):44–57. DOI: 10.1038/nprot.2008.211

[118] Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*. 2004;20(1): 93–99. DOI: 10.1093/bioinformatics/btg382

[119] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene se enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–15550. DOI: 10.1073/pnas.0506580102

[120] Tarca AL, Draghici S, Kathri P, Hasan SS, Mittal P, Kim J, Kim CJ, Kusanovic JP, Romero R. A novel signaling pathway impact analysis. *Bioinformatics*. 2009;25(1):75–82. DOI: 10.1093/bioinformatics/btn577

[121] Shojaie A, Michailidis G. Analysis of gene sets based on the underlying regulatory network. *J Comput Biol*. 2009;16(3):407–426. DOI: 10.1089/cmb.2008.0081

[122] Roge X, Zhang X. RNAseqViewer: visualization tool for RNA-Seq data. *Bioinformatics*. 2014 ;30(6):891–892. DOI: 10.1093/bioinformatics/btt649

[123] Nussbaumer T, Kugler KG, Bader KC, Sharma S, Seidel M, Mayer KF. RNASeqExpressionBrowser–a web interface to browse and visualize high-throughput expression data. *Bioinformatics*. 2014;30(17):2519–2520. DOI: 10.1093/bioinformatics/btu334

[124] An J, Lai J, Wood DL, Sajjanhar A, Wang C, Tevz G, Lehman ML, Nelson CC. RNA-SeqBrowser: a genome browser for simultaneous visualization of raw strand specific RNAseq reads and UCSC genome browser custom tracks. *BMC Genomics*. 2015;16:145. DOI: 10.1186/s12864-015-1346-2

[125] Kinsella M, Harismendy O, Nakano M, Frazer KA, Bafna V. Sensitive gene fusion detection using ambiguously mapping RNASeq read pairs. *Bioinformatics*. 2011;27(8): 1068–1075. DOI: 10.1093/bioinformatics/btr085

[126] McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G et al. deFuse: an algorithm for gene fusion discovery in tumor RNASeq data. *PLoS Comput Biol*. 2011;7(5):e1001138. DOI: 10.1371/journal.pcbi.1001138

[127] Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNASeq data. *BMC Bioinformatics*. 2011;12:480. DOI: 10.1186/1471-2105-12-480

[128] Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucl Acids Res*. 2012;40: e72. DOI: 10.1093/nar/gks001

[129] Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. Efficient experimental design and analysis strategies for the detection of differential expression using RNA sequencing. *BMC Genomics*. 2012;13:484. DOI: 10.1186/1471-2164-13-484

[130] Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. Comprehensive evaluation of differential gene expression analysis methods for RNASeq data. *Genome Biol*. 2013;14(9):R95. DOI: 10.1186/gb-2013-14-9-r95

[131] Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNASeq. *Nat Methods*. 2011;8:469–477. DOI: 10.1038/nmeth.1613

[132] Lu B, Zeng Z, Shi T. Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNASeq. *Sci China Life Sci*. 2013;56(2):143–155. DOI: 10.1007/s11427-013-4442-z

[133] Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics*. 2012;13:734. DOI: 10.1186/1471-2164-13-734

[134] Hu L, Hao C, Fan R, Wu B, Tan L, Wu H. De novo assembly and characterization of fruit transcriptome in black pepper (*Piper nigrum*). *PLoS ONE*. 2015;10(6):e0129822. DOI: 10.1371/journal.pone.0129822

[135] Sudheesh S, Sawbridge TI, Cogan NO, Kennedy P, Forster JW, Kaur S. De novo assembly and characterisation of the field pea transcriptome using RNA-Seq. *BMC Genomics*. 2015;16(1):611. DOI: 10.1186/s12864-015-1815-7

[136] Talhinhas P, Azinheira HG, Vieira B, Loureiro A, Tavares S, Batista D, Morin E, Petitot AS, Paulo OS, Poulain J, Da Silva C, Duplessis S, Silva Mdo C, Fernandez D. Overview of the functional virulent genome of the coffee leaf rust pathogen *Hemileia*

*vastatrix* with an emphasis on early stages of infection. *Front Plant Sci*. 2014;5:88. DOI: 10.3389/fpls.2014.00088

[137] Yang X, Hu H, Yu D, Sun Z, He X, Zhang J, Chen Q, Tian R, Fan J. Candidate resistant genes of sand pear (*Pyrus pyrifolia* Nakai) to *Alternaria alternata* revealed by transcriptome sequencing. *PLoS ONE*. 2015;10(8):e0135046. DOI: 10.1371/journal.pone.0135046

[138] Ropka-Molik K, Zukowski K, Eckert R, Gurgul A, Piorkowska K, Oczkowicz M. Comprehensive analysis of the whole transcriptomes from two different pig breeds using RNA-Seq method. *Anim Genet*. 2014;45(5):674–684. DOI: 10.1111/age.12184

[139] Cui X, Hou Y, Yang S, Xie Y, Zhang S, Zhang Y, Zhang Q, Lu X, Liu GE, Sun D. Transcriptional profiling of mammary gland in Holstein cows with extremely different milk protein and fat percentage using RNA sequencing. *BMC Genomics*. 2014;15:226. DOI: 10.1186/1471-2164-15-226

[140] Sandri M, Stefanon B, Loor JJ. Transcriptome profiles of whole blood in Italian Holstein and Italian Simmental lactating cows diverging for genetic merit for milk protein. *J Dairy Sci*. 2015;98(9):6119–6127.DOI: 10.3168/jds.2014-9049

[141] Wall EH, Bond JP, McFadden TB. Milk yield responses to changes in milking frequency during early lactation are associated with coordinated and persistent changes in mammary gene expression. *BMC Genomics*. 2003;14:296. DOI: 10.1186/1471-2164-14-296

[142] Liu P, Wang Y, Du X, Yao L, Li F, Meng Z. Transcriptome analysis of thermal parthenogenesis of the domesticated silkworm. *PLoS ONE*. 2015;10(8):e0135215. DOI: 10.1371/journal.pone.0135215

[143] Liu R, Wang X, Chen GY, Dalerba P, Gurney A, Hoey T, Sherlock G, Lewicki J, Shedden K, Clarke MF. The prognostic role of a gene signature from tumorigenic breast-cancer cells. *N Engl J Med*. 2007;356(3):217–226. DOI: 10.1056/NEJMoa063994

[144] van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415(6871):530–536. DOI: 10.1038/415530a

[145] van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*. 2002;347(25):1999–2009. DOI: 10.1056/NEJMoa021967

[146] Huang L, Zheng M, Zhou QM, Zhang MY, Yu YH, Yun JP, Wang HY. Identification of a 7-gene signature that predicts relapse and survival for early stage patients with cervical carcinoma. *Med Oncol*. 2012;29(4):2911–2918. DOI: 10.1007/s12032-012-0166-3

[147] Hernández-Prieto S, Romera A, Ferrer M, Subiza JL, López-Asenjo JA, Jarabo JR, Gómez AM, Molina EM, Puente J, González-Larriba JL, Hernando F, Pérez-Villamil B, Díaz-Rubio E, Sanz-Ortega J. A 50-gene signature is a novel scoring system for tu-

mor-infiltrating immune cells with strong correlation with clinical outcome of stage I/II non-small cell lung cancer. *Clin Transl Oncol*. 2015;17(4):330–338. DOI: 10.1007/s12094-014-1235-1

[148] Peng L, Bian XW, Li DK, Xu C, Wang GM, Xia QY, Xiong Q. Large-scale RNA-seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 TCGA cancer types. *Sci Rep*. 2015;5:13413. DOI: 10.1038/srep13413

[149] Koh Y, Park I, Sun CH, Lee S, Yun H, Park CK, Park SH, Park JK, Lee SH. Detection of a distinctive genomic signature in rhabdoid glioblastoma, a rare disease entity identified by whole exome sequencing and whole transcriptome sequencing. *Transl Oncol*. 2015;8(4):279–287. DOI: 10.1016/j.tranon.2015.05.003.

[150] Punt S, Corver WE, van der Zeeuw SA, Kielbasa SM, Osse EM, Buermans HP, de Kroon CD, Jordanova ES, Gorter A. Whole-transcriptome analysis of flow-sorted cervical cancer samples reveals that B cell expressed TCL1A is correlated with improved survival. *Oncotarget*. 2015. DOI: 10.18632/oncotarget.4526

[151] Yu B, Russanova V, Gravina S, Hartley S, Mullikin JC, Ignezweski A, Graham J, Segars JH, DeCherney AH, Howard BH. DNA methylome and transcriptome sequencing in human ovarian granulosa cells links age-related changes in gene expression to gene body methylation and 3′-end GC density. *Oncotarget*. 2015;6(6):3627–3643. DOI: 10.18632/oncotarget.2875

[152] Zhang R, Yu C, Wu R, Zhang L, Zhu L, Xu A, Wang C. RNA-seq-based transcriptome analysis of changes in gene expression linked to human pregnancy outcome after in vitro fertilization-embryo transfer. *Reprod Sci*. 2015;pii: 1933719115597766. DOI: 10.1177/1933719115597766

[153] Jalbrzikowski M, Lazaro MT, Gao F, Huang A, Chow C, Geschwind DH, Coppola G, Bearden CE. Transcriptome profiling of peripheral blood in 22q11.2 deletion syndrome reveals functional pathways related to psychosis and autism spectrum disorder. *PLoS ONE*. 2015;10(7):e0132542. DOI: 10.1371/journal.pone.0132542

[154] Mariani J, Coppola G, Zhang P, Abyzov A, Provini L, Tomasini L, Amenduni M, Szekely A, Palejev D, Wilson M, Gerstein M, Grigorenko EL, Chawarska K, Pelphrey KA, Howe JR, Vaccarino FM. FOXG1-dependent dysregulation of GABA/glutamate neuron differentiation in autism spectrum disorders. *Cell*. 2015;162(2):375–390. DOI: 10.1016/j.cell.2015.06.034

[155] Pecorelli A, Leoni G, Cervellati F, Canali R, Signorini C, Leoncini S, Cortelazzo A, De Felice C, Ciccoli L, Hayek J, Valacchi G. Genes related to mitochondrial functions, protein degradation, and chromatin folding are differentially expressed in lymphomonocytes of Rett syndrome patients. *Mediators Inflamm*. 2013;2013:137629. DOI: 10.1155/2013/137629

[156] Gosalbes MJ, Durban A, Pignatelli M, Abellan JJ, Jimenez-Hernandez N, Perez-Cobas AE, Latorre A, Moya A. Metatranscriptomic approach to analyze the functional human gut microbiota. *PLoS ONE*. 2011;6(3):e17447. DOI: 10.1371/journal.pone.0017447

[157] Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P, Joint I. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE*. 2008;3(8):e3042. DOI: 10.1371/journal.pone.0003042

[158] Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, Delong EF. Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci U S A*. 2008;105(10):3805–3810. DOI: 10.1073/pnas.0708897105

[159] Urich T, Lanzen A, Qi J, Huson DH, Schleper C, Schuster SC. Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE*. 2008;3(6):e2527. DOI: 10.1371/journal.pone.0002527

[160] Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet*. 2013;14:618–630. DOI: 10.1038/nrg3542

[161] Saliba AE, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. *Nucl Acids Res*. 2014;42(14):8845–8860. DOI: 10.1093/nar/gku555

[162] Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, Ueda HR. Quartz-Seq: a highly reproducible and sensitive single-cell RNA-Seq reveals non-genetic gene expression heterogeneity. *Genome Biol*. 2013;14(4):R31. DOI: 10.1186/gb-2013-14-4-r31

[163] Grun D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*. 2015;525:151–255. DOI: 10.1038/nature14966

[164] Londin E, Loher P, Telonis AG, Quann K, Clark P, Jing Y et al. Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proc Natl Acad Sci U S A*. 2015;112(10):E1106–11015. DOI: 10.1073/pnas.1420955112

[165] Cieslik M, Chugh R, Wu YM, Wu M, Brennan C, Lonigro R, Su F, Wang R, Siddiqui J, Mehra R, Cao X, Lucas D, Chinnaiyan AM, Robinson D. The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. *Genome Res*. 2015;25(9):1372–1381. DOI: 10.1101/gr.189621.115

[166] Taylor RC. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics*. 2010;11(Suppl 12):S1. DOI: 10.1186/1471-2105-11-S12-S1

[167] Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP. Large-scale sequencing reveals 21U–RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell*. 2006;127(6):1193–1207. DOI: 10.1016/j.cell.2006.10.040

[168] Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*. 2008;322(5909):1845–1848. 10.1126/science.1162228

[169] Chu C, Qu K, Zhong FL, Artandi SE, Chang HY. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell*. 2011;44(4):667–678. DOI: 10.1016/j.molcel.2011.08.027

[170] Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*. 2009;324(5924):218–223. DOI: 10.1126/science.1168978

[171] Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*. 2009;460:479–486. DOI: 10.1038/nature08170

[172] Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T et al. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*. 2008;44(1):3–12. DOI: 10.1016/j.ymeth.2007.09.009

[173] Churchman LS, Weissman JS. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*. 2011;469:368–373. DOI: 10.1038/nature09652

[174] Jiao Y, Meyerowitz EM. Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control. *Mol Syst Biol*. 2010;6:419. DOI: 10.1038/msb.2010.76

[175] German MA, Pillay M, Jeong DH, Hetawal A, Luo S et al. Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotechnol*. 2008;26(8):941–946. DOI: 10.1038/nbt1417

[176] Pelechano V, Wei W, Steinmetz LM. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*. 2013;497(7447):127–131. DOI: 10.1038/nature12121

[177] Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep*. 2012;2(3):666–673. DOI: 10.1016/j.celrep.2012.08.003

[178] Peng Y, Leung HC, Yiu SM, Lv MJ, Zhu XG, Chin FY. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics*. 2013;29(13):i326–334. DOI: 10.1093/bioinformatics/btt219

[179] Chu HT, Hsiao WW, Chen JC, Yeh TJ, Tsai MH, Lin H, Liu YW, Lee SA, Chen CC, Tsao TT, Kao CY. EBARDenovo: highly accurate de novo assembly of RNASeq with efficient chimera-detection. *Bioinformatics*. 2013;29(8):1004–1010. DOI: 10.1093/bioinformatics/btt092

[180] Maretty L, Sibbesen JA, Krogh A. Bayesian transcriptome assembly. *Genome Biol*. 2014;15(10):501. DOI: 10.1186/s13059-014-0501-4

IntechOpen