

# RNA-seq for revealing the function of the transcriptome

Kavita Goswami and Neeti Sanan-Mishra

*Plant RNAi Biology Group, International Centre for Genetic Engineering and Biotechnology, New Delhi, India*

## 7.1 Introduction

The discovery of RNA as a primary intermediate between the genome and the proteome improved the understanding of key cellular activities and led to the development of techniques related to the recognition of transcripts and the quantification of their expression at the molecular level (Conesa et al., 2016). This greatly boosted the research related to deciphering and understanding the transcriptional structure of a cell. Transcriptome defines the entire set of transcripts in a cell including the protein-coding messenger RNAs (mRNAs), the noncoding structural tRNAs and rRNAs, and the noncoding regulatory small RNAs, such as microRNAs (miRNAs) (Studholme, 2012). The primary objective of transcriptomics is to elucidate the functional constituents of the genome during development and even under various conditions of disease or stress (Wang, Gerstein, & Snyder, 2009).

The original purpose of RNA sequencing (RNA-seq) was to establish the spectrum of genomic loci that are expressed in a cell (population) and quantitating the whole expression at a specific time without the absolute need to predefine their sequences (Liu, Zhou, & White, 2014; Robinson & Oshlack, 2010). Earlier, Sanger sequencing also known as chain termination method was the most extensively used RNA/complementary DNA (cDNA) sequencing method that involved sequencing of single reads. As technology developed, it was substituted by next-generation sequencing (NGS) techniques, which brought a revolution in genomic research (Morozova & Marra, 2008). This method facilitated massive parallel sequencing for millions of DNA strands at the same time (Arsenic et al., 2015; Behjati & Tarpey, 2013). This is not only amplified the speed and accuracy of sequencing but greatly reduced the cost and time (Buermans & Den Dunnen, 2014).

Prior to the discovery of NGS, it was recognized that only ~2%–5% of genomic DNA is translated into protein and rest of the DNA that did not code for any protein was considered as junk or nonfunctional. After 1990s, insights provided by NGS technology facilitated the analysis of the noncoding sequences and the sustained understanding of their crucial role in controlling gene regulation (Barrett, Fletcher, & Wilton, 2013; Ragupathy, You, & Cloutier, 2013). NGS also provided insights into the complexity of the transcriptome by identifying the existence of several kinds of transcript variants and genome-wide alterations in their expression (Trapnell et al., 2013).

Advances in NGS have opened specific applications, such as *de novo* sequencing, methylation sequencing, metagenomics, resequencing, and transcriptome sequencing. RNA-seq comprises of a variety of applications for both coding and noncoding transcripts (Ingolia, Brar, Rouskin, Mcgeachy, & Weissman, 2012). It also provides a quantitative method to assess the expression level of RNA. In addition, it can be used for detecting differential gene expression, profiling of small RNA and other noncoding RNAs (ncRNAs). It also helps in characterizing the alternate splicing patterns and identifying variants or alleles. The fine information offered by sequencing-based transcriptome surveys indicates that RNA-seq is likely to become the preferred medium for steady-state examination of gene expression in any organism (Robinson & Oshlack, 2010).

## 7.2 Next-generation sequencing platforms and their technologies

The four major sequencing giants whose platforms are being widely used as second-generation sequencers are Illumina, Roche 454, Ion torrent, and Sequencing by Oligonucleotide Ligation and Detection (SOLiD). The salient features for each of these are provided below and a comparison between them is presented in Table 7.1.

**TABLE 7.1** Comparison of commonly used second-generation sequencing platforms.

Characteristics	Illumina/HiSeq	Ion torrent	SOLiD ABI	Roche 454
Sequencing principle	Sequencing by synthesis	Sequencing by synthesis	Ligation and two base coding	Pyrosequencing
Instrument cost	\$690,000	\$80,000	\$495,000	\$500,000
Sequencing cost per million base	\$0.07 [\$6000/(30 × ) human genome]	\$5	\$0.13 [\$15,000/100 Gb]	\$10 [\$7000 per run]
Technology	Polonies cleavable dye terminators	emPCR, H + detection	emPCR ligation with cleavable dye terminators	emPCR pyrosequencing
Run time (days per run)	5–6	2–4	7	0.95
Accuracy (%)	99.9	99	99.94	99
Memory (Gb)	48	10	16	48
Read length	100–150	150–200	35–75	450–900
Output range (Gb)	600	1–10	120	0.7
Primary errors	Substitution	Indel	AT bias	Indel
Error rate	0.26	1.71	0.01	0.1
Capacity for paired end	Yes	No	Yes	Yes
Advantage	High throughput	Low cost, fast, optical detection not needed	Accuracy	Read length, fast
Disadvantage	Short read assembly	Higher error rate in case of long reads	Short read assembly	Error rate with polybase >6, high cost, low throughput

### 7.2.1 Illumina

Illumina provides one of the most common and widely used NGS platforms responsible for producing more than 90% of the sequencing data (Quail et al., 2008). It started by acquiring Genome Analyzer (GA) from Solexa and now offers a range of models or sequencing platforms, such as HiSeq 2000/2500, GAIIx, X Ten, NextSeq 500, and MiSeq. The ease of sequencing allows researchers to explore the genome, transcriptome, or epigenome and hence caters to a broad range of applications (Morozova & Marra, 2008). Their method is suitable for sequencing both single-read and paired-end (from both 3' and 5' ends) fragments comprising short and long inserts. The sequencing protocol includes bridge amplification to create clones of each fragment, which expands each pool for sequence by synthesis. The sequencing principle is based on the incorporation of fluorescently tagged nucleotides, each having a unique emission, into the newly synthesized DNA strand. It is capable of sequencing read lengths of 100–150 bases, with a combined regular throughput of 360–500 Gb, which is the largest among the four main sequencing techniques (Liu et al., 2012). Overall, it is a high-performance sequencing system with relatively low operating costs on instruments and other consumables. However, it has a long run time and may take 5–6 days per run with a 48-Gb memory.

### 7.2.2 Roche 454

Roche 454 was the first successful pyrosequencing platform developed and advocated by 454 Life Sciences in the year 2005. Roche acquired it in 2007 but their support for this platform was stopped in 2016. Their method used a spray to break DNA into fragments of 300–800 bp. Different adapters were added at both ends of the DNA and the reaction mix was subjected to emulsion PCR (polymerase chain reaction). Each drop of water contained only one DNA template, a bead coated with oligos complimentary to the adapter, and reagents for synthesis. The technology involved

quantitating the pyrophosphate release during incorporation of nucleotide into the newly synthesized DNA. The base calling and quantification is centered on detecting and measuring proportionate amount of fluorescent signals generated through chemical reactions requiring ATP sulfurylase and luciferase (Jeanneau, Bon, & Martin, 2010). The system was massively parallelized using picotiter plates to produce more than 200,000 reads at 100–150 bp per read with an output of 20 Mb per run in 2005. The upgraded 454 GS FLX Titanium system released by Roche in 2008 improved the average read length to 700 bp with an accuracy of 99.997% and an output of 0.7 Gb of data per run within 24 h (Liu et al., 2012). This technique was suitable for sequencing long reads with a length range of 450–900 bases per read. It was fast when compared to other NGS methods and had low capital cost and cost per experiment. The major disadvantage was its low throughput and high cost per Mb of data. The sequencing reads were error prone in repeats of more than six bases.

### 7.2.3 ION torrent

ION torrent is a product of Thermo Fisher Scientific and is based on Ion-Semiconductor Sequencing. It is a DNA sequencing technique based on the detection of hydrogen ions released into the growing DNA template during the incorporation of new nucleotides (Ribani et al., 2018). In ion-semiconductor sequencing chips, hydrogen ions are detected so that there is a clear correlation between chemical and digital events. Thus bases are called without scanning, camera and light. The four devices marketed by the company are Ion Proton, Ion Personal Genome Machine (PGM), Ion S5, and Ion S5 XL system (Wang, Chen, et al., 2018). The PGM sequencer uses disposable chips containing highly compact array of micro-sized wells for performing massively parallel reactions for nucleotide incorporation. Every well has an individual DNA template and an ion-sensitive layer followed by a proprietary ion-sensor below the well, which recognizes the ions as changes in the solution's pH. The output voltage is doubled if there are two identical bases on the DNA strand, and the chip reports two identical bases. As there is no optical detection needed, the ION Torrent sequence technique is very fast. It can sequence 150–200 bases per read with a regular throughput of 1–10 Gb, but the sequences are prone to errors of up to 1.7%, especially in the case of long reads.

### 7.2.4 SoLid ABI

SoLid ABI involves SOLiD. This NGS technique was developed by Life Technologies and acquired by Applied Biosystems in 2006. It utilizes a fluorescent probe containing a unique pair of two bases on its 3' end (Chi, 2008). Each unique base is assigned one out of four possible colors such that there are 16 unique pairs, such as AA, AC, and AU. The sequencing chemistry requires production of a clonal bead population from a library of DNA fragments. Individual fragments are attached to the magnetic beads using a universal adapter sequence and beads are then covalently bound to a glass slide. Thus only one species of fragment will be present on the surface of each magnetic bead but the starting sequence of every fragment is known and identical. The sequencing is based on the principle of emulsion PCR and primed by oligos complementary to the adapter sequence. A set of di-base probes compete for ligation to the sequencing primer, and their incorporation is governed by the template sequence. Specificity of the di-base probe is achieved by interrogating every first and second bases in each ligation reaction. The sequence extension involves multiple cycles of ligation, detection, and cleavage. The template is reset with the primer complementary to the  $n - 1$  location for a second round of ligation cycles at the end of the extension. For each sequence tag, five rounds of primer reset are completed (Chi, 2008). Through the primer reset process, each base is interrogated in two independent ligation reactions by two different primers. This effectively improves the specificity of the template strand and offers 99.94% accuracy. The chemistry is not hindered by the presence of homopolymer repeat regions. The read length of SOLiD was initially 35-bp reads with an output of 3G data per run, and this was increased to 85-bp reads with an output of 30G data per run. However, the sequence of the fragment can be deduced after five rounds of sequencing only and requires heavy dependence on the computational algorithms. The high capital cost of data delivery makes it more suitable for resequencing or quick read assemblies.

Each of the sequencers described above has been employed in RNA-seq applications by converting RNA into its complementary DNA (cDNA), but they have limitations in read lengths. They are not efficient in detecting structural variants, repetitive elements, extreme guanine–cytosine (GC) content, or sequences with multiple homologous elements in the genome (Salzberg & Yorke, 2005). These drawbacks have strongly driven the search for other methods of sequencing.

With the advent of single molecular and real-time sequencing technology, third-generation sequencing (TGS) was recognized. The three commercially available sequencing platforms in this category include Pacific Biosciences

**TABLE 7.2** Comparison of third-generation sequencing platforms.

Characteristics	Illumina (Tru-seq)	PacBio	Oxford Nanopore
Read length (kb)	1.5–18.5	10–15	Up to 200
Processing time	2–3 days	3–4 h	2 days
Error rate (%)	0.03	10–15	5–15
Cost per run	~\$2500	\$400	\$900
Sequencer	True-seq	RS II	MinION
Advantages	High accurate	Sequence long reads	Sequence long reads, portable device

(PacBio) Single Molecule Real-Time (SMRT) sequencer, Illumina Tru-seq with Synthetic Long-Read technology (SLR), and Oxford (MinION) with Nanopore Technology (Lee et al., 2016). Technologies of the third generation do not break down DNA or amplify it. They sequence a single molecule of DNA directly. A comparison of TGS is provided in Table 7.2. These technologies are mainly used to generate synthetic long reads for *de novo* assembly and genome finishing applications, sequencing traditionally challenging genomes, and perform whole human genome phasing to identify coinherited alleles, haplotype information and phase *de novo* mutations. It is useful for generating full-length RNA or cDNA sequences and for detecting isoforms. A brief overview of the TGS sequencers is provided below.

### 7.2.5 Illumina Tru-seq SLR technology

Illumina Tru-seq SLR technology employs specific library prep and barcode kits to prepare DNA libraries. The technology is not compatible for RNA yet. It fragments DNA into large fragments of approximately 10 kb and marks them with a unique barcode. They are sequenced on HiSeq 2500 or HiSeq 2000 platforms for subsequent assembly into synthetic long reads or whole genome phasing using proprietary informatics. By reducing the need for additional specialized equipment, this method allows researchers to access useful long-read information. Thus at relatively low costs it offers greater insight into the genome.

### 7.2.6 Pacific Biosciences (PacBio) SMRT sequencing

PacBio was commercially launched in 2010 and not based on PCR amplification. This platform captures sequence information during the replication process of the target DNA molecule. It adopts real-time SMRT to monitor fluorescently tagged nucleotides optically as they are incorporated into individual molecules of the template. The template, called a SMRTbell, is a closed, circular single-stranded circular DNA created by ligating hairpin adaptors to both ends of a target double-stranded DNA molecule (Lee et al., 2016; Xiao & Zhou, 2020). The SMRTbell diffuses into a Zero-mode waveguide (ZMW) and the adaptor binds to a polymerase immobilized at the bottom. Four types of nucleotides G, C, T, and A are labeled with red, yellow, green, and blue fluorescent dyes, respectively, to create distinct emission spectrums. The fluorescence output of the color corresponding to the incorporated base (yellow for base C as an example shows here) is elevated by the ZMW. As the dye linker-pyrophosphate product is cleaved from the nucleotide, it diffuses out of the ZMW to end the fluorescence pulse. The current instrument, the PacBio RS II, generates read lengths of up to ~100,000 bp with the maximum throughput [ $\sim 8$  Gb/day of the long-read technologies currently available (Lee et al., 2016)]. PacBio provides longer read lengths than second-generation sequencing (SGS) technologies, making it ideal for genome, transcriptome, and epigenetics studies (Rhoads & Au, 2015).

### 7.2.7 Oxford Nanopore

Nanopore-based devices measure changes in ionic current as biological molecules pass through nanopores or near it (Lee et al., 2016). The information about the change in current can be used to identify that molecule. The concept of nanopore sequencing was envisioned in the early 1990s by David Deamer (UC Santa Cruz) and Daniel Branton (Harvard) and sequencers were commercially launched in 2014. MinION is a pocket-sized, powerful and portable sequencing device for delivering high volumes of long-read sequence data. It was widely used for sequencing

bacterial and viral genomes (Quick et al., 2016). The benchtop GridION Mk1 can run up to five MinION Flow Cells at a time, on-demand, for larger genomics projects. PromethION is the largest format for nanopore sequencing, designed to offer on-demand use of up to 48 Flow Cells—capable of delivering more than 7 Tb of sequence data in a full run, and now is being used in population-scale sequencing projects.

### 7.3 Analyzing the RNA-seq data

The NGS technology, irrespective of the platform used, generates large volume of data that has added new challenges for research. Analysis of this data to obtain meaningful results requires stepwise processing and analysis (Fig. 7.1). The following parameters are essential to improve the accuracy of interpreting the results.

#### 7.3.1 Quality and depth of raw sequencing data

Quality control (QC) of the data emerging from high-throughput sequencers is one of the major challenges. It is critical to conduct QC protocols at various stages of data processing for ensuring correct and effective analysis. Data quality plays a crucial role, particularly in studies involving sequence assembly and gene expression. Quality filtering is carried out in the initial phase to eliminate low-quality reads. Various tools are freely available for quality filtering, such as Fastq\_clean (Zhang et al., 2014) and NGS toolkit (Patel & Jain, 2012). The various parameters used for QC are as follows.

##### 7.3.1.1 Base calling accuracy

Base calling accuracy is the most common metric used to determine the accuracy of a sequencing platform. It indicates the possibility that the sequencer incorrectly calls a given base (Illumina, 2011). Phred quality score is used to indicate the measure of base quality in DNA sequencing. High consistency of a sequenced base is indicated by greater values of Phred. A Phred Score of 20 indicates the likelihood of finding 1 incorrect base call among 100 bases. In other words, the precision of the base call is 99%.  $Q$  scores are classified as a property that is associated logarithmically with the probabilities of base calling error ( $P$ )<sup>2</sup>.

$$Q = -10 \log_{10} P$$

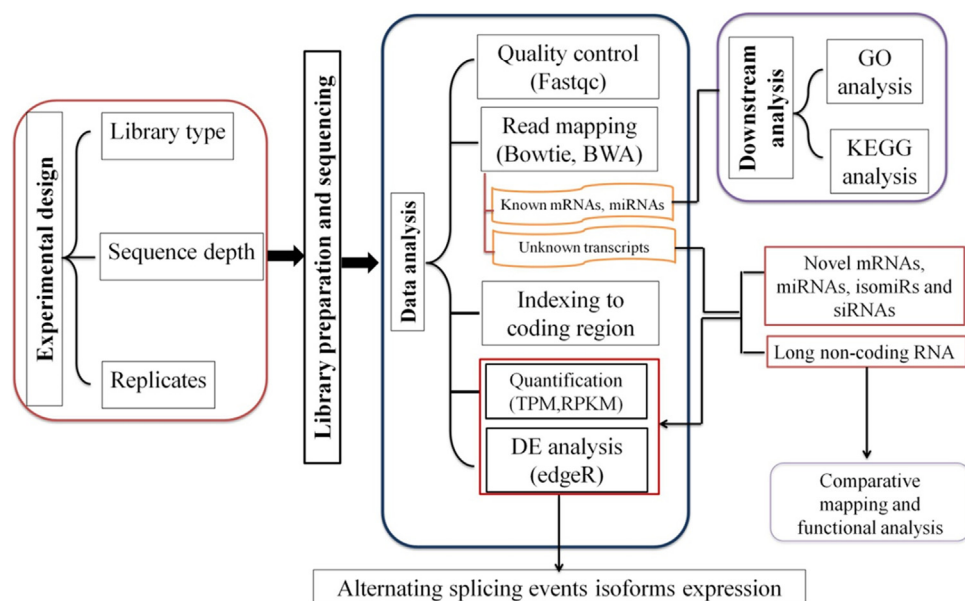


FIGURE 7.1 Pipeline used for RNA-seq data analysis.

### 7.3.1.2 GC content per sequence

GC content per sequence is plotted as a graph and it provides an indication on the contamination. This module analyses the GC content in a file over the entire length of each sequence and compares it to model GC content for normal distribution.

### 7.3.1.3 Adapter sequences

Adapter sequences are a central element of the NGS workflow and need to be removed. Adapter sequence can be specific to sequencer or it may be universal. The adapters contain binding sites of the sequencing primers, the index sequences, and the sites that allow the fragments to be attached to flow cell lawn. Sequencing reads generated by Illumina need adapter trimming only on the 3' ends of reads, since on the 5' ends adapter sequences are not found.

### 7.3.1.4 k-mer value

k-mer value refers to all the potential substrings of length k found in a string, which may imply low-quality or adapter sequence contamination. It is usually measured by checking whether there is a string of length k that occurs more than expected chance in the reads.

### 7.3.1.5 Overrepresented sequences

Overrepresented sequences, is a list of sequences that occur in the file more than expected. Normally, sequences are considered overrepresented if it accounts for 0.1 percent of the total reads. In an attempt to classify such sequences, each overrepresented sequence is compared to a list of common contaminants. The overrepresentation signal is primarily caused by a relatively large fraction of very poor-quality reads or likely by the presence of adaptor sequences.

## 7.3.2 Adapter removal

During sequencing, data are stored in FASTQ format and the sequence reads are contaminated with sequences of the adapters ligated at 5' and 3' ends. Therefore adapter trimming is essential to filter out the transcriptome reads. The NGS QC Toolkit is a helpful tool that provides both Illumina (IlluQC) and Roche 454 (454QC) separately with a strict quality management platform (Patel & Jain, 2012). In addition to adapter trimming and format conversion (FASTQ to FASTA), some other statistical analysis on sequencing data is facilitated. Various tools are available for adapter removal, such as Cutadapt (Martin, 2011), trimmomatic (Bolger, Lohse, & Usadel, 2014), and Adapter Removal (Lindgreen, 2012).

## 7.3.3 Level of alignment

Another important step of total RNA-seq data processing is aligning or mapping the reads to the reference genome and filter out the overlapped ncRNAs (namely rRNA, tRNA, snRNA, snoRNA, long intergenic ncRNA) and coding genes by annotating to exons. Transcriptome assembly is a powerful method for detecting alterations in gene expression and sequences for both model and nonmodel organisms between environments, tissues, or species (Martin & Wang, 2011; Voshall & Moriyama, 2018). This enables the transcriptome assemblers to reconstruct full-length mRNA or transcript reads. The transcriptome assembly can be genome-guided or *de novo* depending upon the availability of genome sequence. The genome-guided transcriptome assembly is more efficient and *de novo* assembly of putative transcripts from short-read sequences is normally performed if reference genome is not available or the available assembly is fragmented and of poor quality.

The format for files, storing read alignments to a reference sequence alignment, can be Sequence Alignment Map (SAM) or Binary Alignment Map (BAM). The SAM file is a tab-delimited text file containing details for alignment with the genome for each individual read. It has a simple format, consisting of a header and an alignment section. BAM file is a binary, compressed version of the SAM file. It can be indexed to allow information to be extracted easily. This allows the alignment viewers not to uncompress the entire BAM file to look at data for a specific coordinate set, anywhere in the file (Li et al., 2009).

### 7.3.4 Redundancy rate of reads

All practical applications are constrained by their accuracy due to genetic redundancy, presence of multiple alleles, and splice variants. Some of these challenges can be addressed by using a splice-ware sequencer aligner followed by resolving alternative splicing variants and exporting consensus transcript sequences (Martin & Wang, 2011; Trapnell et al., 2012).

## 7.4 RNA-seq applications

The key feature of RNA-seq analysis is to identify and analyze the cellular transcriptome. The high-throughput sequencing methods provide the techniques for obtaining a greater coverage to resolve the whole transcriptome including the protein-coding mRNAs, ncRNAs, such as miRNAs, small-interfering RNA (siRNAs), and long ncRNAs (lncRNAs), other RNAs, such as rRNA and tRNA, and different populations of nascent or preprocessed RNAs (Wang et al., 2009). It is also possible to identify novel transcripts, detect allele-specific expression (ASE) and characterize alternative splicing patterns and variations. This provides an important informative link between the genome and functional protein component. Recent developments in the RNA-seq have allowed researchers to elucidate the functional complexity of the transcription process, from sample preparation or library construction to data analysis. Some of the applications of RNA-seq can be summarized as below;

### 7.4.1 Transcriptome assembly

The advancements on RNA-seq and the associated software applications have changed the outlook on mapping the total transcriptome. Now gene expression studies are being initiated by a collection of assembled transcripts. This is particularly advantageous for non model organisms where a sequenced reference genome is not available, as *de novo* transcriptome assemblies have opened a way for experimental studies (Fig. 7.2) (Geniza & Jaiswal, 2017).

#### 7.4.1.1 Genome-guided transcriptome assembly

As more and more genomes are being sequenced, it is possible to use genome-guided assembly approaches to support the assembly process. A reference genome of a species can also be used to assist the assembly of any target species, also called closely related species. Normally this was employed to assemble similar genotypes within the same species (Lischer & Shimizu, 2017). The reference-guided approaches rely on the high level of similarity between two species

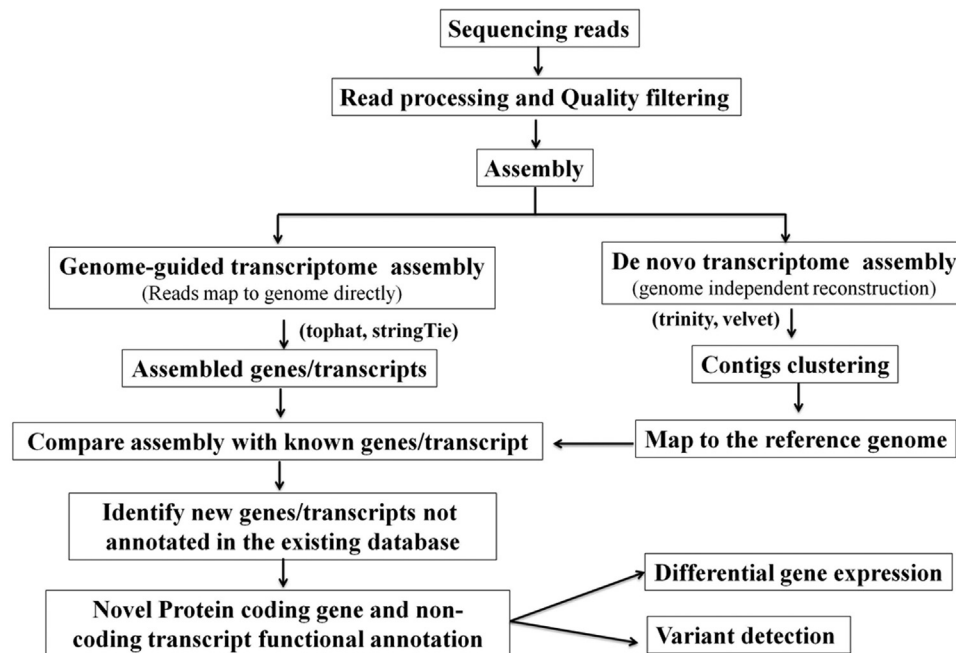


FIGURE 7.2 Flowchart to analyze RNA-seq data and identification of protein-coding genes.

for assembly of the transcriptome (Lischer & Shimizu, 2017; Pop, Phillippy, Delcher, & Salzberg, 2004; Schneeberger et al., 2011).

Two key strategies are employed for reference-guided assembly. As a first strategy, reads were mapped against the reference genome and then used to create an alternate consensus sequence of the target genome (Vezi, Cattonaro, & Policriti, 2011). In the second method, *de novo* assembly is performed and the resulting contigs/scaffolds are then aligned to chromosomes of the reference genome, to validate their annotation or identify possible misassembled contigs or scaffolds (Bao, Jiang, & Girke, 2014; Lischer & Shimizu, 2017; Vezi et al., 2011).

The mapping of transcriptome data to a reference genome has been greatly facilitated by the development of transcriptome assembly computer programs. They have improved the genome-guided assembly, by allowing the mapped reads to be split during alignment to account for presence of introns and other splicing events. It is now possible to align the first portion of the read to one location (an exon) and the other half to the downstream location (Voshall, 2018). Genome-guided assemblers, such as RefShannon (Mao, Pachter, Tse, & Kannan, 2020), StringTie (Pertea et al., 2015), Bayesemmer (Maretty, Sibbesen, & Krogh, 2014), Cufflinks (Ghosh & Chan, 2016), and Scripture (Guttman et al., 2010), typically use an alignment tool, such as TopHat2 (Kim et al., 2013), SpliceMap (Au, Jiang, Lin, Xing, & Wong, 2010), or GSNAP (Wu & Nacu, 2010), to map first short reads to a reference genome to cluster the reads into gene loci based on which the so-called splicing graph or overlap graph can be constructed for individual cluster.

#### 7.4.1.2 *De novo transcriptome assembly*

This method is typically more challenging and less accurate than genome-guided assembly. The *de novo* assembly of short DNA reads into contiguous full-length copies of RNA transcripts is a complex process since there are many sources of variance in the read coverage and read sequences. It is still helpful to produce *de novo* transcript assemblies for model plants, such as *Arabidopsis*, rice, and maize, to identify new transcript isoforms of existing annotated genes, alternate splicing events, and novel transcribed genes from a plant variety, or in response to a specific treatment. To analyze genetic variation, the *de novo* transcriptome assembly may be used to match sequence reads from the same or another sample (Fox et al., 2014; Geniza & Jaiswal, 2017).

Many *de novo* assemblers are available that generate contigs based solely on the RNA-seq data (Robertson et al., 2010). Some such algorithms or software packages are TransAByss (Chopra et al., 2014), Trinity (Grabherr et al., 2011), OASES (Schulz, Zerbino, Vingron, & Birney, 2012), SOAPdenovo-Trans (Xie et al., 2014), Trinity r2013-11-10 (Haas et al., 2013), SOAPdenovo-Trans v1.04 (Xie et al., 2014), and Velvet v1.2.10/Oases v0.2.08 (Schulz et al., 2012; Zerbino & Birney, 2008) and genome-guided assembly problems [Cufflinks (Trapnell et al., 2012) TransComb (Liu, Yu, Jiang, & Li, 2016), StringTie (Pertea et al., 2015) and CLASS2 (Song, Sabunciyan, & Florea, 2016), and Ryuto (Gatter & Stadler, 2018)]. Most assemblers depend on de Bruijn graphs generated in the RNA-seq data from shorter sequences of a given length  $k$  ( $k$ -mers) that are obtained after decomposition of long reads. The overlap of sequencing reads with these  $k$ -mer sequences helps to reconstruct the original sequence (Behera, Voshall, Deogun, & Moriyama, 2017). This need for a  $k$ -mer to start at a location along the initial sequence for the graph to cover the full sequence is also a significant limitation of the de Bruijn graphs. *De novo* transcriptome assemblies are available at GenBank (Benson et al., 2012) and the TreeGenes database (Wegrzyn, Lee, Tearse, & Neale, 2008).

### 7.4.2 Identification of novel protein-coding genes

Earlier, it was speculated that there were on an average 30,000–40,000 protein-coding genes in a plant genome and the remaining region was occupied by the so-called noncoding or junk DNA (Müller et al., 2016). The sequencing was restricted to cloned mRNAs and expressed sequence tag libraries that covered limited genes of a species (Fig. 7.2) (Garber, Grabherr, Guttman, & Trapnell, 2011; Tang, Lomsadze, & Borodovsky, 2015). RNA-seq significantly enhanced the capacity to sequence and identify all the protein-coding transcripts in a cell (Li, Zhang et al., 2011). The reference-based assembly also helps to discover new transcripts, including new isoforms unforeseen by automated genome annotation (Marchant et al., 2016).

A greater role was also played by the stepwise analysis of the sequenced datasets using various tools. To begin with, a large number of short reads were compiled into transcripts by mapping them to reference genome using bowtie. Redundant and highly similar novel genes aligning to other genomic regions were removed to exclude false positive results. This was followed by the assembly of the complete sequence using cufflink to compile the expressed transcripts and count their abundance. Upon gaining the confidence that RNA-seq technology provides reliable, unbiased, and quantifiable data of the transcriptome, the analysis was extended for the identification of novel protein-coding genes

**TABLE 7.3** List of predicted and identified genes from key plant genomes.

Organism	Genome size	Predicted genes	Protein-coding genes
<i>Arabidopsis thaliana</i>	135 Mb	55,398	27,655
<i>Oryza sativa</i>	420 Mb	45,772	37,960
<i>Triticum aestivum</i>	17 Gb	146,597	107,891
<i>Hordeum vulgare</i>	5.3 Gb	251,587	39,841
<i>Zea mays</i>	2.4 Gb	138,424	39,591
<i>Glycine max</i>	1.1–1.15 Gb	88,412	55,897
<i>Solanum lycopersicum</i>	950 Mb	35,825	34,658

(Steijger et al., 2013; Wang et al., 2009; Wilhelm & Landry, 2009). This led to the requirement of different tools and software to predict genes and analyze the vast sequencing data for their presence (Table 7.3). Ensembl is one such genome browser that provides whole information about the genome including coding genes (open reading frames), non-coding genes, evolutionary divergence, sequence variation, and transcriptional regulation (Cunningham et al., 2019).

### 7.4.3 Identification of other classes of RNAs

#### 7.4.3.1 Long noncoding RNAs

Transcripts with lengths exceeding 200 nucleotides but lacking protein-coding capabilities may exhibit regulatory cellular functions and are categorized as lncRNAs (Kung, Colognori, & Lee, 2013). lncRNAs mainly function at posttranscriptional levels and snuggle into the regulation of gene expression pathways mediated by miRNAs (Bischof & Martínez-Zamudio, 2015). However, growing evidence now indicates that lncRNAs are also involved at the transcriptional level, by serving as a decoy for traditional transcriptional factors. They may interact directly with the promoter region, base pair with the transcribed mRNA, or interact with transcription factor proteins (Kumar, Devaux, & Herschkowitz, 2016). By regulating nuclear architecture and transcription in the nucleus and by modulating mRNA stability, translation, and posttranslational changes in the cytoplasm, they emerge as important regulators in gene expression networks (Yao, Wang, & Chen, 2019). A number of computational methods are available, for example, Lncident (Han, Liang, Li, & Du, 2016), CNIT (Guo et al., 2019), LncFinder (Han et al., 2019), and PLEK (Li, Zhang, & Zhou, 2014) to facilitate the detection of lncRNAs from RNA-seq results (Hu & Zhao, 2015).

#### 7.4.3.2 Small noncoding RNAs

The advancement of deep sequencing technologies and the development of computational tools revolutionized the discovery of small RNA molecules. The datasets can now be used to identify and differentiate among small RNA species and also provide an understanding of their biological significance. Housekeeping RNAs, such as tRNAs, rRNAs, snRNAs, and snoRNAs, play structural and catalytic functions among the noncoding groups. The small regulatory RNAs in the size range of 19–25 nucleotides (nt) play significant roles in gene silencing and they can be categorized into specific families including siRNA, piwi RNA, and miRNA (Guleria, Mahajan, Bhardwaj, & Yadav, 2011).

##### 7.4.3.2.1 Endogenous siRNAs

Endogenous siRNAs are 21–25-nt-long duplex having 2-nt overhang at their 3' end. They act as a mediator of RNA silencing pathways at both transcriptional and posttranscriptional levels in the plant and in animals (Chau & Lee, 2007; Yu, Jih, Iglesias, & Moazed, 2014). siRNAs can be classified into different classes, including miRNA directed *trans*-acting siRNAs (ta-siRNAs)/*cis*-acting siRNAs, natural antisense siRNAs (nat-siRNAs), repeat-associated siRNAs, and heterochromatic siRNAs (hc-siRNAs). The various subclasses can be differentiated based on their origin and regulation (Chellappan et al., 2010).

##### 1. Heterochromatic small-interfering RNAs

hc-siRNAs are a 23–24-nt class of siRNAs that contains hc-siRNAs. They are formed from intergenic regions or genomic repeats, such as transposons by the action of RNA Polymerase IV (Pol IV). The RDR2 processes it into

a perfectly complemented duplex on which DCL3 further acts on (Chapman & Carrington, 2007; Matzke, Kanno, Daxinger, Huettel, & Matzke, 2009; Wang, Polydore, & Axtell, 2015). The hc-siRNAs help to maintain genome integrity through AGO4-mediated chromatin modifications (heterochromatin formation) by enhancing DNA and histone methylation (Chellappan et al., 2010; Fei, Xia, & Meyers, 2013). In most plant specimens, the majority of the endogenous regulatory small RNA pools contain heterochromatic siRNAs targeting chromatin that causes DNA methylation (Wang et al., 2015).

## 2. Natural antisense small-interfering RNAs

nat-siRNAs are formed due to the stimulation of partially overlapping transcripts of two adjacent genes under separate biotic and abiotic stresses. They can arise from the sense and antisense transcript, so they can be labeled as *trans-nats* and *cis-nats*, respectively. DCL1/DCL2/DCL3 recognizes the complementary dsRNA for generating 21–24-nt-long nat-siRNAs (Phillips, Dalmay, & Bartels, 2007). Their biogenesis needs several other proteins to interact, such as suppressors of gene silencing (SGS3), RDR2, Pol IV, HEN1, and hyponastic leaf-111 (HYL1). Arabidopsis and rice studies report that the induction of specific biotic and abiotic stresses (especially salt stress) increases the intensity or number of nat-siRNAs. They work mainly at the posttranscriptional stage through either cleavage or translational suppression of target transcripts, although a few instances have been recorded in which they can also direct DNA methylation. (Fei et al., 2013; Wu, Mao, & Qi, 2012). They work by improving the resistance to stress in plants (Naqvi, Choudhury, & Haq, 2011; Zhang et al., 2012).

## 3. *trans*-Acting small-interfering RNAs (ta-siRNAs)

ta-siRNAs are a class of plant-specific endogenous secondary siRNAs that control *trans*-transcriptional mRNAs by increasing AGO4-mediated DNA methylation (Chellappan et al., 2010). Their origin depends on the TAS transcript cleavage driven by miRNA. Transcribed by RNA Polymerase II, the TAS locus is cleaved by the 22-nt-long miRNAs (Felippes & Weigel, 2009; Vazquez et al., 2004). The cleaved transcripts are translated by RDR6 into dsRNA and then processed by DCL4 protein into 23–25-nt siRNAs (Felippes & Weigel, 2009; Vazquez et al., 2004). DCL5 may also process the dsRNA to generate 24-nt phased siRNA. They are involved in controlling the growth and response of plants to different biotic and abiotic stresses (Felippes & Weigel, 2009; Yoshikawa, 2013).

### 7.4.3.2.2 piRNA

piRNAs (23–29-nt long) are derived from transposons and repetitive genomic elements. These are mostly found in animals and involved in providing genome stability by silencing the transposons. (Le Thomas et al., 2014; Weick & Miska, 2014). The first bioinformatics tool for piRNA prediction was developed by applying the Fisher discriminant algorithm to k-mer sequence features using small RNA data (Zhang, Wang, & Kang, 2011). There are various piRNA prediction algorithms available, such as Piano (Wang et al., 2014) and piRNAPredictor (Li, Luo, Zhang, Liu, & Luo, 2016).

### 7.4.3.2.3 miRNAs

miRNAs are 21–24-nt single-stranded small RNA molecules, generated endogenously from RNA transcripts derived from specific genes (Bartel, 2004; Shriram, Kumar, Devarumath, Khare, & Wani, 2016; Xie, Jones, Wang, Sun, & Zhang, 2015). They are also able to control gene expression at transcriptional and posttranscriptional levels during the development of species (Guleria et al., 2011). This is accomplished by controlling the structure of chromatin, chromosome assortment, editing of RNA, stability of RNA, and synthesis of proteins (Carthew & Sontheimer, 2009). During biogenesis they are transcribed by RNA polymerase II (Pol II) as large primary transcripts that are further processed by a protein complex, containing DCL1, TGH, HYL1, and DDL (Tripathi, Goswami, & Sanan-Mishra, 2015). Gene families were examined for gene structures, phylogenetic relationships, and protein motifs, including DCL, HEN1, SE, HYL1, HST, RDR, NRPD1, NRPD2/NRPE2, NRPE1, and AGO. The expression of genes was validated using RNA-seq (Liu, Lu, Dou, Yu, & Zhang, 2014).

For function, the miRNA strands associate with AGO containing RISC complex in the cytoplasm. They bind to their cognate targets in a sequence-specific manner. Many computational approaches have been designed to predict their targets and characterize their putative biological function (Mallory & Vaucheret, 2006). In general, plant miRNAs contain near-perfect complementarity with target sites that occur most frequently in mRNA protein-coding regions (Bonnet, He, Billiau, & Van De Peer, 2010; Llave, Xie, Kasschau, & Carrington, 2002; Rhoades et al., 2002). Degradome offers an opportunity to analyze the miRNA functions, by comparing the miRNA population and their degraded products (Kajal, Kaushal, & Singh, 2019).

During miRNA biogenesis, the indefinite activity of DCL protein leads to the production of different length variants from its canonical processing region. These are known as isomiRs and they can be of different types depending upon the position of the modification at the 5' or 3' end. The type of modification also influences the isomiRs as it may involve the addition or deletion of nucleotides in a template-based or non template-based manner (Neilsen, Goodall, & Bracken, 2012). The functional pathways of the isomiRs are similar to those of the miRNAs and are governed by their sorting with the specific AGO proteins (Cloonan et al., 2011; Goswami, Tripathi, & Sanan-Mishra, 2017; Khan, Goswami, Sopory, & Sanan-Mishra, 2017). Prediction of isomiRs is still controversial because of sequence inconsistency in the datasets. The major limitation is provided by the presence of low-quality reads and sequencing artifacts (Goswami et al., 2017; Khan et al., 2017; Tripathi, Chacon, Lata Singla-Pareek, Sopory, & Sanan-Mishra, 2018).

## 7.4.4 Profiling expression patterns

### 7.4.4.1 Expression level calculation

If the calculated difference for change in read counts between two experimental conditions is statistically important, the gene is considered to be differentially expressed. RNA-seq data are generated as raw read counts, so standardized expression units are necessary to eliminate technical biases in sequenced data that may arise due to sequencing depth (higher depth of sequencing generates more read counts for each transcript) and gene length (differences in gene length create unequal read counts; more read counts for the longer gene). A digital measure of the abundance of transcripts is provided by multiple expression units that calculate read counts as RPM (reads per million), CPM (counts per million), RPKM (reads per kilobase per million mapped reads), FPKM (fragments per kilobase per million mapped reads), TPM (transcript per million), and TMM (trimmed mean of M-values).

TPM or RPM or CPM = abundance of miRNA in experiment/total no. of reads in experiment  $\times 10^6$

Digital expression status of small RNAs like miRNAs is generated by normalizing the expression value as

$$\text{RPKM} = 10^6 \times C \times 10^3 / L \times M$$

Digital expression status of transcripts is generated by normalizing the expression value as RPKM; where  $C$  is the number of reads mapped to gene;  $L$  is the total number of mapped reads; and  $M$  is the gene length in bp.

FPKM is analogous to RPKM and is used especially in paired-end RNA-seq experiments. Two ends (left and right) of the same DNA fragment are sequenced in paired-end RNA-seq. When the paired-end data are subjected to mapping, either both reads or only one read with high quality from a fragment can map to the reference sequence. The fragments that successfully map and not the reads are counted and interpreted for FPKM calculation to avoid uncertainties due to single-read mapping and avoid multiple counting.

$$\text{FPKM} = \text{RM}_g \times 10^9 / \text{RM}_t \times L$$

where  $\text{RM}_g$  is the number mapped reads to the gene;  $\text{RM}_t$  is the total number of read mapped to protein-coding sequences in the alignment; and  $L$  is the length of the gene in base pairs.

$$\text{Fold-change} = \text{NE}/\text{NC}$$

To observe the differential expression pattern, a comparison across the control and treated libraries can be performed. The difference can be calculated as fold-change and values can be represented as chart-plots (Goswami et al., 2017), where NE is the normalized expression (TPM) in the experimental condition and NC is the normalized expression (TPM) in the control condition.

$$\log \text{ fold} = \log(N, 2), \quad \text{where } N \text{ represents the fold-change.}$$

### 7.4.4.2 Comparative analysis and differential expression

Comparative analysis has become a powerful methodology to analyze the gene expression level in different conditions. It shows the genomic characteristics of transcriptional reads and allows us to understand the molecular responses of any gene, protein, miRNA, and siRNA in two conditions to test the differential expression analysis. To screen and classify differentially expressed genes (DEGs) for each entity that may be any cell, tissue, and comparative transcriptome analysis was carried out. Two different conditions or samples are required to perform comparative analysis, such as treatment versus normal tissue/cell/gene and diseased tissue/cell versus normal tissue/cell. Armour (Sanan-Mishra et al., 2018)

provided the significant differential level of miRNAs in different treatments (salt, heat, drought, virus infection, etc.) with respect to normal conditions.

To model RNA-seq data based on the negative binomial distribution, several specialized software packages have been developed. The R package edgeR was developed by Robinson, McCarthy, and Smyth (2010), which initially provides an exact test for two group comparisons and then extends it to allow multifactor designs via a generalized linear model. The R package DESeq2 for differential expression analysis was also developed by Love, Huber, and Anders (2014), which provides shrinkage estimators for both log fold-change and dispersion by forcing a hierarchical model on them (Yu, Fernandez, & Brock, 2017). Several other software are available, such as RNaseqPS (Guo, Zhao, Li, Sheng, & Shyr, 2014), DESeq2 (Love et al., 2014), ALDEX2 (Gloor, 2015), and SigEMD (Wang & Nabavi, 2018). The efficiency of all differential expression methods was assessed for the simulated data using transcript-level abundances.

Analysis of RNA-seq data often deals with large lists of DEGs. To provide useful insights into their collective biological mechanism, annotation enrichment or pathway analysis (Curtis, Orešič, & Vidal-Puig, 2005) is performed. In annotation enrichment analysis, knowledge bases, such as Gene Ontology (GO) or Reactome, use gene or protein annotations to infer specific annotations that are overrepresented in a network. GO analysis has become a common approach for large-scale functional studies of transcriptomic data (Zheng & Wang, 2008).

Enrichment analysis can identify terms that are statistically over- or underrepresented by systematically mapping genes and proteins to their associated biological annotations, such as GO terms (Ashburner et al., 2000) or pathway membership, by comparing the distribution of terms within the gene set of interest with the background. It is assumed that the significant underlying biological mechanism of action can be identified by the enriched terms. There are several functional enrichment analysis tools appropriate for positioning the results (Table 7.4), such as Enrichr (Chen et al., 2013; Kuleshov et al., 2016), WebGestalt (Liao, Wang, Jaehnig, Shi, & Zhang, 2019), Metascape (Zhou et al., 2019), KOBAS (Xie et al., 2011), AgriGO (Tian et al., 2017), ToppGene Suite (Chen, Bardes, Aronow, & Jegga, 2009), and GeneCodis (Nogales-Cadenas et al., 2009).

**TABLE 7.4** List of software and databases useful for analyzing RNA-seq data.

Tool/software	Description	Reference
Cutadapt	A command line tool for the adapter ( <a href="http://code.google.com/p/cutadapt/">http://code.google.com/p/cutadapt/</a> ).	Martin (2011)
FASTX toolkit	Set of command line tools to preprocess Short-Read FASTA/FASTQ data ( <a href="http://hannonlab.cshl.edu/fastx_toolkit/">http://hannonlab.cshl.edu/fastx_toolkit/</a> ).	Gordon and Hannon (2010)
UrQt	Unsupervised Quality trimming of Next-Generation Sequencing reads ( <a href="https://lbbbe.univ-lyon1.fr/-UrQt-.html">https://lbbbe.univ-lyon1.fr/-UrQt-.html</a> ).	Modolo and Lerat (2015)
Fastq_clean	An integrated pipeline for cleaning data from the Illumina sequencing with quality control ( <a href="https://github.com/gaoshanT/Fastq_clean/">https://github.com/gaoshanT/Fastq_clean/</a> ).	Zhang et al. (2014)
NGS toolkit	A quality control toolkit for next-generation data sequencing (NGS) ( <a href="http://www.nipgr.res.in/ngsqctoolkit.html">http://www.nipgr.res.in/ngsqctoolkit.html</a> ).	Patel and Jain (2012)
AdapterRemoval	A comprehensive tool for preprocess of both single- and paired-end NGS data ( <a href="http://code.google.com/p/adapterremoval/">http://code.google.com/p/adapterremoval/</a> ).	Lindgreen (2012)
Trimmomatic	Trimmomatic performs a variety of useful quality control tasks for Illumina paired-end and single-end reads ( <a href="http://www.usadellab.org/cms/index.php?page=trimmomatic/">http://www.usadellab.org/cms/index.php?page=trimmomatic/</a> ).	Bolger et al. (2014)
Bowtie/Bowtie2	Short read alignment to the ref. genome and allow up to three mismatches ( <a href="http://bowtie-bio.sourceforge.net/index.shtml">http://bowtie-bio.sourceforge.net/index.shtml</a> ).	Langmead and Salzberg (2012)
Burrows-Wheeler Aligner (BWA)	BWA is a suite of tools to map a broad reference genome against low-divergent sequences ( <a href="https://github.com/lh3/bwa/">https://github.com/lh3/bwa/</a> ).	Li and Durbin (2010)
Short Oligonucleotide Analysis Package	Alignment tool that provides various tools in a single package. SOAPaligner/soap2, SOAPsnp, SOAPindel, SOAPsv, SOAPdenovo, and SOAP3/GPU ( <a href="http://soap.genomics.org.cn/">http://soap.genomics.org.cn/</a> ).	Li, Li, Kristiansen, and Wang (2008), Li (2009)
SeqMap	It allows up to five mismatches (insertion/deletion), various modification options and input/output formats ( <a href="http://biogibbs.stanford.edu/~jiangh/SeqMap/">http://biogibbs.stanford.edu/~jiangh/SeqMap/</a> ).	Jiang and Wong (2008)

(Continued)

**TABLE 7.4 (Continued)**

Tool/software	Description	Reference
SAMStat	Analyzing unmapped and poorly and correctly mapped sequences separately to infer possible causes of bad mapping ( <a href="http://samstat.sourceforge.net/">http://samstat.sourceforge.net/</a> ).	Lassmann, Hayashizaki, and Daub (2011)
FastQ Screen	Screens FASTQ format sequences to a database to check that the sequences contain the correct details ( <a href="https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/">https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/</a> ).	Wingett and Andrews (2018)
	Aligns short reads with reference sequences and provides particular sensitivity to errors, SNPs, inserts, and deletions ( <a href="http://bfast.sourceforge.net/">http://bfast.sourceforge.net/</a> ).	Homer, Merriman, and Nelson (2009)
RefShannon	A novel genome-guided transcriptome assembler ( <a href="https://github.com/shunfumao/RefShannon/">https://github.com/shunfumao/RefShannon/</a> ).	Mao et al. (2020)
StringTie	A computational method that allows improved transcriptome reconstruction from RNA-seq reads ( <a href="http://ccb.jhu.edu/software/stringtie/">http://ccb.jhu.edu/software/stringtie/</a> ).	Pertea et al. (2015)
Bayesemmer	A novel probabilistic method for transcriptome assembly built on a Bayesian model of the RNA sequencing process ( <a href="https://github.com/bioinformatics-centre/bayesemmer/">https://github.com/bioinformatics-centre/bayesemmer/</a> ).	Shi (2017)
IsoLasso	A LASSO Regression Approach to Transcriptome Assembly Based on RNA-Seq ( <a href="http://alumni.cs.ucr.edu/~liw/isolasso.html">http://alumni.cs.ucr.edu/~liw/isolasso.html</a> ).	Li, Feng, and Jiang (2011)
TopHat	Alignment of throughput reads of shotgun cDNA sequencing created by transcriptomics technologies ( <a href="http://tophat.cbcb.umd.edu/">http://tophat.cbcb.umd.edu/</a> ).	Trapnell, Pachter, and Salzberg (2009)
SpliceMap	Alignment tool and provides great sensitivity and support for arbitrary read lengths for RNA-seq ( <a href="https://web.stanford.edu/group/wonglab/SpliceMap/">https://web.stanford.edu/group/wonglab/SpliceMap/</a> ).	Au et al. (2010)
TransAByss	Tool for <i>de novo</i> transcriptome assembly using short reads ( <a href="http://www.bcgsc.ca/platform/bioinfo/software/">http://www.bcgsc.ca/platform/bioinfo/software/</a> ).	Robertson et al. (2010)
Trinity	Tool for <i>de novo</i> transcriptome assembly of RNA-seq data ( <a href="http://TrinityRNASeq.sourceforge.net/">http://TrinityRNASeq.sourceforge.net/</a> ).	Henschel et al. (2012), Grabherr et al. (2011)
OASES	Rugged <i>de novo</i> RNA-seq assembly around the dynamic range of levels of expression ( <a href="http://www.ebi.ac.uk/~zerbino/oases/">http://www.ebi.ac.uk/~zerbino/oases/</a> ).	Schulz et al. (2012)
rnaSPAdes	A <i>de novo</i> transcriptome assembler and its application to RNA-seq data ( <a href="https://github.com/ablab/spades/">https://github.com/ablab/spades/</a> ).	Bushmanova, Antipov, Lapidus, and Prjibelski (2019)
SOAPdenovo-Trans	A <i>de novo</i> transcriptome assembler designed specifically for RNA-seq ( <a href="http://sourceforge.net/projects/soapdenovotrans/">http://sourceforge.net/projects/soapdenovotrans/</a> ).	Xie et al. (2014)
DESeq	A R package to test high-throughput sequencing assay count data, such as RNA-seq and differential expression checking ( <a href="https://bioconductor.org/packages/DESeq/">https://bioconductor.org/packages/DESeq/</a> ).	Anders (2010)
easyRNASeq	A bioconductor kit for RNA-seq data processing ( <a href="http://bioconductor.org/packages/release/bioc/html/easyRNASeq.html">http://bioconductor.org/packages/release/bioc/html/easyRNASeq.html</a> ).	Delhomme, Padioleau, Furlong, and Steinmetz (2012)
Cufflinks	The Cufflinks suite of tools can be used to perform Transcriptome assembly and differential expression analysis for RNA-seq ( <a href="http://cufflinks.cbcb.umd.edu/">http://cufflinks.cbcb.umd.edu/</a> ).	Trapnell et al. (2012)
edgeR	A Bioconductor package for digital gene expression data processing for differential expression analysis ( <a href="http://bioconductor.org/">http://bioconductor.org/</a> ).	Robinson et al. (2010)
MultiRankSeq	Multiperspective Approach for RNA-seq Differential Expression Analysis and Quality Control ( <a href="https://github.com/slzhao/MultiRankSeq">https://github.com/slzhao/MultiRankSeq</a> ).	Guo, Zhao, Ye, Sheng, and Shyr (2014)
RNAseqPS	A Web Tool for Estimating Sample Size and Power for RNA-seq Experiment ( <a href="http://cqs.mc.vanderbilt.edu/shiny/RNAseqPS/">http://cqs.mc.vanderbilt.edu/shiny/RNAseqPS/</a> ).	Guo, Zhao, Li, et al. (2014)
DEXSeq	A bioconductor package that recognizes the use of differential exons based on RNA-Seq exon counts between samples ( <a href="http://sourceforge.net/projects/differential-sj-usage/files/">http://sourceforge.net/projects/differential-sj-usage/files/</a> ).	Li, Rao, Mattox, Amos, and Liu (2015)

(Continued)

**TABLE 7.4 (Continued)**

Tool/software	Description	Reference
DEXUS	DEXUS is a bioconductor package that recognizes RNA-Seq data with differentially expressed genes in all possible research designs ( <a href="http://www.bioinf.jku.at/software/dexus/">http://www.bioinf.jku.at/software/dexus/</a> ).	Klambauer, Unterthiner, and Hochreiter (2013)
snpMatrix	This class describes objects that carry large arrays of genotypes of single nucleotide polymorphism (SNP) generated using array technologies ( <a href="http://www.bioconductor.org/packages/2.7/bioc/html/snpMatrix.html">http://www.bioconductor.org/packages/2.7/bioc/html/snpMatrix.html</a> ).	Clayton (2010), Clayton and Leung (2007)
Matrix eQTL	Matrix eQTL is designed to handle large genotype and expression datasets ( <a href="http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL">http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL</a> ).	Shabalín (2012)
FastMap	A fast algorithm for indexing, data-mining, and visualization of traditional and multimedia datasets ( <a href="http://cebc.unc.edu/fastmap86.html">http://cebc.unc.edu/fastmap86.html</a> ).	Faloutsos and Lin (1995), Gatti et al. (2009)
Genome Analyzer Toolkit (GATK)	A software program to effectively recover RNA-seq data from raw allelic count data and to evaluate the characteristics of AE data and the sources of errors and technical variations ( <a href="https://gatk.broadinstitute.org/hc/en-us">https://gatk.broadinstitute.org/hc/en-us</a> ).	McKenna et al. (2010)
Enrichr	Integrative research platform for web-based and mobile gene list enrichment that includes more than 30 gene-set libraries ( <a href="http://amp.pharm.mssm.edu/Enrichr">http://amp.pharm.mssm.edu/Enrichr</a> ).	Kuleshov et al. (2016), Chen et al. (2013)
WebGestalt	The Web-based Gene SeT AnaLysis Toolkit is a web tool for functional enrichment analysis. ( <a href="http://www.webgestalt.org">http://www.webgestalt.org</a> ).	Wang, Vasaiakar, Shi, Greer, and Zhang (2017)
EasyGO	Gene ontology-based research platform for functional enrichment ( <a href="http://bioinformatics.cau.edu.cn/easygo">http://bioinformatics.cau.edu.cn/easygo</a> ).	Leng et al. (2015)
AgriGO	AgriGO is a web-based gene ontology research platform and database ( <a href="http://bioinfo.cau.edu.cn/agriGO/">http://bioinfo.cau.edu.cn/agriGO/</a> ).	Du, Zhou, Ling, Zhang, and Su (2010), Tian et al. (2017)
Lncident	A tool to classify long noncoding RNAs rapidly using intrinsic sequence composition and Open Reading Frame information ( <a href="http://csbl.bmb.uga.edu/mirrors/JLU/Lncident/index.php">http://csbl.bmb.uga.edu/mirrors/JLU/Lncident/index.php</a> ).	Han et al. (2016)
CNIT	A fast and accurate web tool based on intrinsic sequence composition for the identification of protein-coding and long noncoding transcripts ( <a href="http://cnit.noncode.org/CNIT">http://cnit.noncode.org/CNIT</a> ).	Guo et al. (2019)
LncFinder	An integrated platform using sequence intrinsic composition for long noncoding RNA identification ( <a href="https://CRAN.R-project.org/package=LncFinder">https://CRAN.R-project.org/package=LncFinder</a> ).	Han et al. (2019)
miRDeep-P	miRDeep-P, an effective tool for plant-specific miRNA characterizing from NSG data ( <a href="http://faculty.virginia.edu/lilab/miRDP/">http://faculty.virginia.edu/lilab/miRDP/</a> ).	Yang and Li (2011)
miRCat2	A new entropy-based approach to detect miRNA loci, part of UEA small RNA workbench ( <a href="http://srna-workbench.cmp.uea.ac.uk/">http://srna-workbench.cmp.uea.ac.uk/</a> ).	Paicu et al. (2017), Stocks et al. (2012)
sRNAtools	sRNAtools is a suite of web-based tools for the identification of miRNA, their targets and expression profiling of small RNAs as well as prediction of ta-siRNAs ( <a href="http://srna-tools.cmp.uea.ac.uk">http://srna-tools.cmp.uea.ac.uk</a> ).	Moxon et al. (2008)
PMRD	Plant miRNA database integrates available plant miRNA data deposited in public databases and contains sequence information, secondary structure, expression profiling, and target genes ( <a href="http://bioinformatics.cau.edu.cn/PMRD/">http://bioinformatics.cau.edu.cn/PMRD/</a> ).	Zhang et al. (2010)
miRBase	miRBase, a searchable database of published miRNA, precursor; it is a central repository for miRNA sequence information ( <a href="http://www.mirbase.org/index.shtml">http://www.mirbase.org/index.shtml</a> ).	Griffiths-Jones, Saini, Van Dongen, and Enright (2007)
PNRD	PNRD is a systematic forum for the study of plant's noncoding RNA (ncRNA) ( <a href="http://structuralbiology.cau.edu.cn/PNRD">http://structuralbiology.cau.edu.cn/PNRD</a> ).	Yi, Zhang, Ling, Xu, and Su (2014)
miRMOD	A tool for the identification of both template and nontemplate-based isomiRs with respect to their canonical miRNA ( <a href="http://bioinfo.icgeb.res.in/miRMOD/">http://bioinfo.icgeb.res.in/miRMOD/</a> ).	Kaushik, Saraf, Mukherjee, and Gupta (2015)

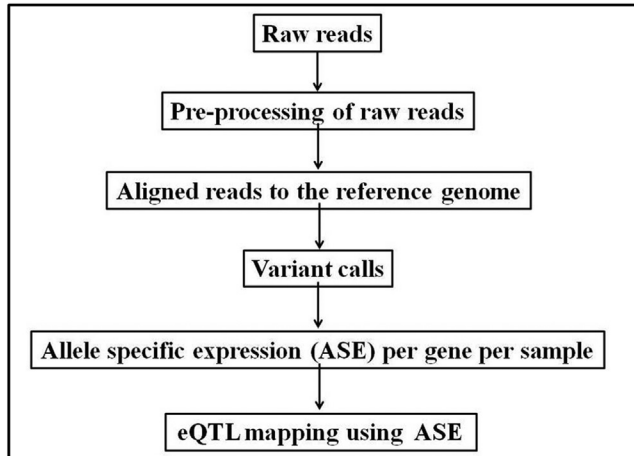
(Continued)

**TABLE 7.4 (Continued)**

Tool/software	Description	Reference
Phasetank	Phasetank, an open-source tool for the prediction of phased siRNA from small RNAs and their target as well as its initiative miRNA ( <a href="http://phasetank.sourceforge.net/">http://phasetank.sourceforge.net/</a> ).	Guo, Qu, and Jin (2014)
NATpipe	A tool for prediction natural antisense transcripts, including the detection and functional analysis of nat-siRNA and phase-distributed nat-siRNAs ( <a href="http://www.bioinfolab.cn/NATpipe/NATpipe.zip">http://www.bioinfolab.cn/NATpipe/NATpipe.zip</a> ).	Yu, Meng, Zuo, Xue, and Wang (2016)
deepBase	Database for annotation and exploration of high-throughput deep sequencing data from small and long ncRNAs ( <a href="http://deepbase.sysu.edu.cn/">http://deepbase.sysu.edu.cn/</a> ).	Zheng et al. (2015)
PMTED	PMTED is used for the retrieval and study of miRNA target expression profiles ( <a href="http://pmted.agrinome.org/">http://pmted.agrinome.org/</a> ).	Sun et al. (2013)
Plant MPSS	It measures the level of expression of most genes under specified conditions and with the support of public HTS data and provides information on potentially new transcripts ( <a href="http://mpss.udel.edu">http://mpss.udel.edu</a> ).	Nakano, Nobuta et al. (2006)
Rfam	A resource of a comprehensive collection of ncRNA families ( <a href="http://rfam.xfam.org/">http://rfam.xfam.org/</a> ).	Griffiths-Jones et al. (2005)
Armour	The Armour database offers a consistent platform and expression for novel and mature miRNAs ( <a href="http://armour.icgeb.trieste.it/">http://armour.icgeb.trieste.it/</a> ).	Sanan-Mishra et al. (2018)
miRNEST	miRNEST is an integrative array of microRNA data from animals, plants, and viruses ( <a href="http://rhesus.amu.edu.pl/mirnest/copy/">http://rhesus.amu.edu.pl/mirnest/copy/</a> ).	Szczęśniak and Makałowska (2014)
Piano	Piano demonstrates excellent piRNA prediction efficiency ( <a href="http://ento.njau.edu.cn/Piano.html">http://ento.njau.edu.cn/Piano.html</a> ).	Wang et al. (2014)
piRNAPredictor	A weighted ensemble genetic algorithm-based approach for the prediction of transposon-derived piRNAs ( <a href="https://github.com/zw9977129/piRNAPredictor">https://github.com/zw9977129/piRNAPredictor</a> ).	Li et al. (2016)
piRNN	piRNN was developed using a deep learning model for piRNA prediction based on the convolution neural network framework ( <a href="https://github.com/bioinfolabmu/piRNN">https://github.com/bioinfolabmu/piRNN</a> ).	Wang, Hoeksema, and Liang (2018)
IsomiR Bank	IsomiR Bank, a database, containing the prediction isomiRs from various samples from plant and animal NGS data set ( <a href="http://mcg.ustc.edu.cn/bsc/isomir/">http://mcg.ustc.edu.cn/bsc/isomir/</a> ).	Zhang et al. (2016)
tasiRNAdb	tasiRNAdb, a database for the identification of miRNAs associated with ta-siRNA regulatory pathway, including TASs, ta-siRNAs, and ta-siRNA targets ( <a href="http://bioinfo.jit.edu.cn/tasiRNADatabase">http://bioinfo.jit.edu.cn/tasiRNADatabase</a> ).	Zhang, Li, Zhu, Zhang, and Fang (2013)
Gene Ontology	The knowledge base of gene ontology is the world's largest source of information on gene functions ( <a href="http://geneontology.org">http://geneontology.org</a> ).	Consortium (2004)
DyNB	DyNB is another MATLAB implementation that models RNA-seq counts and their temporal correlations using NB distribution and GP ( <a href="http://research.ics.aalto.fi/csb/software/">http://research.ics.aalto.fi/csb/software/</a> ).	Äijö et al. (2014)
EBSeqHMM	An EBSeq R package extension. To classify dynamic genes in two steps, it uses an empirical Bayes autoregressive hidden Markov model (AR-HMM) ( <a href="http://www.bioconductor.org/packages/devel/bioc/html/EBSeqHMM.html">http://www.bioconductor.org/packages/devel/bioc/html/EBSeqHMM.html</a> ).	Leng et al. (2015)

### 7.4.5 Degradome sequencing

Degradome sequencing, also known as Parallel Analysis RNA End (PARE), is helpful in the detection of cleaved miRNA targets. It is understood that miRNA guides (AGO) protein from 5' end to 9th–11th nucleotide cleave target mRNA. Degradome sequencing for the cleaved materials is used to validate the miRNA–mRNA target pair. It also provides information about cleavage sites and has helped identify target transcripts of many recognized and novel miRNA and ta-siRNA plants (Addo-Quaye, Miller, & Axtell, 2009; Song et al., 2017; Tripathi, Goswami, Tiwari, Mukherjee, & Sanan-Mishra, 2018).



**FIGURE 7.3** Flowchart for identification of allele-specific expression from RNA-seq.

#### 7.4.6 Variants detection and allele-specific expression

The phenotypic differences between individual organisms can be captured by the analysis of ASE. This quantifies the relative expression of two alleles in a diploid or polyploid genome and is useful for identifying variations in *cis*-regulated gene expression (Fan et al., 2020; Ge et al., 2009; Kwaepila, Burns, & Leong, 2006). Normally a gene may have two variants, derived from each parent, that are located on a chromosome at the same location or genetic locus and these are known as alleles. The ASE refers to the feature when a parental allele is preferentially expressed in the hybrid due to differences in genome regulatory sequences (Gaur, Li, Mei, & Liu, 2013; Shao et al., 2019). The difference in expression caused by ASE can lead to phenotypic or genotypic variations depending on the function of the gene. For instance, within a translated region, ASE can result in a heterozygous variant producing modified or truncated protein (Kukurba et al., 2014). This can result in differential binding of transcription factors or epigenetic modifiers at a regulatory site (Prendergast, Tong, Hay, Farrington, & Semple, 2012; Reddy et al., 2012); or it can affect transcript processing at a splice site or UTR (Li et al., 2012).

The ASE has been proposed as a mechanism behind heterosis, which has dramatically improved the production of many crops globally (Shao et al., 2019). ASE has also been linked to various genetic disorders in mammalian systems. Tumors often observe ASE (Curia et al., 2012; Valle et al., 2008; Walker et al., 2012). In 2002 ASE was first proposed as a direct approach to relate a genotype to susceptibility to disease (Yan, Yuan, Velculescu, Vogelstein, & Kinzler, 2002). Therefore ASE analysis forms an important component of the RNA-seq data (Fig. 7.3). A method ASEP (Allele-Specific Expression analysis in a Population) was the first population-level detection method used for gene-level ASE identification in human diseases (Fan et al., 2020). A new tool in the Genome Analyzer Toolkit (GATK) software package can efficiently recover raw allelic count data from RNA-seq data to analyze the properties of ASE (McKenna et al., 2010).

#### 7.4.7 Expression quantitative trait loci

Genomic loci that contribute to variation in transcript expression levels are represented by expression quantitative trait loci (eQTLs) (Nica & Dermitzakis, 2013). Generally, these are associated with a single gene at a particular chromosomal location. It enables to distinguish between the quantitative characteristics of expression from the most complex characteristics that are not the product of single gene expression.

Standard eQTL analysis requires a direct association test between genetic variations markers with levels of gene expression usually tested in tens or hundreds of individuals. This interaction study can be conducted proximal or distal to the gene and the regulatory variants can usually be characterized as either *cis*- or *trans*-acting. Local eQTLs also referred to as *cis*-eQTLs are located near the gene of origin (a gene that generates the transcript or protein) (Ma, 2018). Those located far from their gene of origin, often on other chromosomes, are referred to as distant eQTLs or *trans*-eQTLs. The nomenclature not only reflects the physical distance from the gene they regulate but also predicts the nature of the interactions involved.

QTL analysis offers data evaluating gene expression in a segregating population, and the abundance of each gene transcript is evaluated with regard to the phenotype. A statistical correlation between genetic markers found in

particular regions of the genome and the degree of expression of the assayed gene is the product of this research. The resulting eQTL plot shows the possible DNA sequence genetic position that causes the observed difference in the abundance of transcripts (West et al., 2007). In addition, eQTL data allow genetic regulatory networks to be modeled and offer a deeper understanding of the underlying phenotypic variation. The GWAS approach enables eQTL mapping to define new functional loci without needing any prior knowledge of existing *cis* and *trans* regulatory regions (Nica & Dermitzakis, 2013). ASE is an essential component for eQTL mapping (Fig. 7.3). There are various tools available that can be employed for this analysis, such as eMap (Chatzinotas & Sampson, 2004), snpMatrix (Clayton, 2010), Plink (Purcell et al., 2007), Merlin (Abecasis, Cherny, Cookson, & Cardon, 2002), FastMap (Gatti et al., 2009), and Matrix eQTL (Shabalin, 2012).

## 7.5 Databases and software for small RNA analysis

After alignment of NGS reads to the reference genome, mapped reads can be directly used for the identification of miRNAs by mapping back to the known miRNA sequences or predicting novel molecules. Fortunately, various valuable computational approaches, tools, and stand-alone programs are freely available online, for the identification and prediction of small RNAs (Table 7.4). The tools can also calculate their expression patterns and predict their target transcripts to indicate their putative role in an organism. Few well-known database/toolkits that provide a comprehensive platform for RNA-seq data analysis include the following.

### 7.5.1 miRBase

This database (<http://www.mirbase.org/>) is the most authentic repository for mature miRNAs and their precursors in plants as well as animals (Kozomara & Griffiths-Jones, 2014). This repository provides complete information on miRNA mature sequences, hairpin precursor's sequences along with genome coordinates community annotation, evidence, and reference. The miRbase release 22 contains 10,414 mature miRNAs and 8615 hairpin precursors for 83 plant species from Viridiplantae kingdom (Kozomara & Griffiths-Jones, 2014).

### 7.5.2 PMRD

It is a publically available database of plant miRNAs that provides complete information on their secondary structure, expression profiles, putative target genes, etc. In PMRD, there are 8433 miRNAs identified from 121 plant species, including *Arabidopsis*, wheat, rice, sorghum, soybean, and maize (Zhang et al., 2010). PMRD also facilitates microarray-based expression profiles of miRNA related to the oxidative stress.

### 7.5.3 Armour

It is a rice-specific miRNA database that contains information on known and novel miRNAs identified using Illumina sequencing and validated experimentally. It facilitates extensive ways to analyze miRNA expression in response to abiotic stresses (heat and salt) across different rice tissues (root, leaf, flag leaf, panicle, and flower). It provides complete information on miRNAs, such as sequence, the secondary structure of precursor, MFE, read counts, and information on targets, including GO/KO annotation (<http://armour.icgeb.trieste.it/>) (Sanan-Mishra et al., 2018).

### 7.5.4 UEA sRNA workbench

UEA sRNA workbench is a collection of tools that provide a comprehensive platform to analyze NGS data (<http://srna-workbench.cmp.uea.ac.uk/>) (Mohorianu, Stocks, Applegate, Folkes, & Moulton, 2017). It includes:

- Adapter removal to remove the adapter from both ends of sequenced reads;
- miRCat, a package for novel miRNA prediction (Paicu et al., 2017);
- TA-SI Prediction tool for prediction of ta-siRNA and phased reads aligned to the genome;
- PAREsnip for target prediction from degradome data and many other tools can be used along with miRNAs prediction; and
- CoLide tool for performing analysis, such as expression profiling of small RNAs.

### 7.5.5 sRNAtoolbox

It is a collection of tools that can be used independently to analyze the small RNAs sequencing data (<http://bioinfo5.ugr.es/srnatoolbox>) (Rueda et al., 2015). It performs expression profiling, isomiRs' identification, and analysis of known miRNAs and other types of small RNAs, produces multiple graphical summaries, and predicts novel miRNAs in both plants and animals.

### 7.5.6 sRNAbench

sRNAbench is a free web server tool and stand-alone application obtained from NGS platforms, such as Illumina or SOLiD, to process small-RNA data. It is used for the prediction of novel miRNAs and their length variants (isomiRs) as well as expression profiling.

### 7.5.7 miRNAconsTarget

It provides an integrated platform for target prediction from both animals and plants and many similar tools are available to analyze sRNA data.

### 7.5.8 Massively parallel signature sequencing database

Massively parallel signature sequencing (MPSS) is a unique signature-based transcription resource for the analysis of mRNA and small RNA. It facilitates the search for gene expression data in model plants, such as *Arabidopsis* and rice (<http://mpss.udel.edu>) (Nakano, Suzuki, Fujimura, & Shinshi, 2006). MPSS database provides various sequencing datasets, such as small RNAs, PARE data, mRNA differential expression, DNA methylated data, and Chip sequencing data. High-throughput sequencing data can be accessed from the GEO short-read archive (<https://www.ncbi.nlm.nih.gov/geo/>). There are many tools available for the analysis of plant small RNAs and target prediction. The MPSS database also facilitates the identification of known and novel ta-siRNA loci in TAS transcripts with respect to many plant species (Nakano, Suzuki, et al., 2006).

Compare is a web resource to sort and examine miRNA-target interaction validated using PARE data ([https://mpss.danforthcenter.org/tools/mirna\\_apps/comPARE.php](https://mpss.danforthcenter.org/tools/mirna_apps/comPARE.php)).

sPARTA, small RNA-PARETargets Analyzer, is a software for the validation of plant miRNAs or sRNA targets can be used for whole genome analysis.

MicroRNA Truncation and Tailing Analysis (miTRATA) is a web-based tool to find the differential of 3' nucleotide modifications in miRNA (3' isomiRs) in respect to the canonical sequences (<https://wasabi.ddpsc.org/~apps/ta/>).

### 7.5.9 CLC Genomics Workbench

It is a powerful solution that uses Cutting-edge technology, unique characteristics, and algorithms commonly used by industry and academic science leaders make it easy to solve data analysis challenges.

## 7.6 Conclusion

The development of various NGS platforms has supported a variety of applications for RNA-seq. It has been particularly useful in the discovery and mapping of the ncRNA species that regulate a variety of biological functions. In this chapter, we have described and compared the different SGS and TGS platforms. The parameters and tools for processing and analyzing the generated data to obtain significant results have also been discussed. The rapid developments in RNA-seq have opened numerous avenues for research. It has allowed the identification of the cellular transcriptome, quantification of the expression of various alleles and transcripts, and understanding of transcript editing processes, such as alternative splicing and gene fusion. It is envisaged that RNA-seq approaches will facilitate global data analysis on genomics, transcriptomics, and regulomics to enhance understanding of the operation and regulation of genetic networks.

## Conflict of interest

The authors declare that the study was performed in the absence of any commercial or financial relationships that could be perceived as a possible conflict of interest.

## References

- Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin—Rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, *30*(1), 97–101.
- Addo-Quaye, C., Miller, W., & Axtell, M. J. (2009). CleaveLand: A pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics (Oxford, England)*, *25*(1), 130–131.
- Äijö, T., Butty, V., Chen, Z., Salo, V., Tripathi, S., Burge, C. B., . . . Lähdesmäki, H. (2014). Methods for time series analysis of RNA-seq data with application to human Th17 cell differentiation. *Bioinformatics (Oxford, England)*, *30*(12), i113–i120.
- Anders, S. (2010). Analysing RNA-Seq data with the DESeq package. *Molecular Biology*, *43*(4), 1–17.
- Arsenic, R., Treue, D., Lehmann, A., Hummel, M., Dietel, M., Denkert, C., & Budczies, J. (2015). Comparison of targeted next-generation sequencing and Sanger sequencing for the detection of PIK3CA mutations in breast cancer. *BMC Clinical Pathology*, *15*(1), 1–9.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Eppig, J. T. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29.
- Au, K. F., Jiang, H., Lin, L., Xing, Y., & Wong, W. H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Research*, *38*(14), 4570–4578.
- Bao, E., Jiang, T., & Girke, T. (2014). AlignGraph: Algorithm for secondary de novo genome assembly guided by closely related references. *Bioinformatics (Oxford, England)*, *30*(12), i319–i328.
- Barrett, L. W., Fletcher, S., & Wilton, S. D. (2013). *Untranslated gene regions and other non-coding elements. Untranslated gene regions and other non-coding elements*. Springer.
- Bartel, D. P. (2004). MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell*, *116*(2), 281–297.
- Behera, S., Voshall, A., Deogun, J. S., & Moriyama, E. N. (2017). Performance comparison and an ensemble approach of transcriptome assembly. In *2017 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 2226–2228). IEEE.
- Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood-Education and Practice*, *98*(6), 236–238.
- Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, *41*(D1), D36–D42.
- Bischof, O., & Martínez-Zamudio, R. I. (2015). MicroRNAs and lncRNAs in senescence: A review. *IUBMB Life*, *67*(4), 255–267.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120.
- Bonnet, E., He, Y., Billiau, K., & Van De Peer, Y. (2010). TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics (Oxford, England)*, *26*(12), 1566–1568.
- Buermans, H., & Den Dunnen, J. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta – Molecular Basis of Disease*, *1842*(10), 1932–1941.
- Bushmanova, E., Antipov, D., Lapidus, A., & Prjibelski, A. D. (2019). rnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, *8*(9), giz100.
- Carthew, R. W., & Sontheimer, E. J. (2009). Origins and mechanisms of miRNAs and siRNAs. *Cell*, *136*(4), 642–655.
- Chapman, E. J., & Carrington, J. C. (2007). Specialization and evolution of endogenous small RNA pathways. *Nature Reviews Genetics*, *8*(11), 884–896.
- Chatzinotas, S., & Sampson, D. (2004). EMAP: Design and implementation of educational metadata application profiles. In *IEEE international conference on advanced learning technologies, 2004. Proceedings* (pp. 876–877). IEEE.
- Chau, B. L., & Lee, K. A. (2007). Function and anatomy of plant siRNA pools derived from hairpin transgenes. *Plant Methods*, *3*(1), 13.
- Chellappan, P., Xia, J., Zhou, X., Gao, S., Zhang, X., Coutino, G., . . . Jin, H. (2010). siRNAs from miRNA sites mediate DNA methylation of target genes. *Nucleic Acids Research*, *38*(20), 6883–6894.
- Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., . . . Ma’ayan, A. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*(1), 128.
- Chen, J., Bardes, E. E., Aronow, B. J., & Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, *37*(suppl\_2), W305–W311.
- Chi, K. R. (2008). *The year of sequencing*. Nature Publishing Group.
- Chopra, R., Burow, G., Farmer, A., Mudge, J., Simpson, C. E., & Burow, M. D. (2014). Comparisons of de novo transcriptome assemblers in diploid and polyploid species using peanut (*Arachis spp.*) RNA-Seq data. *PLoS One*, *9*(12), e115055.
- Clayton, D. (2010). Imputed SNP analyses and meta-analysis with snpMatrix.
- Clayton, D., & Leung, H.-T. (2007). An R package for analysis of whole-genome association studies. *Human Heredity*, *64*(1), 45–51.
- Cloonan, N., Wani, S., Xu, Q., Gu, J., Lea, K., Heater, S., . . . Nourbakhsh, E. (2011). MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biology*, *12*(12), 1–20.

- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... Zhang, X. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1), 13.
- Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(suppl\_2), D258–D261.
- Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., ... Boddu, S. (2019). Ensembl 2019. *Nucleic Acids Research*, 47(D1), D745–D751.
- Curia, M. C., De Iure, S., De Lellis, L., Veschi, S., Mammarella, S., White, M. J., ... Lombardo, M. (2012). Increased variance in germline allele-specific expression of APC associates with colorectal cancer. *Gastroenterology*, 142(1), 71–77.e1.
- Curtis, R. K., Orešič, M., & Vidal-Puig, A. (2005). Pathways to the analysis of microarray data. *Trends in Biotechnology*, 23(8), 429–435.
- Delhomme, N., Padiou, I., Furlong, E. E., & Steinmetz, L. M. (2012). easyRNASeq: A bioconductor package for processing RNA-Seq data. *Bioinformatics (Oxford, England)*, 28(19), 2532–2533.
- Du, Z., Zhou, X., Ling, Y., Zhang, Z., & Su, Z. (2010). agriGO: A GO analysis toolkit for the agricultural community. *Nucleic Acids Research*, 38(suppl\_2), W64–W70.
- Faloutsos, C., & Lin, K.-I. (1995). FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of the 1995 ACM SIGMOD international conference on management of data* (pp. 163–174).
- Fan, J., Hu, J., Xue, C., Zhang, H., Susztak, K., Reilly, M. P., ... Li, M. (2020). ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLoS Genetics*, 16(5), e1008786.
- Fei, Q., Xia, R., & Meyers, B. C. (2013). Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *The Plant Cell*, 25(7), 2400–2415.
- Felippes, F. F., & Weigel, D. (2009). Triggering the formation of tasiRNAs in Arabidopsis thaliana: The role of microRNA miR173. *EMBO Reports*, 10(3), 264–270.
- Fox, S. E., Geniza, M., Hanumappa, M., Naithani, S., Sullivan, C., Preece, J., ... Sage, A. (2014). De novo transcriptome assembly and analyses of gene expression during photomorphogenesis in diploid wheat *Triticum monococcum*. *PLoS One*, 9(5), e96855.
- Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6), 469–477.
- Gatter, T., & Stadler, P. F. (2018). Ryuto: A framework for network-flow based transcriptome reconstruction of RNA-seq data.
- Gatti, D. M., Shabalin, A. A., Lam, T.-C., Wright, F. A., Rusyn, I., & Nobel, A. B. (2009). FastMap: Fast eQTL mapping in homozygous populations. *Bioinformatics (Oxford, England)*, 25(4), 482–489.
- Gaur, U., Li, K., Mei, S., & Liu, G. (2013). Research progress in allele-specific expression and its regulatory mechanisms. *Journal of Applied Genetics*, 54(3), 271–283.
- Ge, B., Pokholok, D. K., Kwan, T., Grundberg, E., Morcos, L., Verlaan, D. J., ... Gagné, V. (2009). Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nature Genetics*, 41(11), 1216–1222.
- Geniza, M., & Jaiswal, P. (2017). Tools for building de novo transcriptome assembly. *Current Plant Biology*, 11, 41–45.
- Ghosh, S., & Chan, C.-K. K. (2016). *Analysis of RNA-Seq data using TopHat and Cufflinks*. *Plant bioinformatics*. Springer.
- Gloor, G. (2015). ALDEx2: ANOVA-Like Differential Expression tool for compositional data. ALDEx manual modular, 20, 1–11.
- Gordon, A., & Hannon, G. (2010). Fastx-toolkit. FASTQ/A short-reads preprocessing tools 5 (unpublished). <[http://hannonlab.cshl.edu/fastx\\_toolkit](http://hannonlab.cshl.edu/fastx_toolkit)>.
- Goswami, K., Tripathi, A., & Sanan-Mishra, N. (2017). Comparative miRomics of salt-tolerant and salt-sensitive rice. *Journal of Integrative Bioinformatics*, 14(1).
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... Zeng, Q. (2011). Trinity: Reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, 29(7), 644.
- Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., & Bateman, A. (2005). Rfam: Annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, 33, D121–D124.
- Griffiths-Jones, S., Saini, H. K., Van Dongen, S., & Enright, A. J. (2007). miRBase: Tools for microRNA genomics. *Nucleic Acids Research*, 36(suppl\_1), D154–D158.
- Guleria, P., Mahajan, M., Bhardwaj, J., & Yadav, S. K. (2011). Plant small RNAs: Biogenesis, mode of action and their roles in abiotic stresses. *Genomics, Proteomics and Bioinformatics*, 9(6), 183–199.
- Guo, J.-C., Fang, S.-S., Wu, Y., Zhang, J.-H., Chen, Y., Liu, J., ... Xu, L.-Y. (2019). CNIT: A fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. *Nucleic Acids Research*, 47(W1), W516–W522.
- Guo, Q., Qu, X., & Jin, W. (2014). PhaseTank: Genome-wide computational identification of phasiRNAs and their regulatory cascades. *Bioinformatics (Oxford, England)*, 31(2), 284–286.
- Guo, Y., Zhao, S., Li, C.-I., Sheng, Q., & Shyr, Y. (2014). RNAseqPS: A web tool for estimating sample size and power for RNAseq experiment. *Cancer Informatics*, 13, S17688.
- Guo, Y., Zhao, S., Ye, F., Sheng, Q., & Shyr, Y. (2014). MultiRankSeq: Multiperspective approach for RNAseq differential expression analysis and quality control. *BioMed Research International*, 2014.
- Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., ... Nusbaum, C. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28(5), 503–510.
- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... Lieber, M. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8), 1494–1512.

- Han, S., Liang, Y., Li, Y., & Du, W. (2016). LncIdent: A tool for rapid identification of long noncoding RNAs utilizing sequence intrinsic composition and open reading frame information. *International Journal of Genomics*, 2016.
- Han, S., Liang, Y., Ma, Q., Xu, Y., Zhang, Y., Du, W., ... Li, Y. (2019). LncFinder: An integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property. *Briefings in Bioinformatics*, 20(6), 2009–2027.
- Henschel, R., Lieber, M., Wu, L.-S., Nista, P. M., Haas, B. J., & Leduc, R. D. (2012). Trinity RNA-Seq assembler performance optimization. In *Proceedings of the 1st conference of the extreme science and engineering discovery environment: Bridging from the eXtreme to the campus and beyond* (pp. 1–8).
- Homer, N., Merriman, B., & Nelson, S. F. (2009). BFAST: An alignment tool for large scale genome resequencing. *PLoS One*, 4(11), e7767.
- Hu, G., & Zhao, K. (2015). *Identification of intergenic long noncoding RNA by deep sequencing. Epigenetic gene expression and regulation*. Elsevier.
- Illumina (2011). Quality scores for next-generation sequencing. *Technical Note: Informatics*, 31.
- Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M., & Weissman, J. S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature Protocols*, 7(8), 1534–1550.
- Jeanneau, M., Bon, M., & Martin, J. (2010). Meaningful application of the new 454 large scale pyrosequencing technology (Roche GS-FLX 454) to the identification of microsatellites for small-scale research projects. In *15th European Weed Research Society symposium proceedings*, Kaspovar, Hungary (pp. 12–15).
- Jiang, H., & Wong, W. H. (2008). SeqMap: Mapping massive amount of oligonucleotides to the genome. *Bioinformatics (Oxford, England)*, 24(20), 2395–2396.
- Kajal, M., Kaushal, N., & Singh, K. (2019). Identification of novel microRNAs and their targets in *Chlorophytum borivilium* by small RNA and degradome sequencing. *Non-coding RNA Research*, 4(4), 141–154.
- Kaushik, A., Saraf, S., Mukherjee, S. K., & Gupta, D. (2015). miRMOD: A tool for identification and analysis of 5' and 3' miRNA modifications in next generation sequencing small RNA data. *PeerJ*, 3, e1332.
- Khan, A., Goswami, K., Sopory, S. K., & Sanan-Mishra, N. (2017). “Mirador” on the potential role of miRNAs in synergy of light and heat networks. *Indian Journal of Plant Physiology*, 22(4), 587–607.
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4), R36.
- Klambauer, G., Unterthiner, T., & Hochreiter, S. (2013). DEXUS: Identifying differential expression in RNA-Seq studies with unknown conditions. *Nucleic Acids Research*, 41(21), e198.
- Kozomara, A., & Griffiths-Jones, S. (2014). miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42(D1), D68–D73.
- Kukurba, K. R., Zhang, R., Li, X., Smith, K. S., Knowles, D. A., Tan, M. H., ... MacArthur, D. G. (2014). Allelic expression of deleterious protein-coding variants across human tissues. *PLoS Genetics*, 10(5), e1004304.
- Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., ... Lachmann, A. (2016). Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1), W90–W97.
- Kumar, M., Devaux, R., & Herschkowitz, J. (2016). Molecular and cellular changes in breast cancer and new roles of lncRNAs in breast cancer initiation and progression. *Progress in Molecular Biology and Translational Science*, 144, 563–586.
- Kung, J. T., Colognori, D., & Lee, J. T. (2013). Long noncoding RNAs: Past, present, and future. *Genetics*, 193(3), 651–669.
- Kwaepila, N., Burns, G., & Leong, A. S. Y. (2006). Immunohistological localisation of human FAT1 (hFAT) protein in 326 breast cancers. Does this adhesion molecule have a role in pathogenesis? *Pathology*, 38(2), 125–131.
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357.
- Lassmann, T., Hayashizaki, Y., & Daub, C. O. (2011). SAMStat: Monitoring biases in next generation sequencing data. *Bioinformatics (Oxford, England)*, 27(1), 130–131.
- Le Thomas, A., Stuwe, E., Li, S., Du, J., Marinov, G., Rozhkov, N., ... Toth, K. F. (2014). Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor processing. *Genes & Development*, 28(15), 1667–1680.
- Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., ... Schatz, M. C. (2016). Third-generation sequencing and the future of genomics. *BioRxiv*, 048603.
- Leng, N., Li, Y., McIntosh, B. E., Nguyen, B. K., Duffin, B., Tian, S., ... Kendziorski, C. (2015). EBSeq-HMM: A Bayesian approach for identifying gene-expression changes in ordered RNA-seq experiments. *Bioinformatics (Oxford, England)*, 31(16), 2614–2622.
- Li, A., Zhang, J., & Zhou, Z. (2014). PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics*, 15(1), 311.
- Li, D., Luo, L., Zhang, W., Liu, F., & Luo, F. (2016). A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. *BMC Bioinformatics*, 17(1), 329.
- Li, G., Bahn, J. H., Lee, J.-H., Peng, G., Chen, Z., Nelson, S. F., & Xiao, X. (2012). Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Research*, 40(13), e104.
- Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5), 589–595.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), 2078–2079.
- Li, R. (2009). *Short oligonucleotide analysis package: SOAPdenovo 1.03*. Beijing: Beijing Genomics Institute.

- Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: Short oligonucleotide alignment program. *Bioinformatics (Oxford, England)*, *24*(5), 713–714.
- Li, W., Feng, J., & Jiang, T. (2011). IsoLasso: A LASSO regression approach to RNA-Seq based transcriptome assembly. *Journal of Computational Biology*, *18*(11), 1693–1707.
- Li, Y., Rao, X., Mattox, W. W., Amos, C. I., & Liu, B. (2015). RNA-seq analysis of differential splice junction usage and intron retentions by DEXSeq. *PLoS One*, *10*(9), e0136653.
- Li, Z., Zhang, Z., Yan, P., Huang, S., Fei, Z., & Lin, K. (2011). RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics*, *12*(1), 540.
- Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., & Zhang, B. (2019). WebGestalt 2019: Gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research*, *47*(W1), W199–W205.
- Lindgreen, S. (2012). AdapterRemoval: Easy cleaning of next-generation sequencing reads. *BMC Research Notes*, *5*(1), 337.
- Lischer, H. E., & Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, *18*(1), 1–12.
- Liu, J., Yu, T., Jiang, T., & Li, G. (2016). TransComb: Genome-guided transcriptome assembly via combing junctions in splicing graphs. *Genome Biology*, *17*(1), 213.
- Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., ... Law, M. (2012). Comparison of next-generation sequencing systems. *BioMed Research International*, 2012.
- Liu, X., Lu, T., Dou, Y., Yu, B., & Zhang, C. (2014). Identification of RNA silencing components in soybean and sorghum. *BMC Bioinformatics*, *15*(1), 4.
- Liu, Y., Zhou, J., & White, K. P. (2014). RNA-seq differential expression studies: More sequence or more replication? *Bioinformatics (Oxford, England)*, *30*(3), 301–304.
- Llave, C., Xie, Z., Kasschau, K. D., & Carrington, J. C. (2002). Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science (New York, N.Y.)*, *297*(5589), 2053–2056.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.
- Ma, T. (2018). *Differential expression and feature selection in the analysis of multiple omics studies*. University of Pittsburgh.
- Mallory, A. C., & Vaucheret, H. (2006). Functions of microRNAs and related small RNAs in plants. *Nature Genetics*, *38*(6), S31–S36.
- Mao, S., Pachter, L., Tse, D., & Kannan, S. (2020). RefShannon: A genome-guided transcriptome assembler using sparse flow decomposition. *PLoS One*, *15*(6), e0232946.
- Marchant, A., Mougél, F., Mendonça, V., Quartier, M., Jacquín-Joly, E., Da Rosa, J., ... Harry, M. (2016). Comparing de novo and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug *Rhodnius prolixus*. *Insect Biochemistry and Molecular Biology*, *69*, 25–33.
- Maretty, L., Sibbesen, J. A., & Krogh, A. (2014). Bayesian transcriptome assembly. *Genome Biology*, *15*(10), 501.
- Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics*, *12*, 671–682.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12.
- Matzke, M., Kanno, T., Daxinger, L., Huettel, B., & Matzke, A. J. (2009). RNA-mediated chromatin-based silencing in plants. *Current Opinion in Cell Biology*, *21*(3), 367–376.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., ... Daly, M. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303.
- Modolo, L., & Lerat, E. (2015). UrQt: An efficient software for the Unsupervised Quality trimming of NGS data. *BMC Bioinformatics*, *16*(1), 137.
- Mohorianu, I., Stocks, M. B., Applegate, C. S., Folkes, L., & Moulton, V. (2017). The UEA small RNA workbench: A suite of computational tools for small RNA analysis. In *MicroRNA detection and target identification*. Springer.
- Morozova, O., & Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, *92*(5), 255–264.
- Moxon, S., Schwach, F., Dalmay, T., Maclean, D., Studholme, D. J., & Moulton, V. (2008). A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics (Oxford, England)*, *24*(19), 2252–2253.
- Müller, R., Weirick, T., John, D., Militello, G., Chen, W., Dimmeler, S., & Uchida, S. (2016). ANGIOGENES: Knowledge database for protein-coding and noncoding RNA genes in endothelial cells. *Scientific Reports*, *6*, 32475.
- Nakano, M., Nobuta, K., Vemaraju, K., Tej, S. S., Skogen, J. W., & Meyers, B. C. (2006). Plant MPSS databases: Signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Research*, *34*(suppl\_1), D731–D735.
- Nakano, T., Suzuki, K., Fujimura, T., & Shinshi, H. (2006). Genome-wide analysis of the ERF gene family in Arabidopsis and rice. *Plant Physiology*, *140*(2), 411–432.
- Naqvi, A. R., Choudhury, N. R., & Haq, Q. M. R. (2011). Small RNA-mediated defensive and adaptive responses in plants. In *Genetics, biofuels and local farming systems*.
- Neilsen, C. T., Goodall, G. J., & Bracken, C. P. (2012). IsomiRs—The overlooked repertoire in the dynamic microRNAome. *Trends in Genetics*, *28*(11), 544–549.
- Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: Present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1620), 20120362.

- Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Vicente, C., Yang, X., Tirado, F., . . . Pascual-Montano, A. (2009). GeneCodis: Interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Research*, *37*(suppl\_2), W317–W322.
- Paicu, C., Mohorianu, I., Stocks, M., Xu, P., Coince, A., Billmeier, M., . . . Moxon, S. (2017). miRCat2: Accurate prediction of plant and animal microRNAs from next-generation sequencing datasets. *Bioinformatics (Oxford, England)*, *33*(16), 2446–2454.
- Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. *PLoS One*, *7*(2), e30619.
- Perteau, M., Perteau, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295.
- Phillips, J. R., Dalmay, T., & Bartels, D. (2007). The role of small RNAs in abiotic stress. *FEBS Letters*, *581*(19), 3592–3597.
- Pop, M., Phillippy, A., Delcher, A. L., & Salzberg, S. L. (2004). Comparative genome assembly. *Briefings in Bioinformatics*, *5*(3), 237–248.
- Prendergast, J. G., Tong, P., Hay, D. C., Farrington, S. M., & Semple, C. A. (2012). A genome-wide screen in human embryonic stem cells reveals novel sites of allele-specific histone modification associated with known disease loci. *Epigenetics and Chromatin*, *5*(1), 6.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Daly, M. J. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575.
- Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., . . . Turner, D. J. (2008). A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, *5*(12), 1005–1010.
- Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., . . . Mikhail, A. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, *530*(7589), 228–232.
- Ragupathy, R., You, F. M., & Cloutier, S. (2013). Arguments for standardizing transposable element annotation in plant genomes. *Trends in Plant Science*, *18*(7), 367–376.
- Reddy, T. E., Gertz, J., Pauli, F., Kucera, K. S., Varley, K. E., Newberry, K. M., . . . Song, L. (2012). Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Research*, *22*(5), 860–869.
- Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., & Bartel, D. P. (2002). Prediction of plant microRNA targets. *Cell*, *110*(4), 513–520.
- Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. *Genomics, Proteomics and Bioinformatics*, *13*(5), 278–289.
- Ribani, A., Schiavo, G., Utzeri, V. J., Bertolini, F., Geraci, C., Bovo, S., & Fontanesi, L. (2018). Application of next generation semiconductor based sequencing for species identification in dairy products. *Food Chemistry*, *246*, 90–98.
- Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., . . . Qian, J. Q. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, *7*(11), 909–912.
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140.
- Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), 1–9.
- Rueda, A., Barturen, G., Lebrón, R., Gómez-Martín, C., Alganza, Á., Oliver, J. L., & Hackenberg, M. (2015). sRNAtoolbox: An integrated collection of small RNA research tools. *Nucleic Acids Research*, *43*(W1), W467–W473.
- Salzberg, S. L., & Yorke, J. A. (2005). Beware of mis-assembled genomes. *Bioinformatics (Oxford, England)*, *21*(24), 4320–4321.
- Sanan-Mishra, N., Tripathi, A., Goswami, K., Shukla, R. N., Vasudevan, M., & Goswami, H. (2018). ARMOUR—A rice miRNA:mRNA interaction resource. *Frontiers in Plant Science*, *9*, 602.
- Schneeberger, K., Ossowski, S., Ott, F., Klein, J. D., Wang, X., Lanz, C., . . . Warthmann, N. (2011). Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proceedings of the National Academy of Sciences of United States of America*, *108*(25), 10249–10254.
- Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, *28*(8), 1086–1092.
- Shabalin, A. A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics (Oxford, England)*, *28*(10), 1353–1358.
- Shao, L., Xing, F., Xu, C., Zhang, Q., Che, J., Wang, X., . . . Chen, L.-L. (2019). Patterns of genome-wide allele-specific expression in hybrid rice and the implications on the genetic basis of heterosis. *Proceedings of the National Academy of Sciences of United States of America*, *116*(12), 5653–5658.
- Shi, X. (2017). *Bayesian modeling for isoform identification and phenotype-specific transcript assembly*. Virginia Tech.
- Shriram, V., Kumar, V., Devarumath, R. M., Khare, T. S., & Wani, S. H. (2016). MicroRNAs as potential targets for abiotic stress tolerance in plants. *Frontiers in Plant Science*, *7*, 817.
- Song, C., Zhang, D., Zheng, L., Zhang, J., Zhang, B., Luo, W., . . . Han, M. (2017). miRNA and Degradome sequencing reveal miRNA and their target genes that may mediate shoot growth in spur type mutant “Yanfu 6.”. *Frontiers in Plant Science*, *8*, 441.
- Song, L., Sabuncuyan, S., & Florea, L. (2016). CLASS2: Accurate and efficient splice variant annotation from RNA-seq reads. *Nucleic Acids Research*, *44*(10), e98.
- Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Akerman, M., Alioto, T., . . . Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, *10*(12), 1177–1184.
- Stocks, M. B., Moxon, S., Mapleson, D., Woolfenden, H. C., Mohorianu, I., Folkes, L., . . . Moulton, V. (2012). The UEA sRNA workbench: A suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics (Oxford, England)*, *28*(15), 2059–2061.
- Studholme, D. J. (2012). Deep sequencing of small RNAs in plants: Applied bioinformatics. *Briefings in Functional Genomics*, *11*(1), 71–85.

- Sun, X., Dong, B., Yin, L., Zhang, R., Du, W., Liu, D., ... Mao, L. (2013). PMTED: A plant microRNA target expression database. *BMC Bioinformatics*, *14*(1), 174.
- Szcześniak, M. W., & Makałowska, I. (2014). miRNEST 2.0: A database of plant and animal microRNAs. *Nucleic Acids Research*, *42*(D1), D74–D77.
- Tang, S., Lomsadze, A., & Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. *Nucleic Acids Research*, *43*(12), e78.
- Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., ... Su, Z. (2017). agriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research*, *45*(W1), W122–W129.
- Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, *31*(1), 46–53.
- Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, *25*(9), 1105–1111.
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., ... Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562–578.
- Tripathi, A., Chacon, O., Lata Singla-Pareek, S., Sopory, S. K., & Sanan-Mishra, N. (2018). Mapping the microRNA expression profiles in glyoxalase overexpressing salinity tolerant rice. *Current Genomics*, *19*(1), 21–35.
- Tripathi, A., Goswami, K., & Sanan-Mishra, N. (2015). Role of bioinformatics in establishing microRNAs as modulators of abiotic stress responses: The new revolution. *Frontiers in Physiology*, *6*, 286.
- Tripathi, A., Goswami, K., Tiwari, M., Mukherjee, S. K., & Sanan-Mishra, N. (2018). Identification and comparative analysis of microRNAs from tomato varieties showing contrasting response to ToLCV infections. *Physiology and Molecular Biology of Plants*, *24*(2), 185–202.
- Valle, L., Serena-Acedo, T., Liyanarachchi, S., Hampel, H., Comeras, I., Li, Z., ... Sadim, M. (2008). Germline allele-specific expression of TGFBR1 confers an increased risk of colorectal cancer. *Science (New York, N.Y.)*, *321*(5894), 1361–1365.
- Vazquez, F., Vaucheret, H., Rajagopalan, R., Lepers, C., Gascioli, V., Mallory, A. C., ... Crété, P. (2004). Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. *Molecular Cell*, *16*(1), 69–79.
- Vezi, F., Cattonaro, F., & Policriti, A. (2011). e-RGA: Enhanced reference guided assembly of complex genomes. *EMBnet journal*, *17*(1), 46–54.
- Voshall, A. (2018). Consensus ensemble approaches improve de novo transcriptome assemblies.
- Voshall, A., & Moriyama, E. N. (2018). Next-generation transcriptome assembly: Strategies and performance analysis. In *Bioinformatics in the era of post genomics and big data* (pp. 15–36).
- Walker, E. J., Zhang, C., Castelo-Branco, P., Hawkins, C., Wilson, W., Zhukova, N., ... Ray, P. (2012). Monoallelic expression determines oncogenic progression and outcome in benign and malignant brain tumors. *Cancer Research*, *72*(3), 636–644.
- Wang, F., Polydore, S., & Axtell, M. J. (2015). More than meets the eye? Factors that affect target selection by plant miRNAs and heterochromatic siRNAs. *Current Opinion in Plant Biology*, *27*, 118–124.
- Wang, J., Vasaikar, S., Shi, Z., Greer, M., & Zhang, B. (2017). WebGestalt 2017: A more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Research*, *45*(W1), W130–W137.
- Wang, K., Hoeksema, J., & Liang, C. (2018). piRNN: Deep learning algorithm for piRNA prediction. *PeerJ*, *6*, e5429.
- Wang, K., Liang, C., Liu, J., Xiao, H., Huang, S., Xu, J., & Li, F. (2014). Prediction of piRNAs using transposon interaction and a support vector machine. *BMC Bioinformatics*, *15*(1), 419.
- Wang, L., Chen, M., Wu, B., Liu, Y. C., Zhang, G. F., Jiang, L., ... Ye, J. (2018). Massively parallel sequencing of forensic STR s using the ion Chef™ and the ion S5™ XL systems. *Journal of Forensic Sciences*, *63*(6), 1692–1703.
- Wang, T., & Nabavi, S. (2018). SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data. *Methods (San Diego, Calif.)*, *145*, 25–32.
- Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63.
- Wegrzyn, J. L., Lee, J. M., Tearse, B. R., & Neale, D. B. (2008). TreeGenes: A forest tree genome database. *International Journal of Plant Genomics*, *2008*.
- Weick, E.-M., & Miska, E. A. (2014). piRNAs: From biogenesis to function. *Development (Cambridge, England)*, *141*(18), 3458–3471.
- West, M. A., Kim, K., Kliebenstein, D. J., Van Leeuwen, H., Michelmore, R. W., Doerge, R., & Clair, D. A. S. (2007). Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics*, *175*(3), 1441–1450.
- Wilhelm, B. T., & Landry, J.-R. (2009). RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. *Methods (San Diego, Calif.)*, *48*(3), 249–257.
- Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research*, *7*.
- Wu, L., Mao, L., & Qi, Y. (2012). Roles of dicer-like and argonaute proteins in TAS-derived small interfering RNA-triggered DNA methylation. *Plant Physiology*, *160*(2), 990–999.
- Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)*, *26*(7), 873–881.
- Xiao, T., & Zhou, W. (2020). The third generation sequencing: The advanced approach to genetic diseases. *Translational Pediatrics*, *9*(2), 163.
- Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., ... Wei, L. (2011). KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Research*, *39*(suppl\_2), W316–W322.
- Xie, F., Jones, D. C., Wang, Q., Sun, R., & Zhang, B. (2015). Small RNA sequencing identifies miRNA roles in ovule and fibre development. *Plant Biotechnology Journal*, *13*(3), 355–369.

- Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., . . . Li, S. (2014). SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics (Oxford, England)*, *30*(12), 1660–1666.
- Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B., & Kinzler, K. W. (2002). Allelic variation in human gene expression. *Science (New York, N.Y.)*, *297*(5584), 1143.
- Yang, X., & Li, L. (2011). miRDeep-P: A computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics (Oxford, England)*, *27*(18), 2614–2615.
- Yao, R.-W., Wang, Y., & Chen, L.-L. (2019). Cellular functions of long noncoding RNAs. *Nature Cell Biology*, *21*(5), 542–551.
- Yi, X., Zhang, Z., Ling, Y., Xu, W., & Su, Z. (2014). PNRD: A plant non-coding RNA database. *Nucleic Acids Research*, *43*(D1), D982–D989.
- Yoshikawa, M. (2013). Biogenesis of trans-acting siRNAs, endogenous secondary siRNAs in plants. *Genes & Genetic Systems*, *88*(2), 77–84.
- Yu, D., Meng, Y., Zuo, Z., Xue, J., & Wang, H. (2016). NATpipe: An integrative pipeline for systematical discovery of natural antisense transcripts (NATs) and phase-distributed nat-siRNAs from de novo assembled transcriptomes. *Scientific Reports*, *6*, 21666.
- Yu, L., Fernandez, S., & Brock, G. (2017). Power analysis for RNA-Seq differential expression studies. *BMC Bioinformatics*, *18*(1), 234.
- Yu, R., Jih, G., Iglesias, N., & Moazed, D. (2014). Determinants of heterochromatic siRNA biogenesis and function. *Molecular Cell*, *53*(2), 262–276.
- Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829.
- Zhang, C., Li, G., Zhu, S., Zhang, S., & Fang, J. (2013). tasiRNADB: A database of ta-siRNA regulatory pathways. *Bioinformatics (Oxford, England)*, *30*(7), 1045–1046.
- Zhang, M., Sun, H., Fei, Z., Zhan, F., Gong, X., & Gao, S. (2014). Fastq\_clean: An optimized pipeline to clean the Illumina sequencing data with quality control. In *2014 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 44–48). IEEE.
- Zhang, X., Xia, J., Lii, Y. E., Barrera-Figueroa, B. E., Zhou, X., Gao, S., . . . Leung, C. (2012). Genome-wide analysis of plant nat-siRNAs reveals insights into their distribution, biogenesis and function. *Genome Biology*, *13*(3), R20.
- Zhang, Y., Wang, X., & Kang, L. (2011). A k-mer scheme to predict piRNAs and characterize locust piRNAs. *Bioinformatics (Oxford, England)*, *27*(6), 771–776.
- Zhang, Y., Zang, Q., Xu, B., Zheng, W., Ban, R., Zhang, H., . . . Li, A. (2016). IsomiR Bank: A research resource for tracking IsomiRs. *Bioinformatics (Oxford, England)*, *32*(13), 2069–2071.
- Zhang, Z., Yu, J., Li, D., Zhang, Z., Liu, F., Zhou, X., . . . Su, Z. (2010). PMRD: Plant microRNA database. *Nucleic acids research*, *38*(suppl\_1), D806–D813.
- Zheng, L.-L., Li, J.-H., Wu, J., Sun, W.-J., Liu, S., Wang, Z.-L., . . . Qu, L.-H. (2015). deepBase v2. 0: Identification, expression, evolution and function of small RNAs, lncRNAs and circular RNAs from deep-sequencing data. *Nucleic Acids Research*, *44*(D1), D196–D202.
- Zheng, Q., & Wang, X.-J. (2008). GOEAST: A web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Research*, *36*(suppl\_2), W358–W363.
- Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., . . . Chanda, S. K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*, *10*(1), 1–10.