Chapter 10
A Survey of Bioinformatics-Based Tools in RNA-Sequencing (RNA-Seq) Data Analysis

Pallavi Gaur and Anoop Chaturvedi

Abstract The capability of next-generation sequencing can be understood by one of its techniques like RNA sequencing (RNA-Seq) that deals with the transcriptome complexity in a powerful and cost-effective way. This technique has emerged as a revolutionary tool with high sensitivity and accuracy over old techniques. Additionally, this technique also gives unprecedented ability to detect novel mRNA transcripts as well as ncRNAs and analyze alternative splicing. Being a high-throughput sequencing technique, it poses a great demand for bioinformatics-based analysis of the generated data. Here, we explain how RNA-Seq data can be analyzed, discuss its challenges, and provide an overview of the data analysis methods/tools. We discuss strategies for quality check, mapping, and differential expression in transcriptomic data along with discussions on lately developed strategies for alternative splicing and isoform quantification. We also mention some useful R/Bioconductor packages for aforementioned tasks.

Keywords RNA-Seq • Mapping • Differential expression • Bioconductor • Galaxy

10.1 Introduction

RNA-Seq is one of the most advanced techniques which use the platform of high-throughput sequencing (HTS) also called the next-generation sequencing (NGS) technologies to decipher the transcriptome. Transcriptome comprises the complete set of transcripts in a tissue, organism, or a specific cell for a given physiological

P. Gaur (✉)
Center of Bioinformatics, Institute of Inter Disciplinary Studies, Nehru Science Center, University of Allahabad, Allahabad 211002, Uttar Pradesh, India
e-mail: palbioinfor@gmail.com

A. Chaturvedi
Department of Statistics, Nehru Science Center, University of Allahabad, Allahabad 211002, Uttar Pradesh, India
e-mail: anoopchaturv@gmail.com

D.-Q. Wei et al. (eds.), *Translational Bioinformatics and Its Application*, Translational Medicine Research, DOI 10.1007/978-94-024-1045-7_10

condition. Transcripts include protein-coding messenger RNA (mRNA) and non-coding RNA like ribosomal RNA (rRNA), transfer RNA (tRNA), and other ncRNAs (Lindberg and Lundeberg 2010; Okazaki et al. 2002). RNA-Seq basically helps us in looking at the regions of genome being transcribed in a sample and quantifying the expression of such transcripts. Transcriptome has the tendency to vary with different physiological conditions that make transcriptomics a significant field of study, thus turning RNA-Seq a powerful tool for dissecting and understanding many biological phenomena like underlying mechanism and pathways controlling disease initiation, development, and progression.

Over the years, several technologies have come to the existence to study transcriptome, but lately developed RNA-Seq has the ability to characterize the transcriptome in a more global and relatively better way than microarrays and other traditional strategies. RNA-Seq uses cDNA sequencing, from RNA sample of interest (Wilhelm et al. 2008). Basically, RNA-Seq starts by library construction, followed by sequencing on a specific NGS platform and subsequent bioinformatic analysis. In a nutshell, library construction requires isolation of RNA which is randomly fragmented into smaller pieces, followed by reverse transcription. Reverse transcription converts RNA fragments into cDNA with ligation of adapter sequences to either one or both ends for amplification. Fragmentation of RNA can be done prior to reverse transcription, or reverse transcription can be done first followed by cDNA fragmentation (Roberts et al. 2011; Wang et al. 2009). This choice plays an important role because it mostly causes a bias in final results. Especially, cDNA fragmentation generates an under-representation of the 5′ of the transcripts, while RNA fragmentation allows a better representation of the transcript body although somehow may end up in delivering depleted transcript end (Mortazavi et al. 2008). Basic steps and strategy executed by RNA-sequencing experiment are almost the same for every platform which is shown in Fig. 10.1.

Fragment size selection and priming the sequence reaction along with the above steps can vary with the implementation of the protocol and introduce some technical biases in the resulting data. The final sequencing step relies on the NGS platform like 454 pyrosequencing system (a subsidiary of Roche), the AB SOLiD system (Life Technologies), and the Illumina Genome Analyzer (Illumina) (Liu et al. 2012; Marguerat and Bahler 2010; Ansorge 2009), each having its own library construction method. Both the 454 and the SOLiD systems employ an innovative emulsion polymerase chain reaction (emulsion PCR) method for clonal amplification. In emulsion PCR, the cDNA fragments from a library are attached to beads followed by compartmentalization in the aqueous droplets called water-in-oil emulsion. This way, each droplet contains a single DNA molecule as well as the segregated template fragments. These fragments are then amplified in very small emulsified aqueous droplets (Dressman et al. 2003).

The Illumina Genome Analyzer (GA) utilizes the strategy of “bridge PCR” amplification where the adapter-linked single-stranded fragments of cDNA are immobilized on a glass slide by oligonucleotide hybridization in a bridging way, followed by clonal PCR amplification (Fedurco et al. 2006). A population of identical templates is resulted from clonal amplification, but it may introduce a

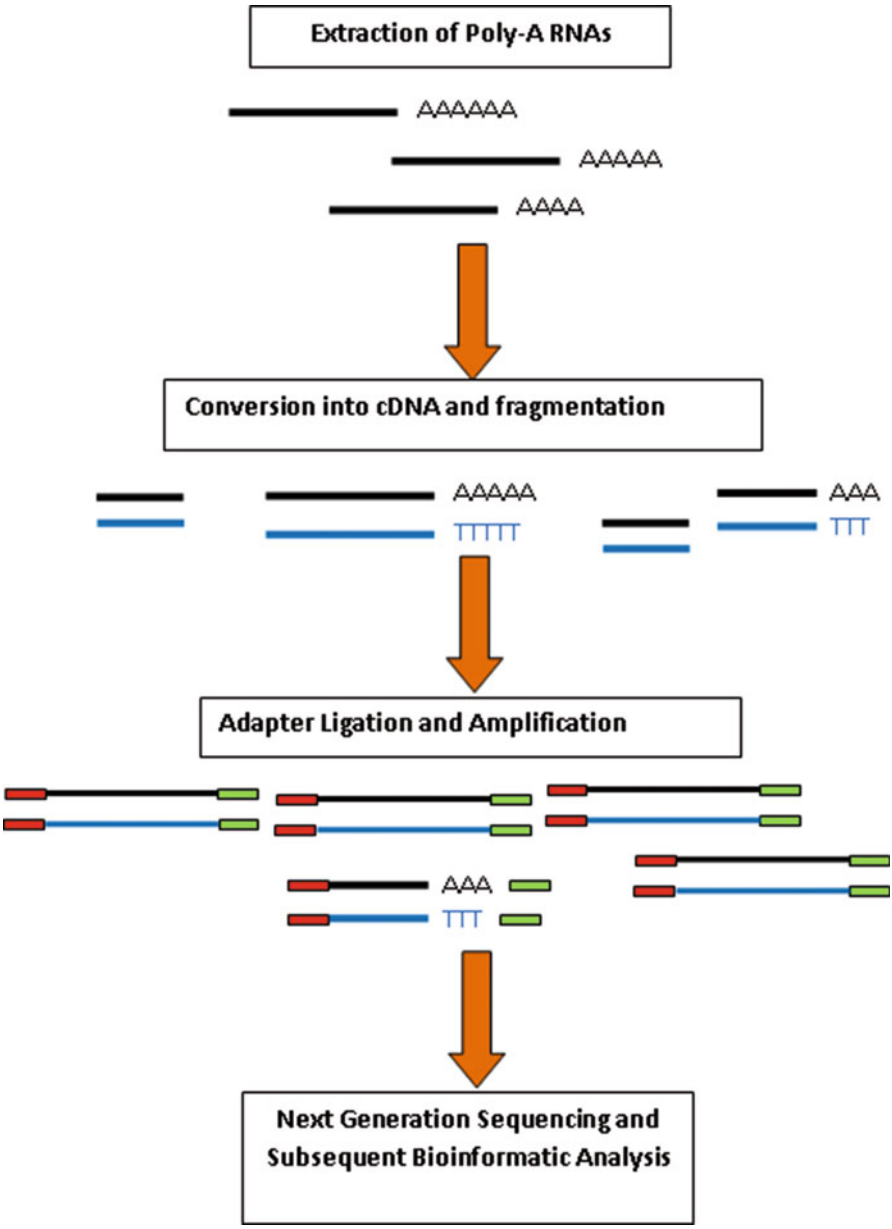

Fig. 10.1 A basic layout of RNA-sequencing experiment

bias in the RNA-Seq result due to PCR artifacts. That is why performances on different biological replicates are needed to determine whether the same short reads are present in different replicates (Wang et al. 2009). Different NGS platforms use different sequencing strategies (Metzker 2009), and several reviews can be found

describing details including mechanisms and comparisons of these NGS technologies (Liu et al. 2012a; Metzker 2009; Shendure and Ji 2008; Ansorge 2009). Sequencing can produce single-end or paired-end reads. In paired-end sequencing, a fragment is sequenced from both ends, while in single-end sequencing, only one end is used. Having the advantage of sequencing from both ends, paired-end sequencing generates data of comparatively high quality.

Since the advent of RNA-Seq in 2008, it has emerged as a superior technique to study transcriptome over traditional methods which were either hybridization (microarray) or sequence based (SAGE, CAGE). Being superior in resolution at the single-base level, this technique can effectively measure the expression level of thousands of genes simultaneously in addition to information on alternative splicing, unannotated exons, allele-specific expression (Heap et al. 2010), microRNAs, variants like SNPs (Quinn et al. 2013), and novel transcripts (gene or noncoding RNAs). Additionally, many significant phenomena such as detection of differential alternative splicing and isoform abundance can be studied in detail with RNA-Seq technique (Park et al. 2013).

Although RNA-Seq is clearly more informative and advantageous, the data produced by this technique are still complex and huge. NGS platforms generate high-throughput data in the form of millions of short sequences termed as "reads." These reads are associated with their base-call quality scores that indicate the reliability of each base call. The length of these short reads depends on the type of NGS platform used for sequencing, but generally they fall within a length of 25–450 bp. The resulting reads are categorized into three types: exonic reads, exon–intron junction reads, and poly(A) reads (Wang et al. 2009). The analysis of this kind of data is not a straightforward task and is usually a bottleneck to deal with. Fortunately, continuous progress in the area of bioinformatics has eased the way to deal with RNA-Seq data. There are now various bioinformatic tools/software, web servers, as well as whole pipelines to tackle and analyze RNA-Seq data. Also, various strategies applicable to RNA-Seq data analysis can be implemented in Bioconductor (Huber et al. 2015; Gentleman et al. 2004) through statistical language "R" (https://www.r-project.org). Bioconductor is free, is open-source, and can deal with analysis of not only RNA-Seq data but other high-throughput genomic data as well. Bioconductor basically works on the basis of different "packages" dedicated to different types of tasks. There are many Bioconductor packages dedicated to the whole RNA-Seq data analysis executable with even a little proficiency in R. Many tools can be combined for analysis of RNA-Seq data, and researchers may form their own custom data analysis pipelines according to their objectives.

Bioinformatic analysis of RNA-Seq data can be divided into several stages. The very first step is experiment/technology dependent, and choice of the methods for downstream analysis is made on the basis of the type of experiment. During sequencing only, the first step of bioinformatic analysis gets started with the transformation of fluorescent measurements into associated nucleotide bases with their quality scores. Base quality score is usually a value representing the confidence of the called bases. The final output of this base-calling step is the short reads (raw data) in FASTQ (FAST-All with quality score) format. The next task is to map

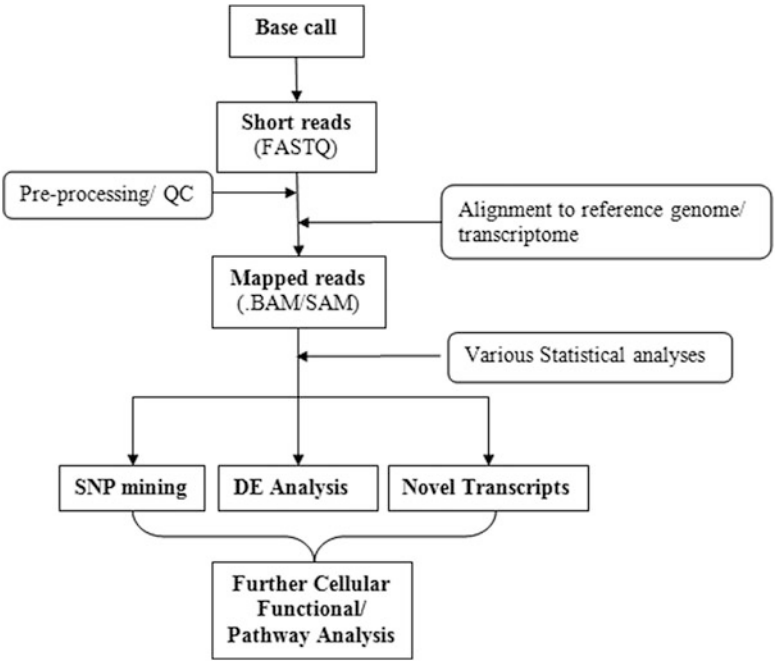

Fig. 10.2 A usual flow chart of bioinformatics-based analysis of RNA-Seq data

these short reads to reference genome (or transcriptome in case of transcriptomic data) in case it's already available or otherwise firstly assemble them de novo. After mapping, further downstream analysis may proceed according to research goals, though a usual work flow of bioinformatics-based analysis associated with RNA-Seq data is shown in the flowchart (Fig. 10.2). During the analysis, different tools/software or strategies may be applied at different steps.

It would not be inappropriate to say that RNA sequencing has a variety of different applications and data analysis strategies depending on the organism under study and research objectives. RNA-Seq has the power of identifying transcripts and quantifying gene expression which is the key to decipher more knowledge on the relationship between genome and proteome. Elucidating RNA isoform expression, alternative splicing, and ncRNA levels are other applications of RNA-Seq having great importance in molecular biology.

10.2 Data Format, Quality Check, and Preprocessing

Raw reads (FASTQ format) obtained after the base-calling step contain nucleotides associated with quality scores. Although different NGS platforms have their own methods of base calling (base-calling software) to evaluate base quality, various

third party groups have also put efforts in developing base-calling methods. The most profitable and notable example is the enhanced ABI base caller, Phred, which played an important role in the Human Genome Project (Ewing and Green 1998; Ewing et al. 1998). Nowadays, most NGS platforms provide the user with a Phred-like score value (Ewing et al. 1998) for base quality evaluation which is based on a logarithmic scale encoding the probability of error in the corresponding base call. This base-calling step is particularly important because its accuracy affects the downstream analysis. The resulting format of base-calling algorithm, i.e., FASTQ, is a FASTA (FAST-All) standard format of biological sequences like format but comes with associated quality score for each nucleotide, usually Phred score.

Reads may be represented in other formats like FASTA and standard flowgram format (SFF) that may be converted to one another, but generally FASTQ format is the most frequent one that can be used as input in many applications. FASTQ files may be so huge in size and also consist of contaminations that need to be eliminated before downstream analysis because contaminated input directly affects the outcome. Preprocessing of data is thus a very important and necessary step before jumping onto the downstream analysis. Preprocessing includes steps like checking the Phred scores, length of reads per base, and read quality and trimming the reads to remove adapters, low-quality sequences, duplicate sequences, and Ns (means no base assigned during the base call). Various available preprocessing tools may be in the form of stand-alone software or accessed with different whole data analysis pipelines, web servers like Galaxy (https://galaxyproject.org/), language platforms like R/Bioconductor, or simply based on command lines.

Some popular tools for quality check and preprocessing of RNA-Seq data are FastQC (Andrews 2010) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit), Cutadapt (https://cutadapt.readthedocs.org/en/stable) (Martin 2011), and Trimmomatic (http://www.usadellab.org/cms/index.php?page=trimmomatic) (Bolger et al. 2014). These tasks are also achievable through some R/Bioconductor packages like "ShortRead" (Morgan et al. 2009). We present a list of some recently developed tools for data quality check and preprocessing (Table 10.1).

10.3 Mapping

Mapping is the most important step in way of analyzing any NGS data. "Mapping" makes each read correspond to a particular position in genome/transcriptome. Since RNA-Seq data may produce reads either from single exon without accessing the exon-exon boundary (unspliced) or from a pair of exon where a read would span the intronic region (spliced), the mapping strategy demands a deeper lookout. If we empirically align the RNA-Seq reads using methods like Burrows–Wheeler transform, we have to consider both the aligned and unaligned reads. Fully aligned reads may be unspliced, but the reads which fail to align may be truly spliced reads spanning an intron. Today, we have many aligners for NGS data using different

Table 10.1 List of recently developed tools/software for data QC and preprocessing

Tool/package	A brief introduction	Input	References
Category: data QC			
AuPairWise	Implemented in R scripts. Measures RNA-Seq replicability by modeling the effects of noise	Expression data	Ballouz and Gillis (2016)
ClinQC	Analysis pipeline. Analyzes both; Sanger and NGS data	Raw reads in any native file format of their sequencing platforms	Pandey et al. (2016)
SinQC	Software tool. Detects technical artifacts in single-cell RNA-seq. Python and R based. R package – ROCR	Gene expression patterns	Jiang et al. Jiang et al. (2016)
TIN (transcript integrity number)	Based on python. Measures RNA degradation	RNA-Seq datasets	Wang et al. (2016)
dupRadar	R package for plotting and analyzing duplication rates dependent on expression levels	BAM file with mapped and duplicate marked reads and a gene model in GTF format	Sayols and Klein (2015)
HTSeq	Python script-based tool	FASTQ, BAM	Anders et al. (2015)
mRIN	Perl- and R-based package. Assess mRNA integrity directly from RNA-Seq data	Coverage profile	Feng et al. (2015)
NOISeq	Bioconductor package. Includes modeling noise distribution of count	Raw and mapped data	Tarazona et al. (2015)
Qualimap 2	Java- and R-based GUI as well as command line interface. Supports multi-sample QC	BAM/SAM, GTF/GFF/BED and read counts table	Okonechnikov et al. (2015)
Rcorrector	Corrects error for Illumina RNA-Seq reads (k-mer-based method). Written in C, C++, and Perl	k-mers based on input reads and counts	Song and Florea (2015)
deepTools	Galaxy-based server	BAM, SAM	Ramirez et al. (2014)
FIXSEQ	R based. Corrects over-dispersed read-count distribution	Read counts	Hashimoto et al. (2014)
QuaCRS	An integrated quality control pipeline for RNA-Seq data. Command line interface	FASTQ, BAM, additional metadata	Kroll et al. (2014)
BlackOPs	Blacklist mismapping in RNA-Seq. Written in Perl	Aligned data	Cabanski et al. (2013)

(continued)

Table 10.1 (continued)

Tool/package	A brief introduction	Input	References
GeneScissors	Detects and corrects spurious transcriptome features leading misalignment. Written in C++, Python, and BamTools	Can be added to any standard pipeline before mapping	Zhang et al. (2013)
HTQC	Toolkit implemented in C++. For graphics – Perl is used	FASTQ	Yang et al. (2013)
IDCheck	RNA-Seq sample identity check	BAM	Huang et al. (2013)
Kraken	Tool package. Pipeline written in Perl and R	FASTQ	Davis et al. (2013)
SEECER	Command line interface. Uses HMMs. Applicable to de novo RNA-Seq	Raw reads	Le et al. (2013)
BM-Map	Software package. Allocates multireads in RNA-Seq data. C++ based	SAM	Yuan et al. (2012)
RSeQC	Python-script-based package. Visualization facilitated through genome browsers like UCSC, IGB, IGV and also using R scripts	SAM, BAM, FASTA, BED or chromosome size file	Wang et al. (2012)
RNA-SeQC	Java based (no installation required). Also integrated in "GenePattern" web interface	One/more BAM	Deluca et al. (2012)
ArrayExpressHTS/AEHTS	R/Bioconductor-based pipeline	Raw reads	Goncalves et al. (2011)
BIGpre	Stand-alone/integrated in Galaxy	FASTQ	Zhang et al. (2011)
NGSQC	Cross platform QC analysis pipeline	FASTQ (IlluQC) or FASTA (454QC)	Dai et al. (2010)
SAMStat	C language-based tool package	SAM, BAM, FASTA, FASTQ	Lassmann et al. (2010)
Category: Trimmers and adapter removers			
ADEPT	Written in Perl5.8. Command line based	One or more FASTQ files	Feng et al. (2016)
Cookiecutter	k-mer-based algorithm. Command line based. Implemented in C++	One or more FASTQ files and a list of k-mers (user provided or cookiecutter generated from FASTA)	Starostina et al. (2015)
NxTrim	For Illumina Nextera Mate Pair (NMP) reads, Command line interface	Raw reads	O'Connell et al. (2015)

(continued)

Table 10.1 (continued)

Tool/package	A brief introduction	Input	References
PEAT	Specifically for paired-end sequencing. Command line interface	FASTQ, no adapter sequence required	Li et al. (2015b)
leeHom	Based on Bayesian maximum a posteriori probability approach. Command-line-based package	One or more FASTQ files, unaligned BAM, adapter sequence	Renaud et al. (2014)
ngsShoRT	Software package written in Perl	FASTQ or Illumina's native QSEQ format	Chen et al. (2014)
QTrim	Stand-alone command line based (python version) as well as a web interface	FASTQ or a FASTA file with its associated quality file (.qual)	Shrestha et al. (2014)
Skewer	"Bit-masked k-difference matching algorithm" based	FASTQ	Jiang et al. (2014)
AlienTrimmer	Command line based	One or more FASTQ files	Criscuolo and Brisse (2013)
NGS QC Toolkit	Implemented in Perl. Command line based, web based	FASTQ, FASTA	Patel and Jain (2012)

Most of the tools shown in table are attributed to RNA-Seq data, but some lately developed tools for NGS data QC and preprocessing are also included in the table. Many data QC tools given in the table are not only limited to raw data QC but to advance stages also like mapping. A brief about basic property of each tool is also included in the table

approaches like seed based (e.g., SHRiMP2;David et al. 2011) , BFAST (Homer et al. 2009), SeqMap (Jiang and Wong 2008), CUSHAW3 (Liu et al. 2014), SOAP (Li et al. 2008a), MAQ (Li et al. 2008b), STAMPY (Lunter and Goodson 2011) or hash based (e.g., MOSAIK; Lee et al. 2014), and HIVE hexagon (Santana et al. 2014). Additionally, a popularly used algorithm in data compression technique, the Burrows–Wheeler transform (BWT), also contributes in providing some excellent mapping tools like BWA (Li and Durbin 2009d), SOAP2 (Li et al. 2009a), and Bowtie (Langmead 2010). Several tools such as TopHat (Trapnell et al. 2009), STAR (Dobin et al. 2013), SpliceMap (Au et al. 2010), and MapSplice (Wang et al. 2010) are available today that perform mapping while considering both the exonic and splicing events.

Mapping refers to locating the short reads onto reference genome/transcriptome which is comparatively feasible with the availability of a reference genome/transcriptome; otherwise a de novo assembly is required to proceed further. Without a reference genome or transcriptome, mapping is not feasible as in such case a de novo assembly of RNA-Seq reads would be required to generate full transcript sequences (Robertson et al. 2010). De novo assembly is usually complex in nature that involves construction of de Bruijn graphs using k-mers. There are many tools for de novo assembly for RNA-Seq data like Trinity (Haas et al. 2013), Velvet (Zerbino and Birney 2008), Bridger (Chang et al. 2015), SOAPdenovo (Li et al.

2010), and Trans-ABySS (Simpson et al. 2009). Here we discuss some useful assemblers for de novo assembly and mappers that are very efficient in RNA-Seq reads mapping.

10.3.1 Trinity

Trinity (Haas et al. 2013) is the first method designed specifically for transcriptome assembly and works on the basis of de Bruijn graphs. It comprises three independent software modules, Inchworm, Chrysalis, and Butterfly, which are used sequentially to produce transcripts. Inchworm assembles the RNA-Seq data into transcript sequences, Chrysalis clusters the Inchworm contigs and constructs complete de Bruijn graphs for each cluster, and then Butterfly processes the individual graphs in parallel to trace the paths of reads within the graph, ultimately reporting full-length transcripts.

10.3.2 Bridger

Bridger is a newer framework for de novo transcript assembly (Chang et al. 2015). It is so named as if to build a bridge between the basic keys of two popular assemblers: Cufflinks (the reference-based assembler (Trapnell et al. 2012)) and Trinity (the de novo assembler (Haas et al. 2013)). It has some advantages over other de novo aligners like it allows the use of different k-mer lengths for different data, while trinity has a fixed k-mer length of 25. It also has a lower false-positive rate and uses less memory and run time compared with Trinity.

On the other hand, the presence of reference genome/transcriptome makes mapping process relatively faster and easier to implement with some web-based/command-line-based tools. In mapping, the problem of multimapping is also usually seen and needs to be taken care of. Generally, mapping utilizes a heuristic first step to find likely candidates followed by local alignment, but alignment is not sufficient for mapping moderate- to large-sized genomes. Thus, the strategy used by most of the aligners/mappers is to somehow enable a fast heuristic method so that the smaller number of local alignments has to be performed. As aforementioned, RNA-Seq mappers should be able to consider the spliced alignment problem, i.e., they should be able to place spliced read across introns and correctly determine exon–intron boundaries. In the present scenario of RNA-Seq research, many aligners work well in this kind of mapping, among which Bowtie2 (Langmead et al. 2009) is a popular one. We discuss a few other tools that have proven their worth.

10.3.3 TopHat

TopHat is a program that aligns RNA-Seq reads to a genome/transcriptome while considering splice junction mapping (Trapnell et al. 2009). It uses the ultrahigh-throughput short read aligner Bowtie and then analyzes the mapping results to identify splice junctions between exons. Using this initial mapping information from Bowtie, TopHat builds a database of possible splice junctions and then again maps the reads against these junctions to confirm them. It runs on Linux and MacOS X and was originally designed to work with reads produced by the Illumina Genome Analyzer, although it is successfully applied with reads from other technologies as well. It also can be implemented in R using some Bioconductor packages as well as on Galaxy server. Moreover, mapping can be visualized through Integrated Genome Viewer (https://www.broadinstitute.org/igv/) (Robinson et al. 2011).

Before performing further downstream analysis, it is also recommended to check the quality of mapping as it greatly influences the downstream analysis. A list of data QC and preprocessing tools capable of checking and processing the data at many stages (including mapping) of data analysis is provided in Table 10.1. Tools like SAMStat (Lassmann et al. 2010) and dupRadar (Sayols and Klein 2015) (R package for QC) are easily accessible and very useful in checking and dealing with mapping quality issues.

10.3.4 STAR

STAR (Spliced Transcripts Alignment to Reference) (Dobin et al. 2013) is one of the important alignment tools that are capable of identifying the alternative splicing junctions in RNA-Seq reads. It is a free, open-source software (under GPLv3 license) that can be downloaded from http://code.google.com/p/rna-star/. It works by indexing the reference genome first, followed by producing a suffix array index to accelerate the alignment step in further processing. STAR has high accuracy like TopHat with comparatively less time consumption. While it can fairly handle single- or paired-end reads, it also increases its accuracy if provided with an annotation (.gtf) file. Advantageously, STAR was not developed as an extension of a short read mapper but a stand-alone C++ code. Being capable of running parallel threads on multi-core systems, STAR is faster in comparison with other tools.

Visualization of mapped reads in a graphical or preferably and advantageously in interactive mode is necessary to closely look at the mapped regions and other factors. There are various tools/software packages such as "SAMtools tview" (Li et al. 2009b), "MapView" (Bao et al. 2009), "Tablet" (Milne et al. 2013), "IGV" (Thorvaldsdóttir et al. 2013), and "Bambino" (Edmonson et al. 2011) that enable the visualization of mapped reads.

In NGS data analysis, the factor of quality control is significant at every single step. Since mapping is the basis for further analysis of data, it is mandatory to check the quality of mapped files to assure the error-free results. Among already available NGS data manipulators like Picard (http://picard.sourceforge.net.) and SAMtools (Li et al. 2009b), some lately developed powerful tools like RseQC and QoRTs assist in quality control, data processing, and management to an excellent level. These tools are included in a package of various utilities that handle the data at different levels.

QoRTs (Hartley and Mullikin 2015) is a fast and portable multifunction toolkit that easily handles cross-comparison of replicates (biological/experimental) and detection of errors, artifacts, and biases. Additionally it can produce count data that can be used in Bioconductor package such as DESeq, DESeq2, and edgeR.

On the other hand, RSeQC (Wang et al. 2012), a comprehensive package of python programs, provides a number of modules to evaluate RNA-Seq data from different aspects. Quality check of raw reads for properties like sequence quality, PCR bias, nucleotide composition bias, and GC bias can be checked with its "basic modules," while "RNA-Seq specific modules" evaluate the quality/status of sequencing saturation of both splice junction detection and expression estimation. RSeQC is written in Python and C and is freely available at http://code.google.com/p/rseqc/.

Mapping is also fundamental in many versatile applications of RNA-Seq like transcript identification and characterization, gene expression quantification, detection of alternatively spliced isoforms, detection of allele-specific expression (ASE), and differential gene expression. Programs like HTSeq-count (Anders et al. 2015) and featureCounts (Liao et al. 2014) use the raw counts of mapped reads for gene quantification. Gene quantification also utilizes a gene transfer format (GTF) file containing the genome coordinates of exons and genes. The number of reads mapped to transcript reference is also the most important information in estimating gene and transcript expression. For expression analysis, only read counts are not sufficient because of other factors like sequence biases, number of reads, and transcript length. These factors are handled by various normalization methods like RPKM (reads per kilobase per million mapped reads) (Mortazavi et al. 2008), FPKM (fragments per kilobase of transcript per million mapped reads) (Trapnell et al. 2010), and TPM (transcripts per million) which are elaborated later in other sections. "Cufflinks" (Trapnell et al. 2012) is a widely used program for estimating transcript level expression from mapping using an EM (expectation–maximization) approach while taking into account biases like nonuniform distribution of reads along the gene length.

The power of identification and quantification of an overall expression of RNAs in a sample is facilitated by RNA-Seq by enabling the genome-wide studies of alternative pre-mRNA splicing which is an important factor to understand the differential expression. Since alternative splicing produces multiple isoforms by skipping or differential joining of exons or introns within a pre-mRNA transcript during transcription (Fig. 10.3), it delivers functional diversity of a gene during posttranscriptional processing and affects gene regulation.

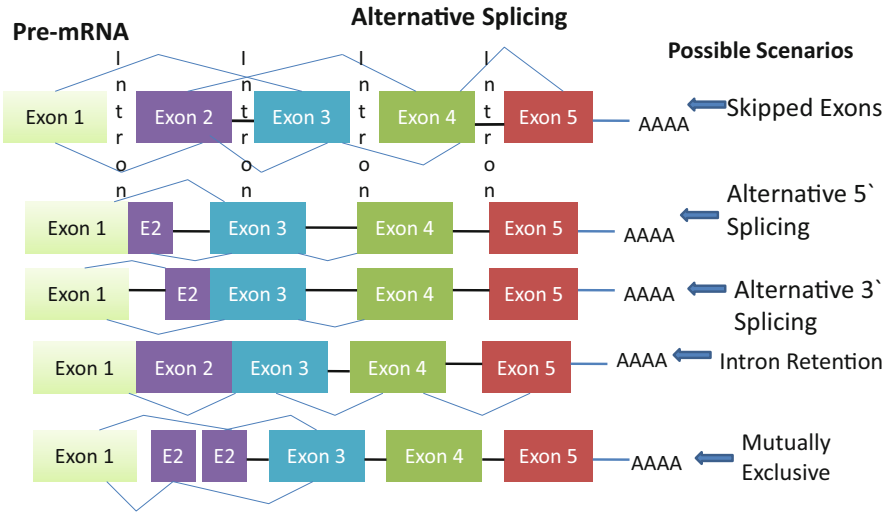

Fig. 10.3 A graphical illustration of alternative splicing event that eventually results in isoforms

Analyzing expression of transcripts at the isoform level is very important in order to understand differential expression. Since many genes may have multiple isoforms, deciphering isoform-specific expression is definitely not straightforward because it is not simple to assign some reads to a particular isoform. The basic approach for dealing with this difficult task was to quantify the transcript isoforms using only those sequences which were unique to particular isoforms (Filichkin et al. 2010). This approach worked on the basis of some already known or predicted transcript isoforms for a given gene that were used to form a set of sequences which in turn could differentiate one isoform from others. Then the mapping of reads to such a set of sequences elaborated the corresponding isoform expression precisely.

Similarly ALEXA-seq (Griffith et al. 2010) method used only those reads that mapped uniquely to one isoform to estimate isoform-specific expression, but these kinds of approaches usually are limited. This is because many isoforms are mostly nonunique or may have minor sequence differences, and also these approaches demand a prior knowledge of precise annotation of splice variants.

The tools related to isoform identification, quantification, abundance estimation, pre-mRNA alternative splicing discovery, and mapping/alignment are already widespread, and the development of new methods is progressing at a very accelerating speed. We present a list (Table 10.2) consisting some recently developed methods/tools dedicated to these tasks along with a brief description of each tool.

Lately, some algorithms like Sailfish, Kallisto, and Salmon have come into existence that use an alignment-free approach to deal with gene/isoform quantification task. These algorithms are considered to be lightweight algorithms that are faster than traditional mapping steps. A succinct overview of all three algorithms is briefed below.

Table 10.2 Recently developed methods/tools for isoform discovery, quantification, abundance estimation, alternative splicing discovery, assembling transcriptome, and alignment of RNA-Seq reads

Tool	A brief description of utility	URL	References
CIDANE	Transcript reconstruction, isoform discovery, and abundance estimation	http://ccb.jhu.edu/software/cidane/.	Canzar et al. (2016)
CLASS CLASS2	Transcriptome assembly. Alternative splicing discovery	http://sourceforge.net/projects/Splicebox	Song and Florea (2013), Song et al. (2016)
Rail-RNA	A cloud-enabled spliced aligner. Analyzes many samples at once. For many samples, Rail-RNA is more accurate than annotation-assisted aligners	http://rail.bio	Nellore et al. (2015)
Rockhopper 2	De novo assembly of bacterial transcriptomes	http://cs.wellesley.edu/~btjaden/Rockhopper	McClure et al. (2013), Tjaden (2015)
JAGuaR	An alignment protocol for RNA-Seq reads. Does not detect novel junctions	http://www.bcgsc.ca/platform/bioinfo/software/jaguar	Butterfield et al. (2014)
MaLTA	Simultaneous transcriptome assembly and quantification from Ion Torrent RNA-Seq data	http://alan.cs.gsu.edu/NGS/?q=malta	Mangul et al. (2014)
HSA	An effective spliced aligner of RNA-Seq reads. Better call rate and efficiency but little less accurate at some attributes	https://github.com/vlcc/HSA	Bu et al. (2013)

10.3.5 Sailfish

Sailfish (Patro et al. 2014) is a free and open-source software, available at http://www.cs.cmu.edu/~ckingsf/software/sailfish. It is a much faster in silico method facilitating the quantification of RNA-isoform abundance by totally avoiding the time-consuming mapping step. Instead of mapping, it inspects k-mers in reads to observe transcript coverage that results in a fast processing of reads. It also maintains the accuracy up to the mark by incorporating an EM procedure that brings a statistical coupling between k-mers. It discards k-mers that overlap inaccurate bases to handle sequencing errors. Overall, it has only a single explicit parameter the k-mer length to rely on. Longer k-mers tend to resolve their origin easier than short k-mers but may be more affected by errors for which Sailfish has implemented an error handling EM procedure. Process wise, Sailfish first builds an index from a FASTA reference transcript file and a chosen k-mer length. Data structures like minimal perfect hash function 9 in the index file play an important role in mapping each k-mer in reference transcript to an identifier in such a way that no two k-mers share an identifier. There is no need to change or rebuild the index unless the reference or the choice of k changes. Next to building index files, the step of quantification is proceeded that takes index and a set of RNA-Seq reads as input

to estimate the isoform abundance, measured in RPKM, KPKM (k-mers per kilobase per million mapped k-mers), and TPM. Sailfish can also be used for non-model organisms in de novo mode. Since Sailfish has an overall parameter of the k-mer counts, it is also computationally efficient that can effectively exploit many CPU cores.

10.3.6 Kallisto

Kallisto (Bray et al. 2016) was developed by Pachter lab with the same lightweight algorithm approach as Sailfish to quantify transcript abundance but improves it with a "pseudoalignment" process. It is a fast software program written mainly in C++. It is considered to be near optimal in speed along with accuracy and tested successfully by its developers in analyzing 30 million unaligned paired-end RNA-Seq reads in less than 5 min on a standard desktop. This software is widely popular because of its accuracy as compared to those of the already existing tools. It does not work on the basis of position in transcript where a read aligns but the compatibility of read with a particular transcript that takes a lot less time than the traditional alignment process.

10.3.7 Salmon

Salmon (Patro et al. 2015) is an open-source software under the GPL v3 license and available at http://combine-lab.github.io/salmon/. Its developers call it a wicked-fast transcript quantification software that requires a set of target transcripts for quantification task and may be run in two modes: the quasi-mapping-based mode and the alignment-based mode. The quasi-mapping-based mode like Sailfish incorporates two phases, indexing and quantification, while the alignment-based mode uses the alignment file (SAM/BAM) provided by the user along with reference transcript FASTA file and does not require indexing.

10.4 Differential Expression

An important application of RNA-Seq technique is to identify genes that change in abundance between conditions, i.e., they differ in counts in different conditions. Differential expression (DE) is simply to compare expression levels of genes between two conditions, e.g., stimulated versus unstimulated or wild type versus mutant or normal versus treated. If there is a statistically significant difference or change in read counts between two conditions, a gene can be affirmed as a differentially expressed gene. The aforementioned steps of data preprocessing

and mapping are mandatory for analysis of differential expression. Also, for differential expression, it is necessary to analyze read-count distributions, typically represented as a matrix N of n × m where N_{ij} is the number of reads assigned to gene in sequencing experiment/condition j. Bioconductor has many packages to support DE analysis of RNA-Seq data. Many packages like DESeq2 (Love et al. 2014), edgeR (Robinson et al. 2010), limma (Ritchie et al. 2015), and baySeq (Hardcastle 2012) have whole RNA-Seq data analysis pipelines which can be of great use. Most of the packages for DE analysis expect input data in the form of matrix of integer values. To prepare the count matrix, SAM/BAM alignment file along with a file specifying the genomic features, e.g., a GFF3 or GTF, can be used. For this, we may use other packages of Bioconductor like Rsubread (Liao et al. 2013) and GenomicAlignments (Lawrence et al. 2013).

Two most popular packages for DE analysis are DESeq2 (Love et al. 2014) and edgeR (Robinson et al. 2010). They are modular in nature that means there are many entry points in the package from where the package can be used. They often give freedom to use an alternative aligner or a different strategy or tool to obtain read counts and then use the package for rest of the analysis. Since there is not any universal standard for DE analysis, it may somewhat be objective oriented and heavily dependent on external data like reference assemblies and annotation. Thus, we can't expect that two different analysis strategies of the same data will end up with the same results, similarity is still expected though.

It is also worth mentioning about the importance of normalization which is a very significant step in the analysis of DE. Normalization is necessary to correct for biases which can arise from technical biases like between-sample differences that denote library size and within-sample gene-specific effects that may be related to gene length and GC-content (Oshlack and Wakefi 2009). There are various normalization methods for DE analysis including Total Count (TC), Upper Quartile (UQ), Median (Med), the DESeq normalization implemented in the DESeq Bioconductor package, Trimmed Mean of M values (TMM) implemented in the edgeR Bioconductor package, Quantile (Q), and RPKM normalization. FPKM normalization is also a popular method and is used by tools like cufflinks (Trapnell et al. 2010). FPKM is analogous to RPKM but does not use read counts.

This overview of DE analysis is superficial and descriptive of basics only used in DE analysis. There are actually a huge number of parameters in each step that can change results. Every step including preprocessing and mapping affects the analysis of subsequent steps. Like other tasks of RNA-Seq data analysis where newer algorithms and tools are making a mark, task of differential expression has also opened up the way for the development of newer and different algorithms/tools. BitSeq (Hensman et al. 2015; Glaus et al. 2012), deGPS (Chu et al. 2015), NOISeq (Tarazona et al. 2015), and XBSeq (Chen et al. 2015) are some of the recently developed tools which are really different in their algorithm and performance and give a broader spectrum to differential expression analysis in RNA-Seq data.

Although in this chapter we elaborate on different approaches and tools for analysis of RNA-Seq data, continuous research in this field has provided us some great whole analysis pipelines to also deal with RNA-Seq data. Since RNA-Seq

technique has unprecedented ability to study transcriptome to a much greater extent than previous technologies, the analyses of ncRNAs have also become more accessible and feasible. Here we succinctly present a list of recently developed pipelines dedicated to RNA-Seq data and also some tools/pipelines dedicated to analyses of ncRNAs obtained through RNA-Seq technique (Table 10.3).

Table 10.3 List of some popular and recently developed pipelines dedicated to whole RNA-Seq and ncRNA data (obtained through RNA-Seq technique) analysis

Tool	A brief introduction of utility	URL	References
Category: RNA-Seq data analysis pipelines			
CANEapp	GUI and an automated server-side analysis pipeline for RNA-Seq	http://psychiatry.med.miami.edu/research/laboratory-of-translational-rna-genomics/CANE-app	Velmeshev et al. (2016)
QuickRNASeq	A pipeline for large-scale RNA-Seq data analyses and visualization	http://sourceforge.net/projects/quickrnaseq/	Shanrong Zhao et al. (2016)
TRAPLINE	Pipeline for RNA sequencing data analysis, evaluation, and annotation	https://usegalaxy.org/u/mwolfien/w/rnaseq-wolfien-pipeline	Wolfien et al. (2016)
BioWardrobe	Integrated pipeline. Analyzes epigenomics and transcriptomic data	https://biowardrobe.com/	Kartashov and Barski (2015)
QuickNGS	Pipeline that analyzes data from multiple NGS projects at a time. Parallel computing resources	http://bifacility.uni-koeln.de/quickngs/web/	Wagle et al. (2015)
RAP	A cloud-computing web application for RNA-Seq analysis	https://bioinformatics.cineca.it/rap/	D'Antonio et al. (2015)
RNAMiner	A multilevel bioinformatics protocol and pipeline for RNA-Seq	http://calla.rnet.missouri.edu/rnaminer/index.html	Li et al. (2015a)
Category: ncRNA analysis tools/pipelines			
isomiR-SEA	Details miRNAs, isomiRs, and conserved miRNA: mRNA interaction. Specialized alignment algorithm	http://eda.polito.it/isomir-sea/	Urgese et al. (2016)
Chimira	An online tool (pipeline) for analyzing large amounts of small RNA-Seq data	http://wwwdev.ebi.ac.uk/enright-dev/chimira/	Vitsios and Enright (2015)
iSRAP	A one-touch integrated small RNA analysis pipeline	http://israp.sourceforge.net/	Quek et al. (2015)
miRA	ncRNA identification tool. Identifies miRNA precursors in plants	https://github.com/mhuttner/miRA	Evers et al. (2015)
mirPRo	A stand-alone pipeline that quantifies known miRNAs and predicts novel miRNAs	http://sourceforge.net/projects/mirpro/	Shi et al. (2015)

(continued)

Table 10.3 (continued)

Tool	A brief introduction of utility	URL	References
miRge	Ultrafast, small RNA-Seq solution pipeline. Decreases computational requirements	http://atlas.pathology.jhu.edu/baras/miRge.html	Baras et al. (2015)
Oasis	Fast and flexible web application. Facilitates online analysis of small-RNA-Seq (smRNA-Seq) data	https://oasis.dzne.de/	Capece et al. (2015)
segmentSeq	Bioconductor package. Identifies robust sets of siRNA precursors	http://www.bioconductor.org/packages/release/bioc/html/segmentSeq.html	Hardcastle (2015), Hardcastle et al. (2012)
sRNAtoolbox	smRNA analysis pipeline. Collection of small RNA research tools	http://bioinfo5.ugr.es/srnatoolbox	Rueda et al. (2015)
SMiRK	Automated pipeline for miRNA analysis	https://github.com/smirkpipeline/SMiRK	Milholland et al. 2015
Tailor	Read aligner for small silencing RNAs. Also captures the tailing events directly from the alignments without extensive post-processing	https://github.com/jhhung/Tailor	Chou et al. (2015)
tDRmapper	t-RNA derived RNA annotation tool. Maps and quantifies tRNA-derived RNAs (tDRs). Includes graphical visualization that facilitates the discovery of novel tRNA.	https://github.com/sararselitsky/tDRmapper	Selitsky and Sethupathy (2015)
YM500v2	A small RNA sequencing (smRNA-Seq) database for human cancer miRNome research	http://ngs.ym.edu.tw/ym500v2/index.php	Cheng et al. (2015), Cheng et al. (2013)
BioVLAB-MMIA-NGS	A whole software pipeline for microRNA-mRNA integrated analysis using high-throughput sequencing data	http://epigenomics.snu.ac.kr/biovlab_mmia_ngs/	Chae et al. (2014)
CAP-miRSeq	Whole pipeline for microRNA sequencing data	http://bioinformaticstools.mayo.edu/research/cap-mirseq/	Sun et al. (2014)
MAGI	Fast microRNA-Seq data analysis in a GPU infrastructure	http://elgar.ucsd.edu/software/magi/	Kim et al. (2014)
mrSNP	Predicts the impact of a SNP in a 3UTR on miRNA binding	http://mrsnp.osu.edu/	Deveci et al. (2014)
piClust	Finds piRNA clusters and transcripts from small RNA-Seq data	http://epigenomics.snu.ac.kr/piclustweb/	Jung et al. (2014)

(continued)

Table 10.3 (continued)

Tool	A brief introduction of utility	URL	References
CoRAL	ncRNA identification tool. Predicts the precursor class of small RNAs present in RNA-sequencing dataset	http://wanglab.pcbi.upenn.edu/coral/	Leung et al. (2013)
ISRNA	Software pipeline designed for storage, visualization, and analysis of small RNA sequencing data	http://omicslab.genetics.ac.cn/resources.php	Luo et al. (2014)
iMir	A modular pipeline for comprehensive analysis of small RNA-Seq data	http://www.labmedmolge.unisa.it/inglese/research/imir	Giurato et al. (2013)
miReader	Detects mature miRNAs directly from next-generation sequencing read data, without any need of reference/genomic sequences	http://scbb.ihbt.res.in/2810-12/miReader.php	Jha and Shankar (2013)
miRDeep	An integrated application tool for miRNA identification from RNA sequencing data	http://sourceforge.net/projects/mirdeepstar/	An et al. (2013)
ShortStack	Processes and analyzes small RNA-Seq data with respect to a reference genome and outputs a comprehensive and informative annotation of all discovered small RNA genes	http://axtell-lab-psu.weebly.com/shortstack.html	Axtell (2013)
SHRiMP2	Software package for aligning genomic reads against a target genome. Works great with small RNA mapping	http://compbio.cs.toronto.edu/shrimp/	David et al. (2011)

10.5 Summary

Today, RNA-Seq is the mainstream tool for analysis of transcriptomes that are so rich in information and progressing day by day. This technique has its wide applications in various areas like clinical diagnostics, pharmacogenomics, and drug development. It can find novel transcripts and identify drug-related genes and microRNAs. Although RNA-Seq technology is still in progressive and developmental stage, yet it has made substantial contributions to our understanding of many transcriptomes from those of simple unicellular organisms to complex mammalian cells, as well as in tissues in normal and disease states. Still, the data from RNA-Seq is complex to analyze and very sensitive to technical biases. This chapter focused mainly with some tools/software for RNA-Seq data analysis and some interesting platforms like R/Bioconductor and Galaxy web server where many of these tools can be accessed and data can be analyzed. It is worth noting that many

tools mentioned in this chapter are not restricted only to RNA-Seq data and may be used for other kinds of NGS data as well. Also, there are several other tools, software, whole analysis pipeline, and statistical strategies for analyzing RNA-Seq data, but they are not discussed here. Still, bioinformatics-based tools are progressing rapidly, and there is a wide opportunity of building new tools and strategies for analyzing RNA-Seq data as well as data derived from other NGS technologies. As NGS technologies are continuously evolving, we can hope for RNA-Seq having more technical and analytical developments with lower cost in the near future.

Acknowledgments The authors would like to thank University Grant Commission, India for the support. The authors express their gratitude to Nimisha Chaturvedi, Dr. Raghvendra Singh, and Swadha Singh for giving valuable suggestions regarding the improvement of this chapter.

References

An J, Lai J, Lehman ML, Nelson CC. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. Nucleic Acids Res. 2013. PMID: 23221645.

Anders S, Pyl PT, Huber W. HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31:166–9.

Andrews S. Fast QC: a quality control tool for high throughput sequence data. 2010. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Ansorge WJ. Next-generation DNA sequencing techniques. N Biotechnol. 2009;25:195–203. Bioinformatics 25:1754–60.

Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNA-seq data by Splice Map. Nucleic Acids Res. 2010;38:4570–8.

Axtell MJ. ShortStack: comprehensive annotation and quantification of small RNA genes. RNA. 2013. PMID: 23610128.

Ballouz S, Gillis J. AuPairWise: a method to estimate RNA-seq replicability through co-expression. bioRxiv. 2016; doi:10.1101/044669.

Bao H, Guo H, Wang J, Zhou R, Lu X, Shi S. MapView: visualization of short reads alignment on a desktop computer. Bioinformatics. 2009. PMID: 19369497.

Baras AS, Mitchell CJ, Myers JR, Gupta S, Weng LC, Ashton JM et al. miRge – a multiplexed method of processing small RNA-Seq data to determine microRNA entropy. PloS one. 2015. PMID: 26571139.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20. doi:10.1093/bioinformatics/btu170.

Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-Seq quantification. Nat Biotechnol. 2016; doi:10.1038/nbt.3519.

Bu J, Chi X, Jin Z. HSA: a heuristic splice alignment tool. BMC Systems Biol. 2013. PMID: 24564867.

Butterfield YS, Kreitzman M, Thiessen N, Corbett RD, Li Y, Pang J et al. JAGuaR: junction alignments to genome for RNA-seq reads. PloS one. 2014. PMID: 25062255.

Cabanski CR, Wilkerson MD, Soloway M, Parker JS, Liu J, Prins JF, et al. BlackOPs: increasing confidence in variant detection through mappability filtering. Nucleic Acids Res. 2013. PMID: 23935067.

Canzar S, Andreotti S, Weese D, Reinert K, Klau GW. CIDANE: comprehensive isoform discovery and abundance estimation. Genome Biol. 2016. PMID: 26831908.

Capece V, Garcia Vizcaino JC, Vidal R, Rahman RU, Pena Centeno T, Shomroni O et al. Oasis: online analysis of small RNA deep sequencing data. Bioinformatics. 2015. PMID: 25701573.

Chae H, Rhee S, Nephew KP, Kim S. BioVLAB-MMIA-NGS: microRNA-mRNA integrated analysis using high throughput sequencing data. Bioinformatics. 2014. PMID: 25270639.

Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, et al. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol. 2015.

Chen C, Khaleel SS, Huang H, Wu CH. Software for pre-processing Illumina next-generation sequencing short read sequences. Source code for biology and medicine. 2014. PMID: 24955109.

Chen HH, Liu Y, Zou Y, Lai Z, Sarkar D, Huang Y, et al. Differential expression analysis of RNA sequencing data by incorporating non-exonic mapped reads. BMC Genomics. 2015; doi:10.1186/1471-2164-16-S7-S14.

Cheng WC, Chung IF, Huang TS, Chang ST, Sun HJ, Tsai CF, et al. YM500: a small RNA sequencing (smRNA-seq) database for microRNA research. Nucleic Acids Res. 2013. PMID: 23203880.

Cheng WC, Chung IF, Tsai CF, Huang TS, Chen CY, Wang SC, et al. YM500v2: a small RNA sequencing (smRNA-seq) database for human cancer miRNome research. Nucleic Acids Res. 2015. PMID: 25398902.

Chou MT, Han BW, Hsiao CP, Zamore PD, Weng Z, Hung JH. Tailor: a computational framework for detecting non-templated tailing of small silencing RNAs. Nucleic Acids Res. 2015. PMID: 26007652.

Chu C, Fang Z, Hua X, Yang Y, Chen E, CowleyJr AW, et al. deGPS is a powerful tool for detecting differential expression in RNA-sequencing studies. BMC Genomics. 2015. doi: 10.1186/s12864-015-1676-0.

Criscuolo A, Brisse S. AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. Genomics. 2013;102:500–6.

Dai M, Thompson RC, Maher C, Contreras-Galindo R, Kaplan MH, Markovitz DM, et al. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics. 2010. PMID: 21143816.

D'Antonio M, D'Onorio De Meo P, Pallocca M, Picardi E, D'Erchia AM, Calogero RA, et al. RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application. BMC Genomics. 2015. PMID: 26046471.

David M, Dzamba M, Lister D, Ilie L, Brudno M. SHRiMP2: sensitive yet practical SHort Read Mapping. Bioinformatics. 2011. PMID: 21278192.

Davis MPA, Dongen SV, Goodger CA, Bartonicek N, Enright AJ. Kraken: A set of tools for quality control and analysis of high-throughput sequence data. Methods. 2013;63(1): 41–9. doi:10.1016/j.ymeth.2013.06.027. PMID 23816787.

Deluca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. 2012;28 (11):1530–2. doi:10.1093/bioinformatics/bts196.

Deveci M, Catalyürek UV, Toland AE. mrSNP: software to detect SNP effects on microRNA binding. BMC Bioinf. 2014. PMID: 24629096.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013. PMID: 23104886.

Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci U S A. 2003;100:8817–22.

Edmonson MN, Zhang J, Yan C, Finney RP, Meerzaman DM, Buetow KH. Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. Bioinformatics. 2011. PMID: 21278191.

Evers M, Huttner M, Dueck A, Meister G, Engelmann JC. miRA: adaptable novel miRNA identification in plants using small RNA sequencing data. BMC Bioinf. 2015. PMID: 26542525.

Ewing B, Green P. Base-calling of automated sequencer traces using Phred II error probabilities. Genome Res. 1998;8(3):186–94.
Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I Accuracy assessment. Genome Res. 1998;8(3):175–85.
Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. 2006;34:e22.
Feng H, Zhang X, Zhang C. mRIN for direct assessment of genome-wide and gene-specific mRNA integrity from large-scale RNA sequencing data. Nat Commun. 2015;6(7816) doi:10.1038/ncomms8816.
Feng S, Lo CC, Li PE, Chain PS. ADEPT, a dynamic next generation sequencing data error-detection program with trimming. BMC Bioinf. 2016; doi:10.1186/s12859-016-0967-z.
Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, et al. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome Res. 2010;20:45–58.
Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5:R80.
Giurato G, De Filippo MR, Rinaldi A, Hashim A, Nassa G, Ravo M, et al. iMir: An Integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by smallRNA-Seq. BMC Bioinf. 2013. PMID: 24330401.
Glaus P, Honkela A, Rattray M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics. 2012. PMID: 22563066.
Goncalves A, Tikhonov A, Brazma A, Kapushesky M. A pipeline for RNA-seq data processing and quality assessment. Bioinformatics. 2011. PMID: 21233166.
Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, et al. Alternative expression analysis by RNA sequencing. Nat Methods. 2010;7:843–7.
Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.
Hardcastle TJ. Discovery of methylation loci and analyses of differential methylation from replicated high-throughput sequencing data. bioRxiv. 2015; doi:10.1101/021436.
Hardcastle TJ. baySeq: eEmpirical Bayesian analysis of patterns of differential expression in count data. R package version 2.8.0. 2012.
Hardcastle TJ, Kelly KA and Baulcombe DC. Identifying small interfering RNA loci from high-throughput sequencing data. Bioinformatics. 2012. PMID: 22171331.
Hartley SW, Mullikin JC. QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. BMC Bioinf. 2015; doi:10.1186/s12859-015-0670-5.
Hashimoto TB, Edwards MD, Gifford DK. Universal count correction for high-throughput sequencing. PLoS Comput Biol. 2014. PMID: 24603409.
Heap GA, Yang JHM, Downes K, Healy BC, Hunt KA, et al. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. Hum Mol Genet. 2010;19:122–34.
Hensman J, Papastamoulis P, Glaus P, Honkela A, Rattray M. Fast and accurate approximate inference of transcript expression from RNA-seq data. Bioinformatics. 2015. PMID: 26315907.
Homer N, Merriman B, Nelson SF. BFAST: an alignment tool for large scale genome resequencing. PLoS One. 2009;4:e7767.
Huang J, Chen J, Lathrop M, Liang L. A tool for RNA sequencing sample identity check. Bioinformatics. 2013. PMID: 23559639.
Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12(2):115–21.
Jha A, Shankar R. miReader: discovering novel miRNAs in species without sequenced genome. PloS one. 2013. PMID: 23805282.

Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics. 2008;24:2395–6.
Jiang H, Lei R, Ding SW, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinf. 2014. PMID: 24925680.
Jiang P, Thomson JA, Stewart R. Quality Control of Single-cell RNA-seq by SinQC. Bioinformatics. 2016; doi:10.1093/bioinformatics/btw176.
Jung I, Park JC, Kim S. piClust: a density based piRNA clustering algorithm. Comput Biol Chem. 2014. PMID: 24656595.
Kartashov AV, Barski A. BioWardrobe: an integrated platform for analysis of epigenomics and transcriptomics data. Genome Biol. 2015. PMID: 26248465.
Kim J, Levy E, Ferbrache A, Stepanowsky P, Farcas C, Wang S, et al. MAGI: a Node.js web service for fast MicroRNA-Seq analysis in a GPU infrastructure. Bioinformatics. 2014. PMID: 24907367.
Kroll KW, Mokaram NE, Pelletier AR, Frankhouser DE, Westphal MS, Stump PA, et al. Quality Control for RNA-Seq (QuaCRS): an integrated quality control pipeline. Cancer Inf. 2014. PMID: 25368506.
Langmead B. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinf Chapter 11, Unit 11.7. 2010.
Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
Lassmann T, Hayashizaki Y, Daub CO. SAMStat: monitoring biases in next generation sequencing data. Bioinformatics. 2010;27(1):130–1. doi:10.1093/bioinformatics/btq614. PMID 21088025.
Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan M, Carey V. Software for computing and annotating genomic RANGES. PLoS Comput Biol 2013;9.
Le HS, Schulz MH, McCauley BM, Hinman VF, Bar-Joseph Z. Probabilistic error correction for RNA sequencing. Nucleic Acids Res. 2013. PMID: 23558750.
Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PloS one. 2014. PMID: 24599324.
Leung YY, Ryvkin P, Ungar LH, Gregory BD, Wang LS. CoRAL: predicting non-coding RNAs from small RNA-sequencing data. Nucleic Acids Res. 2013. PMID: 23700308
Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009d;25(14):1754–60.
Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008a;24:713–4.
Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008b. PMID: 18714091.
Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, et al. SOAP2: an improved ultra-fast tool for short read alignment. Bioinformatics. 2009a;25:1966–7.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009b. PMID: 19505943.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. 1000 genome project data processing subgroup. 2009c.
Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and de novo assembly of the giant panda genome. Nature. 2010;463(7279):311–7.
Li J, Hou J, Sun L, Wilkins JM, Lu Y, Niederhuth CE, et al.. From gigabyte to kilobyte: A bioinformatics protocol for mining large RNA-Seq transcriptomics data. PloS one. 2015a. PMID: 25902288.
Li YL, Weng JC, Hsiao CC, Chou MT, Tseng CW, Hung JH. PEAT: an intelligent and efficient paired-end sequencing adapter trimming algorithm. BMC Bioinf. 2015b. PMID: 25707528
Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 2013;41:e108.

Liao Y, Smyth GK, Shi W. Feature counts: an efficient general-purpose read summarization program. Bioinformatics. 2014;30:923–30.
Lindberg J, Lundeberg J. The plasticity of the mammalian transcriptome. Genomics. 2010;95:1–6.
Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012:251364.
Liu Y, Popp B, Schmidt B. CUSHAW3: sensitive and accurate base-space and color-space short-read alignment with hybrid seeding. PloS one. 2014. PMID: 24466273.
Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersionfor RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011. PMID: 20980556.
Luo GZ, Yang W, Ma YK, Wang XJ. ISRNA: an integrative online toolkit for short reads from high-throughput sequencing data. Bioinformatics. 2014. PMID: 24300438.
Mangul S, Caciula A, Al Seesi S, Brinza D, Măndoiu I, Zelikovsky A. Transcriptome assembly and quantification from Ion Torrent RNA-Seq data. BMC Genomics. 2014. PMID: 25082147.
Marguerat S, Bähler J. RNA-seq: from technology to biology. Cell Mol Life Sci. 2010;67:569–79.
Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2.
McClure R, Balasubramanian D, Sun Y, Bobrovskyy M, Sumby P, Genco CA, et al. Computational analysis of bacterial RNA-Seq data. Nucleic Acids Res. 2013. PMID: 23716638.
Metzker ML. Sequencing technologies – the next generation. Nat Rev Genet. 2009;11:31–46.
Milholland B, Gombar S, Suh Y. SMiRK: an automated pipeline for miRNA analysis. J Genomics. 2015. PMID: 26613105.
Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, et al. Using Tablet for visual exploration of second-generation sequencing data. Brief Bioinf. 2013. PMID: 22445902.
Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. Short read: a Bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics. 2009;25:2607–8.
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods. 2008. PMID: 18516045.
Nellore A, Collado-Torres L, Jaffe AE, Morton J, Pritt J, Alquicira-Hernández J, et al. Rail-RNA: Scalable analysis of RNA-seq splicing and coverage. bioRxiv. 2015. doi:10.1101/019067.
O'Connell J, Schulz-Trieglaff O, Carlson E, Hims MM, Gormley NA, Cox AJ. NxTrim: optimized trimming of Illumina mate pair reads. Bioinformatics. 2015. PMID: 25661542.
Okazaki Y, Furuno M, Kasukawa T, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature. 2002;420:563–73.
Okonechnikov K, et al. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2015. PMID: 26428292.
Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct. 2009;4:14.
Pandey RV, Pabinger S, Kriegner A, Weinhäusel A. ClinQC: a tool for quality control and cleaning of Sanger and NGS data in clinical research. BMC Bioinf. 2016; doi:10.1186/s12859-016-0915.
Park JW, Tokheim C, Shen S, Xing Y. Identifying differential alternative splicing events from RNA sequencing data using RNASeq-MATS. Methods Mol Biol. 2013. PMID: 23872975.
Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PloS one. 2012. PMID: 22312429.
Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32:462–4. PMID: 23912058
Patro R, Duggal G, Kingsford C. Salmon: accurate, versatile and ultrafast quantification from RNA-seq data using lightweight-alignment. bioRxiv. 2015. http://dx.doi.org/10.1101/021592
Quek C, Jung CH, Bellingham SA, Lonie A, Hill AF. iSRAP – a one-touch research tool for rapid profiling of small RNA-seq data. J Extracell Vesicles. 2015. PMID: 26561006.

Quinn EM, Cormican P, Kenny EM, et al. Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. PLoS One. 2013;8(3):e58815.
Ramirez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014. PMID: 24799436.
Renaud G, Stenzel U, Kelso J. leeHom: adaptor trimming and merging for Illumina sequencing reads. Nucleic Acids Res. 2014. PMID: 25100869.
Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43 (7):e47.
Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 2011;12:R22.
Robertson G, Schein J, Chiu R, et al. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010;7:909–12.
Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, et al. Integrative Genomics Viewer. Nat Biotechnol. 2011;29:24–6.
Rueda A, Barturen G, Lebrón R, Gómez-Martín C, Alganza Á, Oliver JL, et al. sRNAtoolbox: an integrated collection of small RNA research tools. Nucleic Acids Res. 2015. PMID: 26019179.
Santana-Quintero L, Dingerdissen H, Thierry-Mieg J, Mazumder R, Simonyan V. HIVE-hexagon: high-performance, parallelized sequence alignment for next-generation sequencing data analysis. PLoS ONE. 2014;9(6):e99033. doi:10.1371/journal.pone.0099033.
Sayols S, Klein H. dupRadar: assessment of duplication rates in RNA-Seq datasets. R package version 1.1.0. 2015.
Selitsky SR, Sethupathy P. tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data. BMC Bioinf. 2015; doi:10.1186/s12859-015-0800-0.
Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26:1135–45.
Shi J, Dong M, Li L, Liu L, Luz-Madrigal A, Tsonis PA et al. mirPRo-a novel standalone program for differential expression and variation analysis of miRNAs. Scientific Rep. 2015. PMID: 26434581.
Shrestha RK, Lubinsky B, Bansode VB, Moinz MB, McCormack GP and Travers SA. QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform. BMC Bioinf. 2014. PMID: 24479419.
Simpson JT, Wong K, Jackman SD, et al. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009;19:1117–23.
Song L, Florea L. CLASS: constrained transcript assembly of RNA-seq reads. BMC Bioinf. 2013. PMID: 23734605.
Song L, Florea L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. GigaScience. 2015; doi:10.1186/s13742-015-0089-y.
Song L, Sabunciyan S, Florea L. CLASS2: accurate and efficient splice variant annotation from RNA-seq reads. Nucleic Acids Res. 2016. PMID: 26975657.
Starostina E, Tamazian G, Dobrynin P, O'Brien S, Komissarov A. Cookiecutter: a tool for kmer-based read filtering and extraction. bioRxiv. 2015. doi:10.1101/024679.
Sun Z, Evans J, Bhagwate A, Middha S, Bockol M, Yan H, et al. CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. BMC Genomics. 2014. PMID: 24894665.
Tarazona S, Furió-Taríl P, Turrà D, Pietro AD, José Nueda M, Ferrer A, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. 2015; doi:10.1093/nar/gkv711.
Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92.

Tjaden B. De novo assembly of bacterial transcriptomes from RNA-seq data. Genome Biol. 2015. PMID: 25583448.

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al.. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7:562–78.

Urgese G, Paciello G, Acquaviva A, Ficarra E. isomiR-SEA: an RNA-Seq analysis tool for miRNAs/isomiRs expression level profiling and miRNA-mRNA interaction sites evaluation. BMC Bioinf. 2016. PMID: 27036505.

Velmeshev D, Lally P, Magistri M, Faghihi MA. CANEapp: a user-friendly application for automated next generation transcriptomic data analysis. BMC Genomics. 2016. PMID: 26758513.

Vitsios DM, Enright AJ. Chimira: analysis of small RNA sequencing data and microRNA modifications. Bioinformatics. 2015. PMID: 26093149.

Wagle P, Nikolić M, Frommolt P. QuickNGS elevates next-generation sequencing data analysis to a new level of automation. BMC Genomics. 2015. PMID: 26126663.

Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.

Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al.. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010. PMID: 20802226.

Wang, L, Wang, S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012;28(16): 2184–2185. http://doi.org/10.1093/bioinformatics/bts356

Wang L, Nie J, Sicotte H, Li Y, Eckel-Passow JE, Dasari S, et al. Measure transcript integrity using RNA-seq data. BMC Bioinf. 2016;17(1):1–16. http://doi.org/10.1186/s12859-016-0922-z Rseqc

Wilhelm BT, Marguerat S, Watt S, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008;453:1239–43.

Wolfien M, Rimmbach C, Schmitz U, Jung JJ, Krebs S, Steinhoff G, et al.. TRAPLINE: a standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. BMC Bioinf. 2016. PMID: 26738481

Yang X, Liu D, Liu F, Wu J, Zou J, Xiao X, et al.. HTQC: a fast quality control toolkit for Illumina sequencing data. BMC Bioinf. 2013. PMID: 23363224.

Yuan Y, Norris C, Xu Y, Tsui KW, Ji Y and Liang H. BM-Map: an efficient software package for accurately allocating multireads of RNA-sequencing data. BMC Genomics. 2012. PMID: 23281802.

Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9.

Zhang T, Luo Y, Liu K, Pan L, Zhang B, Yu J, et al. BIGpre: a quality assessment package for next-generation sequencing data. Genom Proteom Bioinform. 2011;9:238–44. PMID: 22289480.

Zhang Z, Huang S, Wang J, Zhang X, Pardo Manuel de Villena F, McMillan L, et al. GeneScissors: a comprehensive approach to detecting and correcting spurious transcriptome inference owing to RNA-seq reads misalignment. Bioinformatics. 2013;29:i291–9. . PMID: 23812996

Zhao S, Xi L, Quan J, Xi H, Zhang Y, Schack DV, et al. QuickRNASeq lifts large-scale RNA-seq data analyses to the next level of automation and interactive visualization. BMC Genomics. 2016; doi:10.1186/s12864-015-2356-9.