Chapter 1

RNA-seq: Applications and Best Practices

Michele Araújo Pereira, Eddie Luidy Imada and
Rafael Lucas Muniz Guedes

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.69250

Abstract

RNA-sequencing (RNA-seq) is the state-of-the-art technique for transcriptome analysis that takes advantage of high-throughput next-generation sequencing. Although being a powerful approach, RNA-seq imposes major challenges throughout its steps with numerous caveats. There are currently many experimental options available, and a complete comprehension of each step is critical to make right decisions and avoid getting into inconclusive results. A complete workflow consists of: (1) experimental design; (2) sample and library preparation; (3) sequencing; and (4) data analysis. RNA-seq enables a wide range of applications such as the discovery of novel genes, gene/transcript quantification, and differential expression and functional analysis. This chapter will encompass the main aspects from sample preparation to downstream data analysis. It will be discussed how to obtain high-quality samples, replicates amount, library preparation, sequencing platforms and coverage, focusing on best recommended practices based on specialized literature. Basic techniques and well-known algorithms are presented and discussed, guiding both beginners and experienced users in the implementation of reliable experiments.

Keywords: RNA-seq, next-generation sequencing, transcriptome, data analysis, best practices

1. Introduction

A transcriptome represents the entire repertoire of RNA content from an organism, a tissue or a cell and it is dynamic, changing in response to genetic and environmental factors. Several approaches have been developed for transcriptome analysis: hybridization-based (DNA microarray [1]) or sequence-based (ESTs—Expressed Sequence Tags [2], SAGE—Serial Analysis of Gene Expression [3], CAGE—Cap Analysis of Gene Expression [4] and MPSS—Massively Parallel Signature Sequencing [5]). The first sequence-based methods relied on

Sanger sequencing [6], but with advances in next-generation sequencing technology (NGS), transcriptomic studies have evolved considerably and RNA-seq [7, 8] became the state-of-art for transcriptome analysis.

RNA-seq consists of the direct sequencing of transcripts by NGS. Several NGS platforms [9–11] are commercially available nowadays. In general, an RNA set of interest is converted to a library of complementary DNA (cDNA) fragments and sequenced in a high-throughput manner. Compared to ESTs, RNA-seq provides better resolution and representativeness, whereas when compared to microarrays, the independence of reference sequences facilitates the discovery of novel genes and isoforms [8].

RNA-seq experiments harbors challenges from the experimental design to data analysis. Since a complete comprehension of each step is critical to make right decision, this chapter will encompass essential principles required for a successful RNA-seq experiment, focusing on best recommended practices based on specialized and recent literature. Basic techniques and well-known algorithms are presented and discussed, guiding both beginners and experienced users in the implementation of reliable experiments.

2. Experimental design

In order to obtain a successful RNA-seq experiment, it is critical to have a good experimental design. Despite its importance, a proper planning is not always done. There are many experimental options available, and to fully comprehend each step, it is essential to make right decisions, avoiding inconclusive results. These choices depend on extrinsic (e.g., cost, time, samples availability) and intrinsic (e.g., experimental design complexity, transcriptional variability among tissues, samples and organisms) factors. The amount of available resources is usually the main extrinsic limiting factor driving researchers' decisions. First, it is necessary to identify the main goal of an RNA-seq experiment in order to be able to choose the best approach. Qualitative (e.g., annotation) and quantitative (e.g., differential gene expression—DGE) data analyses have some different requirements such as those related to the starting RNA amount, the number and type of replicates, library type and preparation, sequencing platforms, throughput, coverage and depth, and read length. Scotty [12], RNASeqPower [13] and RnaSeqSampleSize [14] are statistical tools designed to aid in the conception of the experimental design, adjusting many of these variables to the main objective and taking into account the financial limitations. A detailed workflow from experimental design to library sequencing is presented in **Figure 1**.

2.1. Starting sample amount

The necessary starting amount of an RNA sample varies between kits and platforms, and the amount of available RNA is one of the limiting factors for an RNA-seq experiment. The majority of library construction kits require micrograms of RNA, sometimes limited to high-quality samples. Takara Bio USA Inc presents some kits for low quantity and/or quality RNA samples: SMARTer Ultra Low mRNA-seq kits (as little as 1 cell or 10 pg of total RNA), SMARTer

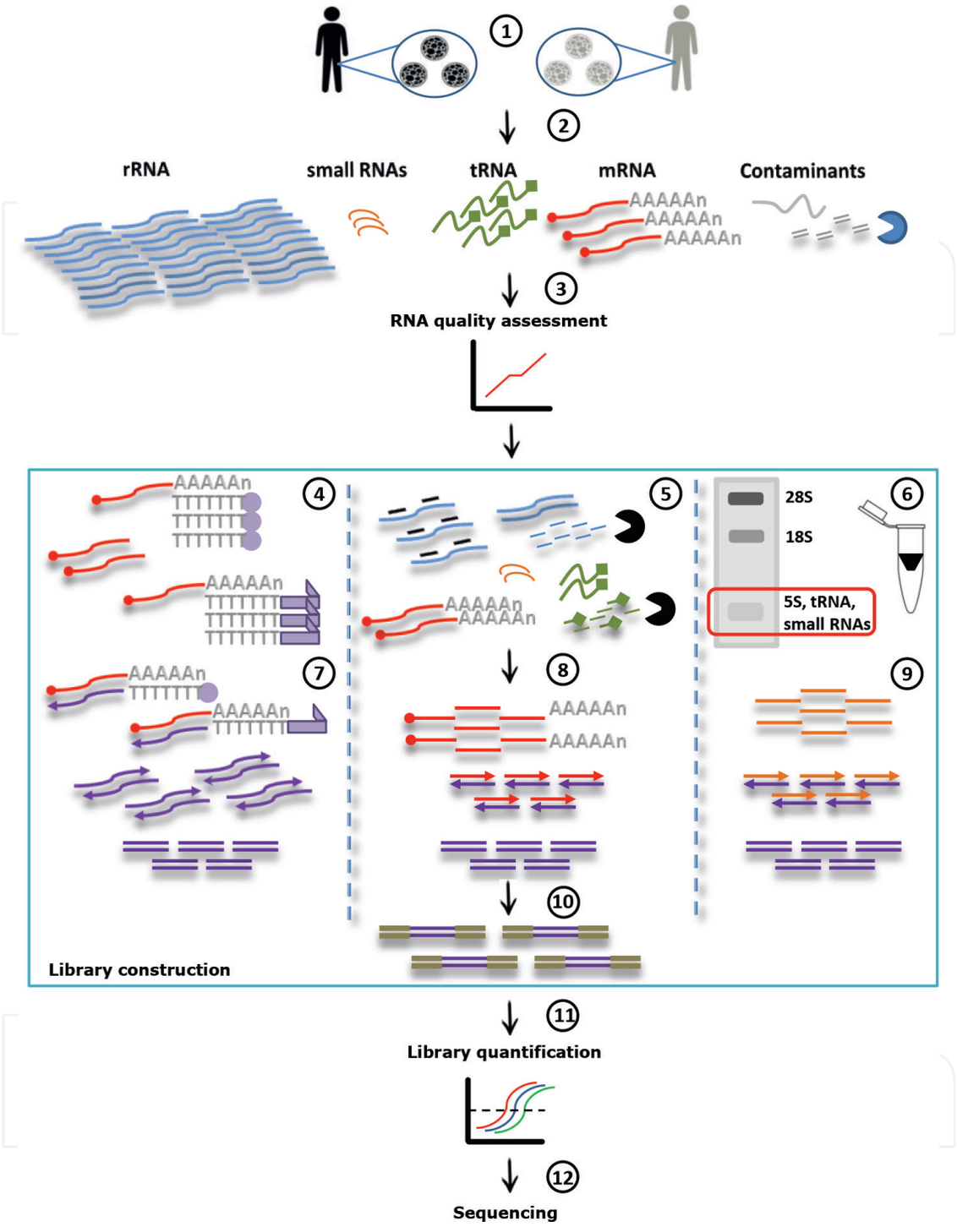

Figure 1. A typical RNA-seq workflow. (1) Experimental design definition of qualitative and quantitative goals. Differential gene expression among different conditions is exemplified; (2) Sample selection, RNA extraction and elimination of contaminants such as genomic DNA; (3) Assessment of RNA integrity; (4-6) RNA enrichment. (4) mRNA enrichment using magnetic or cellulose beads coated with oligo(dT) molecules or oligo(dT) priming; (5) mRNA enrichment through rRNA depletion with conserved probes or Selective Depletion of abundant RNA (SDRNA); (6) Small RNA size-selection through electrophoresis or based on solid phase extraction; (7-9) cDNA single/double strand synthesis. (7) cDNA synthesis followed by fragmentation; (8) mRNA fragmentation followed by cDNA synthesis; (9) cDNA synthesis for small RNA without fragmentation; (10) Adapters ligation; (11) Library quantification and (12) Library sequencing with NGS technology.

Stranded kits (100 pg, regardless of RNA quality) and SMARTer Universal kits (200 pg, regardless of RNA quality). These kits are compatible with both Illumina and Ion Torrent platforms. NuGEN company has also some kits with input RNA levels of 10 pg (Ovation Ultralow Library System V2 and Ovation SoLo RNA-Seq System) available only for Illumina. For a comparison study of four commercially available RNA amplification kits using low-input RNA samples, see Ref. [15].

2.2. Replicates

The variability of an RNA-seq experiment depends on the organism, the biological question under investigation and the available laboratory techniques, and it can be measured by technical and biological variances. Technical replication consists on the repeated analysis of the same sample to infer the variance associated with the technology, that is, equipment and protocols [16]. If only experimental errors analysis is desired, technical replication is satisfactory. Otherwise, biological replicates are necessary [17]. Three biological replicates are the minimum suggested for any inferential analysis [18], although the minimum amount required for a reliable RNA-seq experiment depends on the desired statistical power. For example, in DGE analysis, performing more biological replication is recommended over increasing the sequencing depth [19, 20], and from 6 to 12 biological replicates have been suggested [21]. Biological replication is often preferable to enrich the inferential analysis and increase your statistical power. Statistical knowledge helps to understand the different statistical analysis methods required for different levels of replication [16, 17, 22].

2.3. Sequencing platforms

There are several sequencing platforms available with diverse data formats, throughputs and qualities [9–11]. Two commonly used approaches are sequencing by synthesis (e.g., Illumina, Helicos and PacBio) and ion semiconductor sequencing (Ion Torrent). They can also be classified as clonal amplification-based sequencing (e.g., Illumina and Ion Torrent) or single-molecule-based sequencing (e.g., Helicos, PacBio, Nanopore). For RNA-seq experiments, the most popular platform is Illumina due to its high throughput and low-error rates. PacBio has gained attention due to read length increases since its reads can be long enough to recapitulate a full-length cDNA transcript [23–26]. RNA-seq approaches can also be combined to take advantage of each method benefits. Further information and comparison studies are available in Refs. [11, 27–29].

2.4. Sequencing depth

The required sequencing depth for RNA-seq experiments varies over several degrees. Transcripts are expressed at different levels within the cell, and their coverage differs considerably in any RNA-seq experiment. A deeper sequencing is required to detect low abundance transcripts and rare splicing events, but their relevance can only be assessed with a good biological replication [30]. However, deeper sequencing may increase the detection of off-target RNA species and the number of false positives in differential expression calls [31]. A

correlation between sequencing depth and accuracy demonstrated that as low as one million reads can provide similar information of transcript abundance as more than 30 million reads for highly expressed genes. This result was consistently shown in all six widely used model organisms (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus* and *Saccharomyces cerevisiae*) that represent a wide range of genome sizes [32]. For the majority of human tissue genes, the amount required was about 15–50 million reads [33]. It is noteworthy that there is a point of sequencing depth saturation where a deeper sequencing results in only a small gain of information. More about the impact of sequencing depth on gene detection, gene expression quantification and structural variants discovery can be found in Ref. [33].

2.5. Read length

Short-read sequencing is cheaper than long-read sequencing. RNA-seq experiments usually make use of short-reads; however, longer reads can be helpful and more informative. Reads are usually shorter than full-length transcripts, and a single read may map to multiple positions in the genome stickling expression analysis and transcriptome assembly. Longer read length reduces mapping bias and ambiguity in assigning reads to genomic elements [34] and improves splicing detection [35, 36] and complex transcriptome analysis [37, 38]. However, some studies question the advantages of long reads sustaining that for humans, there are no substantial improvements in transcriptome assembly quality with reads over 150 base pairs [39] and in differential expression analysis with reads over 50 base pairs [35].

2.6. Library type

Standard RNA-seq library protocols do not retain the strand orientation for each original transcript, making it difficult to discriminate gene expression from overlapping genes. Therefore, it is often desirable to construct strand-specific libraries [40–42]. There are several strand-specific protocols available, and they can be performed by two main alternatives. One method consists of marking the second strand by chemical modification, preventing it from being amplified by PCR and leading to the amplification of the first strand only. The deoxy-UTP (dUTP) approach [43] is a well-known example, and it is one of the leading protocols. The other method involves adapter's ligation in a known orientation in the RNA molecule such as Illumina RNA ligation method [44]. A comparison between seven library-construction protocols reveals strong differences and substantial variation in the experimental complexity [40]. Stranded RNA-seq provides more accurate downstream expression analysis, and it is the recommend approach for RNA-seq studies [40, 42]. Moreover, the dUTP and the Illumina RNA ligation methods were identified as the best overall protocols [40, 45].

2.7. Spike-in

The External RNA Control Consortium (ERCC) [46] has developed a set of 92 polyadenylated synthetic spike-in controls for normalization and noise reduction of gene expression. ERCC spike-ins mimic eukaryotic mRNAs and can be added ('spiked') equally to each sample prior

to library construction [47]. Ambion ERCC spike-in control mixes (Thermo Fisher Scientific) are commercially available. Sequins, another set of spike-in RNA standards, can also be used as internal controls and are freely available for non-profit research upon request [48]. Normalization methods should be carefully chosen to ensure that spike-in will behave as expected. The R package *erccdashboard* [49] and Anaquin [50] can be used for spike-in analysis.

3. Sample preparation and library construction

After defining the experimental design, a typical RNA-seq experiment workflow consists of (i) RNA preparation, (ii) cDNA library construction, (iii) sequencing and (iv) bioinformatic analysis. Each step will be briefly discussed below.

3.1. RNA preparation

Since RNA is more labile than DNA and RNases are ubiquitous and very stable enzymes, special precautions and more stringent working practices should be taken to obtain pure and high-quality RNA. Best practices can be found at [51] or spread on diverse companies' websites such as Thermo Fisher Scientific, Qiagen and Ambion.

In an RNA-seq experiment, the RNA preparation consists basically of isolation/extraction and enrichment. Many RNA sample preparation techniques and commercial kits are available. No unique method is optimal for every application, and combination of methods may vary depending on the sample type and the study goals. It is always recommended to carefully follow manufacturer's instructions.

3.1.1. RNA isolation and extraction

In order to isolate high-quality RNA, the samples need to be processed immediately after harvest. If an immediate isolation is not possible, samples can be stabilized in an intermediary solution to preserve RNA integrity and allow storage. Commonly used stabilizers are RNA*later* (Thermo Fisher Scientific and Qiagen) and RNAstable (Sigma-Aldrich). RNA isolation and extraction methods can be manual (e.g., TRIzol–Thermo Fisher Scientific) or automated (e.g., RNeasy—Qiagen), and different types of samples require different approaches, although all of them comprise: (i) sample solubilization in the presence of detergent and chaotropic agents, (ii) sample homogenization for complete cell disruption and (iii) RNA recovery from the lysate with organic or solid-phase extraction. It is also important to have a final RNA free of genomic DNA (gDNA) contaminants. Some protocols can carry over some gDNA into total RNA samples that can be removed by a DNAse treatment. gDNA contamination can lead to a counting bias in downstream analysis and can be detected by reads background over the whole genome (false positive signal). Further information about sample preparation techniques and some commercial kits available can be found in Ref. [52]. Different commercial kits demonstrated satisfactory RNA yield, but differences in the quality of extracted RNA were observed, which can interfere on the downstream analysis [53].

RNA quality can be assessed by gel electrophoresis (agarose or polyacrylamide) or through Agilent Bioanalyzer. RNA quantity can be assessed using spectrophotometer (e.g., Nanodrop), fluorometer (e.g., Qubit) or Agilent Bioanalyzer. No single RNA quantification and quality control method are ideal, and it is necessary to know the limits of each method. We recommend Bioanalyzer since it measures the RNA integrity and level of degradation by the RNA Integrity Number (RIN) score that allows sample quality comparison by a scale with a range from 1 (most degraded) to 10 (most intact) [54, 55]. There is no consensus about the RIN cut-off for sample inclusion or exclusion in a study, but RIN ≥ 6 are commonly acceptable. DGE analysis could be performed even with RIN scores around 4 [56], but non-degraded RNA is preferred for a successful transcriptome analysis. It is also important to highlight that some organisms do not present typical rRNAs peaks and cannot be evaluated by RIN value. Most insect RNA shows a cleavage of 28S rRNA into two similar fragments (28Sα and 28Sβ) that comigrate with 18S rRNA depending on pretreatment and electrophoresis conditions. This comigration is due to the disruption of the hydrogen bonds responsible for maintaining the two 28S fragments together. This profile should not be misinterpreted as low integrity and degradation [57]. In these cases, check the overall Bioanalyzer trace. More information about each method and a comparison study can be found in Refs. [58, 59], respectively.

3.1.2. RNA enrichment

The type of the desired RNA molecule drives the RNA enrichment approach. Selection of mature mRNAs by their poly(A) tails is the most common application and can be carried out with magnetic or cellulose beads coated with oligo(dT) molecules or through oligo(dT) priming for reverse transcription (RT). Therefore, since RNAs from formalin-fixed and paraffin-embedded (FFPE) are degraded and mRNA-seq poorly captures degraded mRNAs, it is not an appropriate method to use with FFPE samples [42], unless adapted protocols are applied such as the recently described protocol based on *in vitro* T7 transcription for linear amplification of mRNA [60]. In order to surpass this limitation, rRNA depletion protocols have been developed based on hybridization in highly conserved ribosomal regions, including the selective depletion of abundant RNA (SDRNA) with RNase H [61, 62], Ribominus (Thermo Fisher Scientific), Ribo-Zero (Illumina), GeneRead (Qiagen) and RiboGone (Takara). Another approach is the duplex-specific nuclease (DSN) normalization by depletion of abundant transcripts, such as rRNAs and tRNAs [63, 64]. Samples can be also enriched of small ncRNAs (e.g., miRNA, siRNA and piRNA) via size-selection through electrophoresis or based on solid phase extraction with commercial kits such as mirVana (Thermo Fisher Scientific) and miRNeasy (Qiagen). For comparison studies between these methods, see Refs. [42, 65]. rRNA depletion is recommended rather than oligo(dT) because it can capture a complete view of the transcriptome and can be used for low-quality RNA samples [65].

3.2. cDNA Library construction

The library construction includes four steps: (i) RNA/cDNA fragmentation, (ii) cDNA synthesis, (iii) adapters ligation and (iv) quantification. Some specific points will be briefly discussed below, but additional information can be found in Refs. [41, 45].

3.2.1. RNA/cDNA fragmentation

The length of your RNA insert is a key factor for library construction and sequencing. Since most current platforms sequence only short reads, most protocols incorporate an RNA or cDNA fragmentation step that allows amplification and sequencing. For short RNAs (under 200 pb), no fragmentation is required. There are three main ways to fragment the nucleic acid samples: physical (e.g., sonication, nebulization), enzymatic (e.g., RNase III, DNase I or Fragmentase) and chemical (e.g., heat, metal ion) shearing. Little information is known about which is the best method for each application. A comparison study of nebulization, sonication and enzymatic digestion showed that all three methods presented equal performance and that fragmentation is indicated [66]. In most cases, RNA is fragmented before conversion into cDNA. Furthermore, it is important to highlight that due to FFPE samples degradation, cDNA fragmentation must be performed instead of RNA fragmentation when using oligo(dT) priming for first-strand synthesis.

3.2.2. cDNA synthesis

After an adequate RNA preparation, RNA must be converted to double complementary DNA (cDNA) via RT, generating a cDNA:RNA hybrid. This process is known as first-strand cDNA synthesis and requires an oligonucleotide primer. Three options are available: oligo(dT) priming, random priming or gene-specific priming. The first two are the mainly used for RNA-seq. Oligo(dT) priming is one of the oldest methods for first-strand synthesis and involves oligo(dT) primer to capture the poly(A) tail of mature mRNA. Because of their specificity for poly(A) tails, oligo(dT) priming is not compatible with fragmented RNA, such as FFPE samples, nor for RNAs that lack poly(A) tails, such as non-mRNAs (e.g., microRNAs (miRNAs)). If using this methodology, cDNA fragmentation must be performed instead of RNA fragmentation. Besides that, RTs are not highly processive polymerases and can prematurely terminate the strand biosynthesis, leading to 3′ end bias and under-representation of the 5′ ends. Random priming involves oligonucleotides with random base sequences that prime at random positions along the RNA (i.e., no template specificity), and it is preferable to oligo(dT) priming. This approach allows recovery of non-poly(A) RNAs and prevents 3′ end bias, resulting in a more uniform transcript coverage. However, it was shown that random priming is not completely random leading to a nucleotide bias across the first reads positions [67, 68].

The first-strand cDNA is used as a template to generate double-stranded cDNA. Second-strand cDNA synthesis can be performed by (i) RNA nicking of the RNA template by RNase H and synthesis with *E. coli* DNA polymerase I and T4 DNA ligase [69], (ii) using an oligo that is complementary to an adapter located in the 5′ end of the RNA template or by (iii) Clontech's SMART (Switching Mechanism At 5′ end of RNA Transcript) technology [70]. RNase H method presented a better performance for low-quality RNA when compared to four other methods (Ribo-Zero, NuGEN, SMART and DSN-lite) [65].

3.2.3. Adapters sequences and ligation

Adapters sequences must be ligated at the ends of every single molecule during library preparation, and this process varies depending upon the sequencing platform. It can contain one

or more extra functional elements such as barcode/index to allow multiplexing and a second sequencing-priming site to allow paired-end sequencing. The addition of adapter via Y-adapter PCR is the most commonly used technique. Adapters can also be added via RT/PCR during the first- and second-strand synthesis process or via ligation.

3.2.4. Library quantification

To ensure the maximum yield (i.e., data output) and quality from your RNA-seq experiment, it is important to have a precise quantification of your NGS libraries. Inaccurate quantification may lead to lower throughput, lower sequences qualities and poor samples balance within your multiplex. There are many ways to quantify your libraries, but the most accurate and effective method is quantitative real-time PCR (qPCR). qPCR is more sensitive and only quantifies amplifiable DNA molecules (i.e., molecules that contain both adaptor sequence), providing a more precise estimation. Some commercial kits available are KAPA Library Quantification Kit (Kapa Biosystem), GeneRead Library Quant System (Qiagen), Ion Library TaqMan Quantitation Kit (Thermo Fisher Scientific), QPCR NGS Library Quant Kit (Agilent), PerfeCTa NGS Quantitation Kit (Quantabio) and NEBNext Library Quant Kit (New England BioLabs). Other methods are similar to the previously mentioned for RNA quantification: spectrophotometer (e.g., Nanodrop), fluorometer (e.g., Qubit) and Agilent Bioanalyzer. However, since these methods measure total nucleic acid concentrations, including non-amplifiable DNA, they can lead to inaccurate results. It is also recommended to verify the libraries fragment size distribution, which can be performed by electrophoresis, preferably Agilent Bioanalyzer. Bioanalyzer electropherogram needs to show a narrow distribution with a peak height of the average size fragmentation value. After quantification, the library must be sequenced with the platforms discussed in Section 2.3, and data must be analyzed through bioinformatic tools. RNA-seq data analysis will be discussed below.

4. Data analysis

RNA-seq data analysis involves many different strategies that depend on the goals and biological questions established at the time of the study design. A typical data analysis includes quality control, reads preprocessing, alignment to a reference or *de novo* assembly and downstream analysis such as transcripts annotation, DGE, gene fusion analysis and alternative splicing. In the following topics, we will emphasize common steps and applications of this technology. A detailed workflow for data analysis is presented in **Figure 2**. Bioinformatic tools discussed in this chapter are compiled at **Table 1**, and a more exhaustive list of available tools can be found in Ref. [71]. For those with limited access for computational resources or little experience with command-line execution of these bioinformatic tools, free online (Galaxy [72]) and commercial (Illumina BaseSpace [73] and Geneious [74]) platforms can be very helpful and intuitive.

4.1. Quality control and reads preprocessing

A complete pipeline for an RNA-seq analysis demands some checkpoints in order to ensure the quality of the results and elimination of noise from the biological samples. After sequencing,

the analysis starts with files containing the raw reads. The FASTQ [75] is the standard format used to store the nucleotide sequences along with a per base quality score in Phred log scale. The qualities, typically with scores from 0 to 40, are represented by single letters encoded with pre-defined ranges of characters from the American Standard Code for Information Interchange (ASCII) table. Currently, there are two patterns: Phred + 64, used in initial Illumina versions 1.3 + and 1.5 + and Phred + 33, the default encoding for Sanger and more recent sequencers. The FASTQ is widely accepted and used in most downstream software, although the unmapped BAM (uBAM) format has been recently encouraged as it is capable of storing important sequencing metadata not present in FASTQ, and for being binary, it demands less disk storage. Some sequencing platforms, like Ion Torrent, have already included uBAM as default output format in their pipelines. Both formats are interchangeable by using Picard [76], BamUtil [77] and BamTools [78].

The first step is to perform a quality control (QC) of the data, checking parameters like amount of reads per sample, general read and base qualities, mean reads length, G + C content, presence of unclipped adapters or PCR primers and unexpected repetitive sequences. This general overview will indicate if library construction and sequencing were properly performed, or if errors like contaminants, poor ribosomal RNA depletion or low sequencing output will demand a new round of experiments. The most common software used to retrieve these basic statistics is FastQC [79] and PRINSEQ [80]. The first was mainly designed for Illumina, while the later for 454/Roche technology and may be also used for preprocessing. Both programs are available with intuitive graphical user interfaces (GUI), accept other sequencing technologies input files and generate graphical reports, which are very useful for guiding the choice of filtering thresholds.

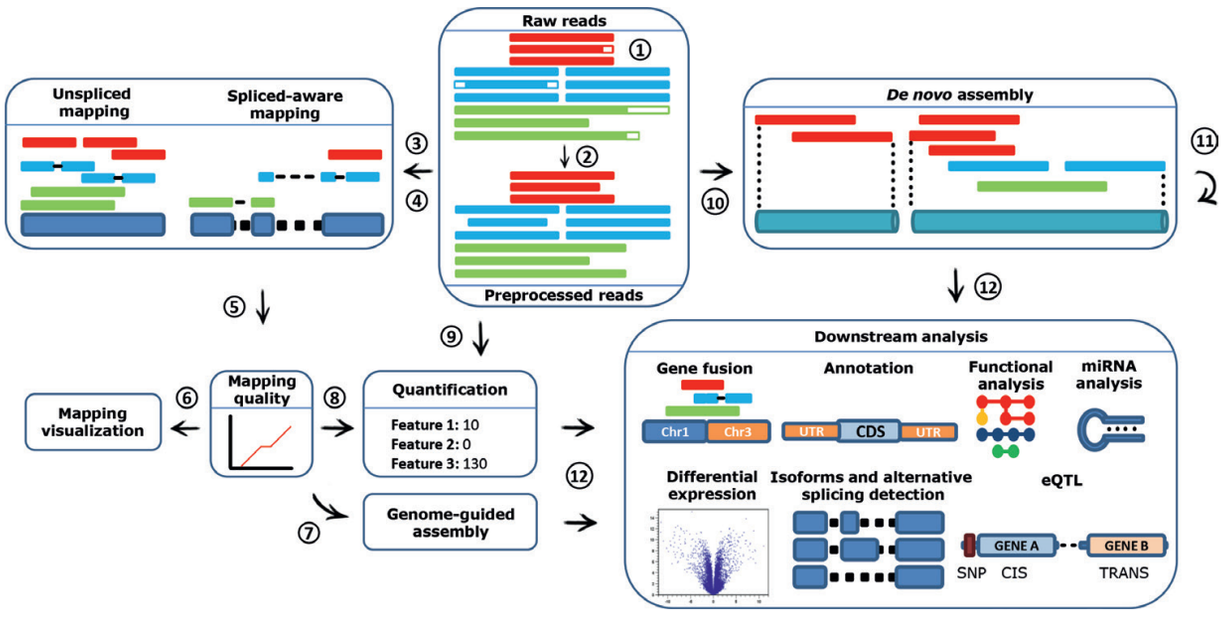

Figure 2. RNA-seq data analysis. (1) Raw single-end and paired-end reads obtained from NGS sequencing; (2) Adapters clipping and base quality trimming. Alternatively error correction can be performed; (3) Mapping without preprocessing using soft-clipping; (4) Unspliced or spliced-aware reads mapping; (5) Assess mapping quality and biases; (6) Mapping visualization; (7) Transcriptome genome-guided assembly; (8) Per feature quantification using mapped reads; (9) Per feature quantification using quasi-mapping approach; (10) Transcriptome de novo assembly; (11) Mapping reads to de novo assembled transcriptome; (12) Downstream data analysis.

Category		Tools
Experimental design		Scotty [12]
Raw reads quality control		FASTQC [75]
Reads preprocessing	Read clipping/trimming	Picard [72], BamUtil [73], BamTools [74], PRINSEQ [76], Cutadapt [78], FASTX-Toolkit [79], Trimmomatic [80]
	Paired-end reads overlapping detection	FLASh [84], PEAR [85]
	Reads error correction	SEECER [86], Rcorrector [87]
Unspliced mapping	Hash Table index based	BFAST [90], MAQ [92], Mosaik [93], Novoalign [94], RMAP [95], SHRiMP [96]
	FM-index based	Bowtie2 [97], BWA [100]
Spliced-aware mapping	Hash Table index based	GSNAP [91], RNASEQR [103]
	FM-index-based	TopHat2 [98], HISAT2 [99], SOAP-splice [101], STAR [102], RNASEQR [103]
Alignment quality assessment		Picard [72], BamUtil [73], BamTools [74], Samtools [106], Qualimap2 [107], BAMstats [108], SAMstat [109]
Assembly	Genome-guided	Cufflinks [111], Scripture [112], StringTie [113]
	De novo	Rnnotator [115], Trans-ABySS [116], Trinity [119], Oases [120]
Assembly quality assessment		Detonate [122], TransRate [123], BUSCO [124]
Alignment visualization		IGV [125], Tablet [126], UCSC [127]
Raw read counts	Mapped-based	featureCounts [129], HTSeq-count [130], RSEM [136]
	Pseudoalignment	Kallisto [140], Salmon [141]
Raw read counts quality assessment		NOISeq [131]
Differential expression		DESeq [132], DESeq2 [133], edgeR [134], CuffDiff2 [137], BitSeq [138], Ballgown [139]
Annotation		BLAST [145, 146], DIAMOND [147], InterProScan [148], tRNAscan-SE [149], RNAmmer [150], Blast2GO [151], Annocript [152], TRAPID [153], Trinotate [154]
Enrichment analysis		GSEA [155]
Alternative splicing		Cufflinks [111], Scripture [112], StringTie [113]
Differential alternative splicing		CuffDiff [111], Ballgown [139], DEXSeq [161], rMATS [162], SpliceR [163], MISO [164], DiffSplice [165]
Fusion genes		SOAPfuse [170], FusionCatcher [171], JAFFA [172]
miRNA		miRdeep2 [179], miReNA [180], miRanalyzer [181]

Table 1. Tools for RNA-seq data analysis.

The following preprocessing step is crucial and can greatly influence the data analysis [81]. Besides PRINSEQ, other tools like Cutadapt [82], FASTX-Toolkit [83] and Trimmomatic [84] are efficient in preprocessing reads, but FASTX-Toolkit cannot be used with paired-end reads. Generally, due to problems inherent in sequencing technologies, the bases in 3'end of reads have lower quality, and one may choose to filter off reads with low mean quality or trim only the low-quality ends. Trimming in most cases may improve mappability, although shorter reads have a higher probability of erroneous mapping. Therefore, it is recommended to remove short reads in conjunction with non-aggressive base-quality trimming to avoid spurious mapping and incorrect inferences [85, 86]. Adapter removal and trimming low-quality ends improve RNA-seq assembly, single nucleotide polymorphism (SNP) detection and gene-expression analysis.

Modern mapping tools (see next section) are capable of labeling the unaligned read ends, a process known as soft-clipping, without actually removing them (hard-clipping). There is no consensus on which approach is the best, but it has been considered that keeping as much as information as possible would be better for downstream analysis. For example, the soft-clipped reads are important for detection of genomic structural variants [87].

When the goal is to perform RNA-seq *de novo* assembly, supplementary tools can be used to join overlapping paired-end reads, like FLASh [88] and PEAR [89]. Additionally, base error correction can be applied as an alternative to read trimming and filtering, increasing the amount of useful data and consequently the contig sizes. SEECER [90] and Rcorrector [91] were specifically designed for this task. Both strategies will likely improve assembly qualities.

In summary, preprocessing is beneficial, but there is no best tool for any experiment or general rule for filtering thresholds. All software has its own standard parameters, advantages and limitations, being recommended a case-by-case analysis and a thorough software comparison.

4.2. Mapping, assembling and visualizing mapped reads

Now that the raw reads have been preprocessed, alternative approaches can be chosen according to the availability of a reference sequence. If present, reads can be mapped to the genome and the gene that originated the transcript from which the reads were derived may be inferred and expression quantified. The genome may also be used to guide transcriptome assembly, resulting in several contigs representing the genes and its isoforms. On the other hand, if the studied species still lacks a reference sequence, reads can be *de novo* assembled, and transcripts can now be used as a mapping reference.

4.2.1. Mapping to a reference

Mapping reads to a reference can be also seen as a traditional pair-wise sequence alignment, as observed in common Basic Local Alignment Search Tool (BLAST) [92], but with the main difference that a vast amount of reads are compared with a database composed of fewer and longer sequences instead of several thousand nucleotides/proteins. This is a field under constant development with plenty of tools available [93]. These tools have to deal with inherent mapping challenges, such as sequencing errors, natural sequence variability like SNPs and

indels, reads spanning exon junctions and repetitive regions or pseudogenes in references. To guarantee reproducibility, it is highly recommended reporting alignment parameter details, such as mapper and reference versions and sources, allowed seed mismatches, minimal alignment score and treatment given to multi mapping reads.

Mappers can be roughly divided by the algorithm chosen to create indexes and by the ability to recognize exon-exon junctions. Indexes have the purpose of making the alignments significantly faster and are mainly divided into Hash Table or compressed prefix or suffix array-like structures (FM-index). Their principle is to quickly find small local alignments representing substrings of whole reads—designated as seeds—in the reference and then extend those alignments surpassing a defined quality threshold toward the read ends, assigning a Phred-based mapping quality score for each read. Unfortunately, most mappers have developed their own mapping quality formulas, creating a non-uniform mapping qualification. Some well-known Hash Table-based algorithms are BFAST [94], GSNAP [95], MAQ [96], Mosaik [97], Novoalign [98], RMAP [99] and SHRiMP [100], while Bowtie2 [101], TopHat2 [102], HISAT2 [103], BWA [104], SOAP-splice [105] and STAR [106] are examples of FM-index based algorithms.

Regarding the splicing events, they can be divided into unspliced and splice-aware aligners. Most recent mappers are capable of using reference annotation files to deal with known exon-exon junctions and to predict new splice sites, which is essential when analyzing RNA-seq data from most eukaryotes. GSNAP, SOAP-splice, RNASEQR [107], STAR and TopHat2 are some recommended options for spliced alignments, but for intronless species, miRNA and transcriptomes, unspliced aligners can be used. Comparative evaluations showed that FM-index-based mappers are preferable [108] and that, again, no tool is the best for every performance parameters like speed, alignment yield, exon discovery and accuracy [109].

The standard alignment output is the Sequence Alignment/Map (SAM) format or its binary version BAM and they are essential inputs for many downstream applications. Picard [76] and Samtools [110] are frequently used to manipulate these files. It is advisable to assess the alignment quality from SAM/BAM files with tools like Qualimap2 [111], BAMstats [112] and SAMstat [113] for general characterization or for comparing mappers' performances.

4.2.2. Genome-guided assembly

Short RNA-seq reads represent only a small portion of most transcripts, and therefore, overlaps have to be detected in order to fully reconstruct the original molecules. Paralogous genes, alternative splicing, alternative transcription initiation and termination sites increase the complexity and impose computational challenges in Eukaryotic assembly analysis [114]. For Bacteria, Archaea and lower eukaryotes, the absence or smaller amount of introns makes the assembly more straightforward.

RNA-seq assemblers greatly differ from DNA-seq algorithms because a wide range of transcripts coverage is expected, and several gene isoforms can be observed resulting in thousands of contigs stead of ideally one per chromosome. When a good quality reference genome is available, the usual procedure is to use the coordinates of aligned reads to separate them into clusters and perform a *de novo* alignment individually for each locus, from which individual

isoforms can be inferred. Cufflinks [115], Scripture [116] and StringTie [117] are recommended tools, and their algorithm strategies have been reviewed [118], with StringTie [117] presenting better transcript reconstruction performance. Paired-end, strand-specific libraries and longer reads are highly encouraged for better assemblies and to allow distinction in overlapping transcripts from opposite strands for gene-dense species and antisense transcription. Genome-guided assembled transcriptomes can be used to improve gene structures annotation through detection of transcription boundaries and splice-sites.

4.2.3. De novo assembly

In the absence of a reference sequence or if only a fragmented draft genome is available, overlaps have to be detected from the complete read set in a *de novo* assembly approach. The independence from a good quality reference and mapping procedures can be also seen as an advantage. The counterpart is that sequencing depth must be obtained in a higher coverage, estimated around 30× [119], while genome-guided approach requires about 10× [120, 121] to find full-length transcripts. The higher throughput increases the processing requirements, so data digital normalization is recommended in order to remove redundancy without impacting the assembly outcome [122]. Although the *de novo* approach is usually more error prone and computationally intensive, it allows the discovery of novel splicing events, unpredicted genes and exons, chromosomal rearrangements and trans-splicing. Trinity [123], Oases [124], Rnnotator [119] and Trans-ABySS [120] are advised for this task. Whenever possible, a combined genome-guided/*de novo* strategy is recommended, as enhanced performance is observed [125]. A comprehensive overview of transcriptome assembly can be found in Ref. [121]. Evaluation of the assembly quality and transcriptome completeness can be assessed with Detonate [126], TransRate [127] and BUSCO [128].

4.2.4. Visualization

Alignment output SAM files are hard to be interpreted with common text editors, and therefore, a number of graphical browsers have been developed to inspect NGS sequencing data at any specific loci at nucleotide level. IGV [129], Tablet [130], Browser Genome [131] and UCSC [132] are extremely useful when validating novel transcripts and gene junctions, checking the coverage support for genomic variants and spot read piles, which may represent repetitive regions.

4.3. Downstream analyses

After conducting these general steps, the experiments can be directed to specific applications in order to address the scientific questions, designated as downstream analysis.

4.3.1. Quantification and differential expression

The primary goal of most RNA-seq projects is to quantify and compare the gene expression under different conditions and infer biological function to differential expression at gene or transcript level. Intra-sample abundance comparisons were commonly performed with

Reads Per Kilobase per Million (RPKM mapped reads) or Fragments Per Kilobase per Million (FPKM mapped reads) metrics. Their principle is to count the amount of raw reads mapped to each genomic feature and normalize considering the gene length and library depth. Although still widely applied, these normalization metrics should be avoided as RPKM has shown to be inconsistent and Transcripts Per Million (TPM) is preferable [133]. Raw reads counting can be obtained with feature counts [134] and HTSeq-count [135], which are capable of detecting multi-mapping reads, exon junctions and overlapping reference features. NOISeq [136] can be used to assess the count quality parameters, such as saturation and specificity, in a set of comprehensive plots.

DESeq [137], DESeq2 [138] and edgeR [139] packages are recommended for between-sample comparisons to detect differences in the relative abundances of genes [140]. Quantification at transcript level can be analyzed with Cufflinks [115] and RSEM [141] and compared with DESeq2, CuffDiff2 [142], BitSeq [143] or Ballgown [144]. Variations in expression between different conditions are usually measured in log2 fold-change units. DESeq2 can also perform pair-wise and time series analysis.

Generally, a control set of housekeeping genes should present non-differential expression and a high between replicates correlation (Spearman $R^2 \geq 0.9$) observable in Principal Component Analysis (PCA) plots [18]. For a set of 12 or less replicates, at gene level, edgeR or DESeq2 is recommended to detect differential expression and DESeq when more than 12 replicates are available [21]. Thresholds in log2 fold-change should be applied to increase the true positive and decrease the false positive rates, but this parameter is highly dependent on the amount of biological replicates, varying from 0.1 to 0.5 [21].

Recently, quasi-mapping (or pseudoalignment) approaches have been proposed for RNA-seq quantification, like Kallisto [145] and Salmon [146]. Their main difference is that reads are assigned to reference sequences without base-to-base alignment, making analyses usually considerably faster. They have shown comparable performance over complete mapping-based methods, can incorporate information from multi-mapping reads, and provide counts and abundances already as normalized TPM values, which can be used as input for differential expression analysis. These are promising although under development tools.

Although RNA-seq provides a precise and accurate estimation of RNA abundance, these findings are still widely required to be further validated through quantitative PCR, also known as qPCR or real-time PCR as it is still considered the gold standard for gene expression quantification. However, it is still questionable whether qPCR validation is still necessary for RNA-seq studies. High correlation between RNA-seq and qPCR results has been observed in previous studies [7, 147, 148]. Due to this high consistency, qPCR may be more useful when performed on different biological replicate samples from those already sequenced, confirming the DGE findings and validating the biological conclusions.

4.3.2. Annotation

In computational biology, annotation is the process of identifying the location and sequence of genomic elements and/or assigning biological function to them. Despite the annotation process

being mostly carried over genomic sequences, such as newly sequenced genomes, RNA-seq data can provide valuable information to improve existing annotations [149] or create novel transcript annotations for an unsequenced organism [123].

The major drawback of using genome sequences for annotation is that only features with patterns or conservation with annotated elements, such as open reading frames (ORFs), tRNAs and rRNAs can be inferred from it. On the other hand, RNA-seq data provide a new layer of information that allows precise identification of pattern less features such as untranslated regions (UTRs), non-coding RNAs and post-transcriptional events. Even though some features can be somewhat inferred through DNA sequences, for example, Transcription Start Site (TSS), TATA box/CpG islands and splicing sites, transcriptomic data still provide a more reliable annotation.

Transcriptome assembly, *de novo* or reference-guided, often reveals new potential transcripts whose functions are unknown. Before any further step can be made, it is crucial to gather information on these transcripts function in order to extract any meaningful answer.

The most common approach to annotate a transcript is to look for similar known transcripts or protein sequences in large databases. This is usually done using versatile tools like BLAST/BLASTX [150, 151] or DIAMOND [152] when looking for similar nucleotide or protein sequences. It is often better to perform searches at protein level since it is easier to find homology, as they tend to be more conserved than nucleotide sequences, especially if the study subject has no close species sequenced.

InterProScan [153] can be used to search for conserved protein signatures. This is especially useful when it is difficult to find full sequence homologs given that the study organism might be too divergent from species sequences available in the database. Protein families often present signature domains that are well conserved even among divergent species, so these signatures can give insights into the putative function of the protein. The process for annotating non-coding transcripts differs from protein coding transcripts. They usually present poor sequence conservation since their function relies on factors, such as secondary structure, rather than amino acid sequences. Therefore, their annotation process requires specialized software to detect those intrinsic characteristics of a given class of non-coding transcripts, for example, tRNAscan-SE [154] for tRNAs and RNAmmer [155] for rRNAs.

Given the importance of annotation, there are plenty of tools and pipelines developed to streamline this process. Some annotation tools like Blast2GO [156] are generic and very user-friendly, although it requires a paid license to use it. Others like Annocript [157], TRAPID [158] and Trinotate [159] are pipelines developed specifically for annotating transcriptomes. It is important to note that although automatic pipelines often ease and speed up the analysis, it comes at a cost of lesser control of the annotation process.

4.3.3. Enrichment analysis

Functional enrichment analysis is a computational method capable of determining whether a pre-defined set of genes shows significant differences between samples. The GSEA software from Broad Institute runs the original GSEA algorithm [160]. Although alternative algorithms

have been published since then, the original algorithm is still the most widely used. In order to perform an enrichment analysis from RNA-Seq data, the GSEAPreranked software is recommended and it requires two types of data: a gene set list and a ranked list.

A gene set is a set of genes related to the feature to be tested for enrichment. A variety of features can be tested from general features such as pathways and chromosome location, to more specific features such as cancer signatures or miRNAs targets. Gene sets can be obtained from the Molecular Signatures Database (MSigDB) that comprehends thousands of pre-defined gene sets, or it can be created by the user.

A ranked list of the genes needs to be provided to test if the chosen gene set is significantly enriched at either end of the ranking. The list can be ranked according to any quantitative feature such as gene expression or fold-change results from DGE analysis.

4.3.4. Alternative splicing

Alternative splicing (AS) is a post-transcriptional mechanism present in the majority of eukaryotes that greatly increases the diversity of proteins that can be encoded by a determined genome. This process occurs when particular regions of a gene are included or excluded, through splicing, from the final processed mRNA sequence. AS can occur in several ways, such as exon skipping, intron retention, alternative 5′ donor and 3′ receptor sites [161, 162], analysis of new AS events or patterns is relevant since many traits, especially genetic diseases such as cancers, are related with disorders in splicing patterns that generates aberrant variants [162, 163].

AS analysis by deep sequencing requires splice-aware programs capable of aligning transcripts reads to a reference genome while performing the difficult task of placing spliced reads across introns by determining the exon-intron boundaries. A systematic evaluation of splice-aware alignment programs for RNA-seq data performed by the RNA-seq Genome Annotation Assessment Project (RGASP) Consortium [109] tested 26 RNA-seq alignment protocols and concluded that, in general, GSNAP [164], MapSplice [165] and STAR [106] compared favorably to other methods. Still, two of this software (GSNAP and STAR) presented many false exons junctions in the output if they were not filtered based on the number of supporting alignments.

Following the alignment step, software like cufflinks [115], scripture [116] and StringTie [117] can be used to perform transcript reconstruction, which can reveal new splicing isoforms evidenced by the alignments. This step usually yields an updated GTF annotation file as output that can be used in subsequent steps.

If data from different conditions are available, differential AS analysis can be performed. With the alignment results (SAM file) and a GTF annotation file at hand, differential exon usage analysis can be performed with DEXSeq [166] and differential analysis of AS events, such as skipped exon, alternative 5′ and 3′ splice site, mutually exclusive exons, and retained intron events can be performed with rMATS [167]. There are plenty of other software specialized in performing differential AS analysis each one with their advantages and disadvantages, such as CuffDiff [115], Ballgown [144], SpliceR [168], MISO [169] and DiffSplice [170].

4.3.5. Fusion genes

Fusion genes or chimeras are aberrant alterations commonly found in tumor cells [171] that can be useful biomarkers or therapeutic targets [172]. They may originate from chromosomal rearrangements, insertions, deletions and inversions or even by trans-splicing events. The increasing throughput and reads length from NGS technologies have facilitated their detection and supported the development of several bioinformatic tools [173]. For fusion detection, most and more accurate methods rely on good quality read alignments supporting discordant mappings (read segments aligning to different genes) and both single- or paired-end sequencing, although paired data increase the probability of fusion detection [174]. A recent evaluation defined SOAPfuse [175], FusionCatcher [176] and JAFFA [177] the best tools among 18 options for real and simulated data, and their combination has shown increased performance, albeit high false-positive rates are still a reality in this field, with space for improvements [178].

4.3.6. miRNA

MicroRNAs (miRNAs) are a subset of small non-coding RNAs, usually 21–23 nt long that play a post-transcriptional regulatory role in several pathogenic and developmental processes [179]. These molecules are part of an RNA-induced silencing complex (RISC) containing Dicer, Argonaute and many associated proteins that can cause enhanced decay/cleavage of mRNA target, elongation and ribosomal binding inhibition, thus acting at transcriptional and translational levels [180].

A common miRNA pipeline follows the same steps as the conventional RNA-seq: (i) raw data must be preprocessed as previously described where adapters and low quality bases are trimmed with a minimum length filter (e.g., 18–21 nt for miRNAs), (ii) sequences are mapped to a reference (genome, RefSeq, miRBase) and raw counts are estimated, (iii) the raw count of mapped reads is normalized and (iv) downstream analysis is conducted to investigate biologically relevant questions. Due to its small nature, miRNA sequencing analysis has some caveats that require attention especially in steps (ii) and (iii).

The read mapping step is crucial for accurate miRNA abundance estimation, and therefore, the alignment algorithm must be carefully selected and adjusted to deal with its small size. Although a wide range of software are available to perform this task, some aligners are designed and optimized for specific tasks (e.g., SNP calling, splicing detection, gapped alignment) that might not be appropriate for the task at hand [181]. Compared with conventional RNA-seq, indels and splicing events are usually not relevant to miRNA alignment, and therefore, splice-aware aligners are not required for this task. To these extent general purpose aligners such as BWA-MEM, bowtie [182] and STAR [106] can be used. Most aligners default settings are set for conventional longer RNA-seq reads, and since miRNAs are very short, aligners' parameters should be tweaked. The default seed size for these aligners is longer than miRNA sizes and therefore should be set to a value that is at least shorter than the smallest read size. Given that sequencing errors might occur and the fact that many miRNAs often does not present an exact match with their target, it is recommended to allow at least one mismatch in the seeding and alignment process as well [183]. Also during the mapping step, it

is very common to find multi-mapping reads since we are dealing with very small sequences. Similarly to conventional RNA-seq, multi-mapping reads are usually not taken into account for the abundance estimation, since it is impossible to know from where the read was originated. As long as these aligners are properly set, they should yield similar results [181].

Please note that for the aforementioned pipeline, miRNA annotations or sequences are usually required for raw counting estimation. If annotations are not available for the study subject or looking for novel miRNAs candidates, algorithms such as miRdeep2 [184], miReNA [185] and miRanalyzer [186] can be used to annotate novel canonical and non-canonical miRNA.

After raw miRNAs abundances are estimated, a normalization step is required in order to remove bias of non-biological origin (e.g., sequencing depth, sample handling, library preparation). A good normalization technique should reduce those biases without generating noise, so that the remaining differences between samples are truly of biological origin. Previous comparative studies on normalization procedures for miRNA data resulted in conflicting results. A study from Garmire and Subramaniam [187] supported the use of quantile and Lowess normalization, while Tam et al. [181] and Dillies et al. [140] advocated for the use of Trimmed Mean of M (TMM) and Upper quartile (UQ) normalization. Nevertheless, the results from any of these methods and also DESeq2 normalization [138] method should be highly similar, while other normalization methods such as CPM, total count scaling and linear regression should be avoided since they tend to present higher variance and bias [181]. Several R/Bioconductor packages can be used to normalize the data and also run differential expression, such as edgeR [139] (TMM and UQ), DESeq2 [138] (DESeq normalization) and limma [188] (quantile and cyclic loess).

After all these processing steps, the resulting miRNA estimation is ready to conduct downstream analysis. This can be done with useful databases. Being the primary miRNA sequence repository, miRBase [189] contains several features that may help to investigate the roles for miRNAs of interest, such as annotations for a wide range of species, references links for studies and deep sequencing evidence.

4.3.7. eQTL

Quantitative trait loci (QTLs) are genomic regions that contain sequence variants that can affect any given trait. Since genome-wide association studies (GWAS) started [190], thousands of variants have been associated with complex traits and diseases. The process of assigning variants to a gene is relatively straightforward when variants are located in coding regions that can have a direct effect on a gene product; however, most variants are found in non-coding regions making difficult to identify the causal genes [191]. By integrating transcriptomic data, it is possible to identify causal genes for non-coding variants that affect its expression. When the trait in question is gene expression, they are referred as expression quantitative trait loci (eQTLs) that, similarly to other QTLs, are sequence variants capable of affecting the expression level of one or more genes that will ultimately result in different phenotypes. eQTLs can be classified according to the location of the QTL itself and its targeted gene, and according to the mechanism that affects the expression [192].

Regarding the eQTL-Gene position, when they are located close to the genes, they influence they are called local eQTLs. Local eQTLs can affect a gene in two ways: in *cis* (cis-eQTL) when the variant affects only the gene that is located on the same chromosome and not affecting the copy of the homologous chromosome, thus causing an allelic imbalance; and in *trans* (trans-eQTL) when the eQTLs do not affect the target expression directly, but instead affect an intermediate factor that will ultimately affect its target expression. Since the intermediate factor acts equally for both alleles, it does not cause allelic imbalance. On the other hand, eQTLs located further away from their target genes are referred as distant eQTLs, usually act in *trans* and are harder to find [192]. Several eQTL-mapping studies published in the past few years showed that many variants often affect gene expression levels of nearby and distant genes [193–197] highlighting the importance of integrating transcriptomic and genomic data.

Despite the mapping process for eQTL analysis being conceptually simple, since this analysis is dealing with allelic specific expression, some caution is required during its counting estimation. For the aligning process, general purpose aligners or variant aware aligners such as GSNAP [164] can be used. After the alignment, some steps are recommended for retrieving allelic-specific counts, such as removing duplicate reads that may arise from PCR artifacts. However, it is important that the choice for discarding a duplicate read is not done by mapping score as this might bias toward the reference allele [198]. Also, mapping bias should be controlled by filtering sites with likely bias [199]. Some tools like ASEReadCounter from GATK for allelic-specific expression implement these filters by default [200].

The GTEx portal is a valuable resource to study human gene expression and regulation related to genetic variation. It hosts data from several eQTL studies and much information on laboratory and analysis methods for eQTL [201].

5. Concluding remarks

In the past few years, recent advances in sequencing technologies allowed the cost-efficient generation of an unprecedented amount of biological information. Similarly, RNA-seq techniques are under continuous improvements allowing wide range applications and development of high level resolution experiments such as those based on the emergent single-cell RNA sequencing (scRNA-seq) field. To couple with this ever increasing data, several tools and pipelines have been constantly developed. The bioinformatics field changes in an astonishing pace, in a way that it is almost impossible to keep up with all the new tendencies, the overwhelming amount of available software and the controversial opinions in the scientific community. For some aspects, it is difficult to find a consensus on the best pipeline to be applied. This chapter goal was to guide RNA-seq users through its complex steps, providing a brief overview of the complete workflow, highlighting accessible protocols and currently available tools, most of which correlated with supporting benchmark studies.

Author details

Michele Araújo Pereira[1]*, Eddie Luidy Imada[2] and Rafael Lucas Muniz Guedes[1]

*Address all correspondence to: michele.pereira@hermespardini.com.br

1 Hermes Pardini Group, Vespasiano, Brazil

2 Federal University of Minas Gerais, Belo Horizonte, Brazil

References

[1] Ramsay G. DNA chips: State-of-the art. Nature Biotechnology. Jan 1998;**16**(1):40-44

[2] Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. Science. 21 Jun 1991;**252**(5013):1651-1656

[3] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science (80-). 20 Oct 1995;**270**(5235):484-487

[4] Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, et al. CAGE: Cap analysis of gene expression. Nature Methods. 3 Mar 2006;**3**(3):211-222

[5] Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D. et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nature Biotechnology. Jun 2000;**18**(6):630-634

[6] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences. 1 Dec 1977;**74**(12):5463-5467

[7] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science (80-). 6 Jun 2008; **320**(5881):1344-1349

[8] Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics. Jan 2009;**10**(1):57-63

[9] Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. Nature Reviews Genetics. 17 May 2016 May;**17**(6):333-351

[10] van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends in Genetics. Sep 2014;**30**(9):418-426

[11] Metzker ML. Sequencing technologies—The next generation. Nature Reviews Genetics. Jan 2010;**11**(1):31-46

[12] Scotty. Available from: http://bioinformatics.bc.edu/marthlab/scotty/scotty.php [Accessed: 3 February 2017]

[13] Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher J-P. Calculating sample size estimates for RNA sequencing data. Journal of computational biology. Dec 2013;**20**(12):970-8

[14] Zhao S, Li C, Guo Y, Sheng Q and Shyr Y. RnaSeqSampleSize: RnaSeqSampleSize. R package version 1.8.0. 2017

[15] Shanker S, Paulson A, Edenberg HJ, Peak A, Perera A, Alekseyev YO, et al. Evaluation of commercially available RNA amplification kits for RNA sequencing using very low input amounts of total RNA. Journal of Biomolecular Techniques. Apr 2015;**26**(1):4-18

[16] Blainey P, Krzywinski M, Altman N. Points of significance: replication. Nature Methods.28 Aug 2014;**11**(9):879-880

[17] Fang Z, Cui X. Design and validation issues in RNA-seq experiments. Briefings in Bioinformatics. 1 May 2011;**12**(3):280-287

[18] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A. et al. A survey of best practices for RNA-seq data analysis. Genome Biology. 26 Dec 2016;**17**(1):13

[19] Liu Y, Zhou J, White KP. RNA-seq differential expression studies: More sequence or more replication? Bioinformatics. 1 Feb 2014;**30**(3):301-304

[20] Ching T, Huang S, Garmire LX. Power analysis and sample size estimation for RNA-Seq differential expression. RNA. Nov 2014;**20**(11):1684-1696

[21] Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V. et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. Jun 2016;**22**(6):839-851

[22] Gu X. Statistical detection of differentially expressed genes based on RNA-seq: From biological to phylogenetic replicates. Briefings in Bioinformatics. Mar 2016;**17**(2):243-248

[23] Rhoads A, Au KF. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics. Oct 2015;**13**(5):278-289

[24] Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, Williams BA. et al. Characterization of the human ESC transcriptome by hybrid sequencing. Proceedings of the National Academy of Sciences. 10 Dec 2013;**110**(50): E4821-E4830

[25] Love KR, Shah KA, Whittaker CA, Wu J, Bartlett MC, Ma D. et al. Comparative genomics and transcriptomics of *Pichia pastoris*. BMC Genomics. 5 Dec 2016;**17**(1):550

[26] Gao S, Ren Y, Sun Y, Wu Z, Ruan J, He B. et al. PacBio full-length transcriptome profiling of insect mitochondrial gene expression. RNA Biology. 16 Sep 2016;**13**(9):820-825

[27] Liu L, Li Y, Li S, Hu N, He Y, Pong R. et al. Comparison of next-generation sequencing systems. Journal of Biomedicine and Biotechnology. 2012;**2012**:1-11

[28] GLENN TC. Field guide to next-generation DNA sequencers. Molecular Ecology Resources. Sep 2011;**11**(5):759-769

[29] Chu Y, Corey DR. RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic Acid Therapeutics. Aug 2012;**22**(4):271-274

[30] Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: Key considerations in genomic analyses. Nature Reviews Genetics. 17 Jan 2014;**15**(2):121-132

[31] Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: A matter of depth. Genome Research. 1 Dec 2011;**21**(12):2213-2223

[32] Lei R, Ye K, Gu Z, Sun X. Diminishing returns in next-generation sequencing (NGS) transcriptome data. Gene. Feb 2015;**557**(1):82-87

[33] Hou R, Yang Z, Li M, Xiao H. Impact of the next-generation sequencing data depth on various biological result inferences. Science China Life Sciences. 8 Feb 2013;**56**(2):104-109

[34] Cho H, Davis J, Li X, Smith KS, Battle A, Montgomery SB. High-resolution transcriptome analysis with long-read RNA sequencing. Buratti E, editor. PLoS One. 24 Sep 2014;**9**(9): e108095

[35] Chhangawala S, Rudy G, Mason CE, Rosenfeld JA. The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome Biology. 23 Dec 2015;**16**(1):131

[36] Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F. et al. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nature Biotechnology. 18 May 2015;**33**(7):736-742

[37] Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nature Biotechnology. 13 Oct 2013;**31**(11):1009-1014

[38] Tilgner H, Raha D, Habegger L, Mohiuddin M, Gerstein M, Snyder M. Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. G3: Genes|Genomes|Genetics. Mar 2013;**3**(3):387-397

[39] Chang Z, Wang Z, Li G. The impacts of read length and transcriptome complexity for De Novo assembly: A simulation study. Papavasiliou FN, editor. PLoS One. 15 Apr 2014;**9**(4):e94825

[40] Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N. et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nature Methods. 15 Sep 2010;**7**(9):709-715

[41] Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR. et al. Library construction for next-generation sequencing: Overviews and challenges. Biotechniques. 1 Feb 2014;**56**(2):61-77

[42] Zhao W, He X, Hoadley KA, Parker JS, Hayes D, Perou CM. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics. 2014;**15**(1):419

[43] Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S. et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Research. 1 Oct 2009;**37**(18):e123-e123

[44] Illumina. Directional mRNA-Seq Sample Preparation Guide. Part # 15018460 Rev. A. Oct 2010. Available from: https://support.illumina.com/downloads/directional_mrna-seq_sample_preparation_guide.html [Accessed: 16 May 2017]

[45] van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: Tone down the bias. Experimental Cell Research. Mar 2014;**322**(1):12-20

[46] Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J. et al. The external RNA controls consortium: A progress report. Nature Methods. Oct 2005;**2**(10):731-734

[47] Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M. et al. Synthetic spike-in standards for RNA-seq experiments. Genome Research. 1 Sep 2011;**21**(9):1543-1551

[48] Hardwick SA, Chen WY, Wong T, Deveson IW, Blackburn J, Andersen SB. et al. Spliced synthetic genes as internal controls in RNA sequencing experiments. Nature Methods. 8 Aug 2016;**13**(9):792-798

[49] Munro SA, Lund SP, Pine PS, Binder H, Clevert D-A, Conesa A. et al. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. Nature Communications. 25 Sep 2014;**5**:5125

[50] Wong T, Deveson IW, Hardwick SA, Mercer TR. ANAQUIN: A software toolkit for the analysis of spike-in controls for next generation sequencing. Bioinformatics. 27 Jan 2017;btx038

[51] Nielsen H. Working with RNA. Methods in Molecular Biology. 2011;**703**:15-28

[52] Thatcher SA. DNA/RNA Preparation for molecular detection. Clinical Chemistry. 1 Jan 2015;**61**(1):89-99

[53] Sellin Jeffries MK, Kiss AJ, Smith AW, Oris JT. A comparison of commercially-available automated and manual extraction kits for the isolation of total RNA from small tissue samples. BMC Biotechnology. 14 Dec 2014;**14**(1):94

[54] Mueller O, Lightfoot S, Schroeder A. RNA Integrity Number (RIN) -Standardization of RNA quality control. Agilent Technologies. 2004;1-8. 5989-1165EN

[55] Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M. et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Molecular Biology. 31 Jan 2006;**7**(1):3

[56] Gallego Romero I, Pai AA, Tung J, Gilad Y. RNA-seq: Impact of RNA degradation on transcript quantification. BMC Biology. 2014;**12**(1):42

[57] Winnebeck EC, Millar CD, Warman GR. Why does insect RNA look degraded? Journal of Insect Science. Sep 2010;**10**(159):1-7

[58] Wieczorek D, Delauriere L, Schagat T. Methods of RNA quality assessment. Promega Corporation Website. October 2012;1-14. Available from: http://www.promega.com.br/resources/pubhub/methods-of-rna-quality-assessment [Accessed: 16 May 2017]

[59] Aranda R, Dineen SM, Craig RL, Guerrieri RA, Robertson JM. Comparison and evaluation of RNA quantification methods using viral, prokaryotic, and eukaryotic RNA over a 104 concentration range. Analytical Biochemistry. Apr 2009;**387**(1):122-127

[60] Ferreira EN, de Campos Molina G, Puga RD, Nagai MA, Campos AHJFM, Guimarães GC. et al. Linear mRNA amplification approach for RNAseq from limited amount of RNA. Gene. Jun 2015;**564**(2):220-227

[61] Sinicropi D, Morlan J, City F. Methods for depleting RNA from nucleic acid samples. US20110111409. Vol. 1; 2011

[62] Morlan JD, Qu K, Sinicropi D V. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. Dadras SS, editor. PLoS One. 10 Aug 2012;**7**(8):e42882

[63] Zhulidov PA. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Research. 13 Feb 2004;**32**(3):37e-37

[64] Yi H, Cho Y-J, Won S, Lee J-E, Jin Yu H, Kim S. et al. Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. Nucleic Acids Research. 1 Nov 2011;**39**(20): e140-e140

[65] Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM. et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nature Methods. 19 May 2013;**10**(7):623-629

[66] Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. Gilbert MTP, editor. PLoS One. 2011 30 Nov;**6**(11):e28240

[67] Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Research. 1 Jul 2010;**38**(12):e131-e131

[68] van Gurp TP, McIntyre LM, Verhoeven KJF. Consistent errors in first strand cDNA due to random hexamer mispriming. Gibas C, editor. PLoS One. 30 Dec 2013;**8**(12):e85583

[69] Gubler U, Hoffman BJ. A simple and very efficient method for generating cDNA libraries. Gene. Nov 1983;**25**(2-3):263-269

[70] Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD. Reverse transcriptase template switching: A SMART approach for full-length cDNA library construction. Biotechniques. Apr 2001;**30**(4):892-897

[71] RNA-seq data analysis bioinformatic tools. Available from: https://omictools.com/rna-seq-category [Accessed: 3 February 2017]

[72] Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M. et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research. 8 Jul 2016;**44**(W1):W3-10

[73] Illumina BaseSpace. Available from: https://basespace.illumina.com [Accessed: 3 February 2017]

[74] Geneious. Available from: http://www.geneious.com [Accessed: 3 February 2017]

[75] FASTQ description. Available from: https://en.wikipedia.org/wiki/FASTQ_format [Accessed: 3 February 2017]

[76] Picard. Available from: http://broadinstitute.github.io/picard [Accessed: 3 February 2017]

[77] BamUtil. Available from: http://genome.sph.umich.edu/wiki/BamUtil [Accessed: 3 February 2017]

[78] Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. BamTools: A C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 15 Jun 2011;**27**(12): 1691-1692

[79] Andrews S. FastQC: A quality control tool for high throughput sequence data. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc [Accessed: 3 February 2017]

[80] Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 15 Mar 2011;**27**(6):863-864

[81] Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on illumina NGS data analysis. Seo J-S, editor. PLoS One. 23 Dec 2013;**8**(12):e85024

[82] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2 May 2011;**17**(1):10

[83] FASTX-toolkit. Available from: http://hannonlab.cshl.edu/fastx_toolkit/index.html [Accessed: 3 February 2017]

[84] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 1 Aug 2014;**30**(15):2114-2120

[85] Williams CR, Baccarella A, Parrish JZ, Kim CC. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics. 25 Dec 2016;**17**(1):103

[86] MacManes MD. On the optimal trimming of high-throughput mRNA sequence data. Frontiers in Genetics. 31 Jan 2014;**5**:13

[87] Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. Frontiers in Bioengineering and Biotechnology. 25 Jun 2015;**3**:92

[88] Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 1 Nov 2011;**27**(21):2957-2963

[89] Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. 1 Mar 2014;**30**(5):614-620

[90] Le H-S, Schulz MH, McCauley BM, Hinman VF, Bar-Joseph Z. Probabilistic error correction for RNA sequencing. Nucleic Acids Research. 1 May 2013;**41**(10):e109-e109

[91] Song L, Florea L. Rcorrector: Efficient and accurate error correction for Illumina RNA-seq reads. Gigascience. 19 Dec 2015;**4**(1):48

[92] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. Oct 1990;**215**(3):403-410

[93] Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. Bioinformatics. 1 Dec 2012;**28**(24):3169-3177

[94] Homer N, Merriman B, Nelson SF. BFAST: An alignment tool for large scale genome resequencing. Creighton C, editor. PLoS One. 11 Nov 2009;**4**(11):e7767

[95] Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 1 Apr 2010;**26**(7):873-881

[96] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research. 1 Nov 2008;**18**(11):1851-1881

[97] Lee W-P, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. MOSAIK: A hash-based algorithm for accurate next-generation sequencing short-read mapping. Hsiao CK, editor. PLoS One. 5 Mar 2014;**9**(3):e90581

[98] Novoalign. Available from: http://www.novocraft.com [Accessed: 3 February 2017]

[99] Smith AD, Xuan Z, Zhang MQ. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics. 2008;**9**(1):128

[100] Rumble SM, Lacroute P, Dalca A V., Fiume M, Sidow A, Brudno M. SHRiMP: Accurate mapping of short color-space reads. Wasserman WW, editor. PLoS Computational Biology. 22 May 2009;**5**(5):e1000386

[101] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 4 Mar 2012;**9**(4):357-359

[102] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology. 2013;**14**(4):R36

[103] Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. Nature Methods. 9 Mar 2015;**12**(4):357-360

[104] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 15 Jul 2009;**25**(14):1754-1760

[105] Huang S, Zhang J, Li R, Zhang W, He Z, Lam T-W, et al. SOAPsplice: Genome-wide ab initio detection of splice junctions from RNA-Seq data. Frontiers in Genetics. 7 Jul 2011;**2**:46

[106] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 1 Jan 2013;**29**(1):15-21

[107] Chen LY, Wei K-C, Huang AC-Y, Wang K, Huang C-Y, Yi D, et al. RNASEQR—A streamlined and accurate RNA-seq sequence analysis program. Nucleic Acids Researc. 1 Mar 2012;**40**(6):e42-e42

[108] Lindner R, Friedel CC. A comprehensive evaluation of alignment algorithms in the context of RNA-Seq. Salzberg SL, editor. PLoS One. 26 Dec 2012;**7**(12):e52403

[109] Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Alioto T. et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nature Methods. 3 Nov 2013;**10**(12):1185-1191

[110] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N. et al. The sequence alignment/Map format and SAMtools. Bioinformatics. 15 Aug 2009;**25**(16):2078-2079

[111] Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 1 Oct 2015;btv566

[112] BAMstats. Available from: http://bamstats.sourceforge.net/ [Accessed: 3 February 2017]

[113] Lassmann T, Hayashizaki Y, Daub CO. SAMStat: Monitoring biases in next generation sequencing data. Bioinformatics. 1 Jan 2011;**27**(1):130-131

[114] Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics. 2 Dec 2008;**40**(12):1413-1415

[115] 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. Nature Biotechnology. 2 May 2010;**28**(5):511-515

[116] Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X. et al. Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature Biotechnology. 2 May 2010;**28**(5):503-510

[117] Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology. 18 Feb 2015;**33**(3):290-295

[118] Florea LD, Salzberg SL. Genome-guided transcriptome assembly in the age of next-generation sequencing. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 10(5):1234-1240

[119] Martin J, Bruno VM, Fang Z, Meng X, Blow M, Zhang T. et al. Rnnotator: An automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics. 2010;**11**(1):663

[120] Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD. et al. De novo assembly and analysis of RNA-seq data. Nature Methods. 10 Nov 2010;**7**(11):909-912

[121] Martin JA, Wang Z. Next-generation transcriptome assembly. Nature Reviews Genetics. 7 Sep 2011;**12**(10):671-682

[122] Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. A reference-free algorithm for computational normalization of shotgun sequencing data. 21 Mar 2012. arXiv:1203.4802v2 [q-bio.GN]. 1-18

[123] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 15 May 2011;**29**(7):644-652

[124] Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 15 Apr 2012;**28**(8): 1086-1092

[125] Jain P, Krishnan NM, Panda B. Augmenting transcriptome assembly by combining de novo and genome-guided tools. PeerJ. 15 Aug 2013;**1**:e133

[126] Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R. et al. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biology. 21 Dec 2014;**15**(12):553

[127] Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Research. 26 Aug 2016;(**8**): 1134-1144

[128] Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 1 Oct 2015;**31**(19):3210-3212

[129] Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Briefings in Bioinformatics. 1 Mar 2013;**14**(2):178-192

[130] Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L. et al. Using tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics. Mar 2013;**14**(2):193-202

[131] Schmid-Burgk JL, Hornung V. BrowserGenome.org: web-based RNA-seq data analysis and visualization. Nature Methods. 29 Oct 2015;**12**(11):1001-1001

[132] Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Briefings in Bioinformatics. Mar 2013;**14**(2):144-161

[133] Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences. 8 Dec 2012;**131** (4):281-285

[134] Liao Y, Smyth GK, Shi W. Feature counts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 1 Apr 2014;**30**(7):923-930

[135] Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 15 Jan 2015;**31**(2):166-169

[136] Tarazona S, Furió-Tarí P, Turrà D, Pietro A Di, Nueda MJ, Ferrer A. et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Research. 16 Jul 2015;gkv711

[137] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010;**11**(10):R106

[138] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 5 Dec 2014;**15**(12):550

[139] Robinson MD, McCarthy DJ, Smyth GK. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 1 Jan 2010;**26**(1): 139-140

[140] Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics. 1 Nov 2013;**14**(6):671-683

[141] Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;**12**(1):323

[142] Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology. 9 Dec 2012;**31**(1):46-53

[143] Glaus P, Honkela A, Rattray M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics. 1 Jul 2012;**28**(13):1721-1728

[144] Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, Leek JT. Ballgown bridges the gap between transcriptome assembly and expression analysis. Nature Biotechnology. 6 Mar 2015;**33**(3):243-236

[145] Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 4 Apr 2016;**34**(5):525-527

[146] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods. Apr 2017;**14**(4):417-419

[147] Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD. et al. Alternative expression analysis by RNA sequencing. Nature Methods. 12 Oct 2010;**7**(10):843-847

[148] Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME. et al. Quantitative assessment of single-cell RNA-sequencing methods. Nature Methods. 20 Oct 2013; **11**(1):41-46

[149] Boley N, Stoiber MH, Booth BW, Wan KH, Hoskins RA, Bickel PJ. et al. Genome-guided transcript assembly by integrative analysis of RNA sequence data. Nature Biotechnology. 16 Mar 2014;**32**(4):341-346

[150] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W. et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research. 1 Sep 1997;**25**(17):3389-3402

[151] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K. et al. BLAST+: Architecture and applications. BMC Bioinformatics. 2009;**10**(1):421

[152] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature Methods. 17 Nov 2014;**12**(1):59-60

[153] Zdobnov EM, Apweiler R. InterProScan—An integration platform for the signature-recognition methods in InterPro. Bioinformatics. Sep 2001;**17**(9):847-848

[154] Lowe TM, Eddy SR. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research. 1 Mar 1997;**25**(5):955-964

[155] Lagesen K, Hallin P, Rodland EA, Staerfeldt H-H, Rognes T, Ussery DW. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research. 16 Apr 2007;**35**(9):3100-3108

[156] Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 15 Sep 2005;**21**(18):3674-3676

[157] Musacchia F, Basu S, Petrosino G, Salvemini M, Sanges R. Annocript: A flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs. Bioinformatics. 1 Jul 2015;**31**(13):2199-2201

[158] Van Bel M, Proost S, Van Neste C, Deforce D, Van de Peer Y, Vandepoele K. TRAPID: An efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. Genome Biology. 2013;**14**(12):R134

[159] Trinotate. Available from: https://trinotate.github.io/ [Accessed: 3 February 2017]

[160] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 25 Oct 2005;**102**(43):15545-15550

[161] Black DL. Mechanisms of alternative pre-messenger RNA splicing. Annual Review of Biochemistry. Jun 2003;**72**(1):291-336

[162] Matlin AJ, Clark F, Smith CWJ. Understanding alternative splicing: Towards a cellular code. Nature Reviews Molecular Cell Biology. May 2005;**6**(5):386-398

[163] Roy BM, Haupt LR. Griffiths L. Review: Alternative Splicing (AS) of genes as an approach for generating protein complexity. Current Genomics. 1 Apr 2013;**14**(3):182-194

[164] Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. GMAP and GSNAP for genomic sequence alignment: Enhancements to speed, accuracy, and functionality. 2016;**1418**:283-334

[165] 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 Research. 1 Oct 2010;**38**(18):e178-e178

[166] Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Research. 1 Oct 2012;**22**(10):2008-2017

[167] Shen S, Park JW, Lu Z, Lin L, Henry MD, Wu YN. et al. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proceedings of the National Academy of Sciences. 23 Dec 2014;**111**(51):E5593-E5601

[168] Vitting-Seerup K, Porse B, Sandelin A, Waage J. SpliceR: An R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. BMC Bioinformatics. 2014;**15**(1):81

[169] Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods. 7 Nov 2010;**7**(12):1009-1015

[170] Hu Y, Huang Y, Du Y, Orellana CF, Singh D, Johnson AR. et al. DiffSplice: The genome-wide detection of differential splicing events with RNA-seq. Nucleic Acids Research. 1 Jan 2013;**41**(2):e39-e39

[171] Mertens F, Johansson B, Fioretos T, Mitelman F. The emerging complexity of gene fusions in cancer. Nature Reviews Cancer. 22 May 2015;**15**(6):371-381

[172] Capdeville R, Buchdunger E, Zimmermann J, Matter A. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. Nature Reviews Drug Discovery. Jul 2002; **1**(7):493-502

[173] Liu S, Tsai W-H, Ding Y, Chen R, Fang Z, Huo Z. et al. Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. Nucleic Acids Research. 18 Mar 2016;**44**(5):e47-e47

[174] Bashir A, Volik S, Collins C, Bafna V, Raphael BJ. Evaluation of Paired-End Sequencing strategies for detection of genome rearrangements in cancer. Ouzounis CA, editor. PLOS Computational Biology. 25 Apr 2008;**4**(4):e1000051

[175] Jia W, Qiu K, He M, Song P, Zhou Q, Zhou F. et al. SOAPfuse: An algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biology. 2013;**14**(2):R12

[176] Nicorici D, Satalan M, Edgren H, Kangaspeska S, Murumagi A, Kallioniemi O, et al. FusionCatcher—A tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv. 19 Nov 2014. 1:11

[177] Davidson NM, Majewski IJ, Oshlack A. JAFFA: High sensitivity transcriptome-focused fusion gene detection. Genome Medicine. 11 Dec 2015;**7**(1):43

[178] Carrara M, Beccuti M, Lazzarato F, Cavallo F, Cordero F, Donatelli S. et al. State-of-the-Art Fusion-Finder algorithms sensitivity and specificity. BioMed Research International. 2013;**2013**:1-6

[179] Lai EC. Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. Nature Genetics. 18 Apr 2002;**30**(4):363-364

[180] Iorio M V, Croce CM. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. EMBO Molecular Medicin. Mar 2012;**4**(3):143-159

[181] Tam S, Tsao M-S, McPherson JD. Optimization of miRNA-seq data preprocessing. Briefings in Bioinformatics. 1 Nov 2015;**16**(6):950-963

[182] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009;**10**(3):R25

[183] Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Research. Jan19 2009;(1):92-105

[184] Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Research. Jan 2012;**40**(1):37-52

[185] Mathelier A, Carbone A. MIReNA: Finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. Bioinformatics. Sep 15 2010;**26**(18):2226-2234

[186] Hackenberg M, Rodriguez-Ezpeleta N, Aransay AM. miRanalyzer: An update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic Acids Research. 1 Jul 2011;**39**(Suppl):W132-W138

[187] Garmire LX, Subramaniam S. Evaluation of normalization methods in mammalian microRNA-Seq data. RNA. Jun 2012;**18**(6):1279-1288

[188] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 20 Apr 2015;**43**(7):e47

[189] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: MicroRNA sequences, targets and gene nomenclature. Nucleic Acids Research. 1 Jan 2006;**34**(Database issue):D140-D144

[190] Klein RJ. Complement factor H polymorphism in age-related macular degeneration. Science (80-). 15 Apr 2005;**308**(5720):385-389

[191] Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature Genetics. 8 Sep 2013;**45**(10):1238-1243

[192] Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nature Reviews Genetics. Apr 2015;**16**(4):197-212

[193] Fehrmann RSN, Jansen RC, Veldink JH, Westra H-J, Arends D, Bonder MJ. et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genetics. Aug 2011;**7**(8):e1002197

[194] Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. PLoS Genetics. 1 Apr 2010;**6**(4):e1000888

[195] Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 1 Apr 2010;**464**(7289):768-772

[196] Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A. et al. Multiple common variants for celiac disease influencing immune gene expression. Nature Genetics. Apr 2010;**42**(4):295-302

[197] Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs K V. et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 5Aug 2010;**466**(7307):714-749

[198] Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. Tools and best practices for data processing in allelic expression analysis. Genome Biology. 17 Sep 2015;**16**:195

[199] Stevenson KR, Coolon JD, Wittkopp PJ. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. BMC Genomics. 2013;**14**(1):536

[200] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A. et al. The genome analysis toolkit: A map reduce framework for analyzing next-generation DNA sequencing data. Genome Research. Sep 2010;**20**(9):1297-1303

[201] GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. Jun 2013;**45**(6):580-585