A Guide to RNAseq Data Analysis Using Bioinformatics Approaches

12

Preeti Sharma, B. Sharan Sharma, and Ramtej J. Verma

Abstract

The emergence of Next Generation Sequencing (NGS), such as DNA, RNA and other small RNA sequencing technologies, gave rise to a huge amount of raw data on a massive scale. To analyse that data and to obtain the biological interpretation as a challenging act, advancements in computational biology and bioinformatics applications emerged as the need of the hour. RNAseq accounts for exploration of comprehensive expression profile of genes and quantifies the presence of RNA content in the biological sample. In addition to this, RNAseq also provides information for alternative splice variants, novel gene identification, differentially expressing genes, etc. The workflow for RNAseq data analysis requires quality check of the data, mapping onto a reference genome/transcriptome, read quantification, differential expression analysis and functional annotation. Various tools and softwares with different algorithms have been developed to provide biological understanding of the data and to meet the demands of the analyst. An overview of the tools and softwares has been provided in the chapter that can be exploited to analyse the data for different investigations. Also, a glimpse of

P. Sharma (✉)
Department of Zoology, Biomedical Technology and Human Genetics, University School of Sciences, Gujarat University, Ahmedabad, Gujarat, India

PanGenomics International Pvt Ltd, Sterling Accuris Diagnostics, Ahmedabad, Gujarat, India

B. S. Sharma
Genexplore Diagnostics and Research Centre, Ahmedabad, Gujarat, India

Rivaara Labs Pvt Ltd, KD Hospital, Ahmedabad, Gujarat, India

R. J. Verma
Department of Zoology, Biomedical Technology and Human Genetics, University School of Sciences, Gujarat University, Ahmedabad, Gujarat, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-33-6191-1_12

other RNAseq techniques such as single cell RNAseq and small RNA sequencing has been discussed as an introduction to newer forms of RNA sequencing.

Keywords

Next generation sequencing · Transcriptome · Pre-processing · Quantification · Normalization

12.1 Introduction

With the advent of NGS technologies, RNA sequencing (RNAseq) occurred as a pivotal approach to evaluate the expression of a whole genomic profile. Sooner, the technique was exploited tremendously for certain advantages over others, such as identification of novel genes, unlike microarrays, detection of alternative splice variants, differentially expressing transcripts, etc.

The vast and varied applicability of RNAseq by offering results in multiple forms led to the generation of huge loads of data, also referred to as 'Big Data'. Resultantly, the technological expansion in the era of NGS also directed the evolution in the field of computational biology. Different tools and softwares were developed to analyse and interpret the results from the data generated on different platforms, such as SoLiD sequencing, Ion Torrent Platform, Illumina sequencing, etc. The procedure for RNAseq data analysis takes place in a number of steps which involves cDNA preparation, fragmentation followed by adapter ligation, cDNA library preparation and amplification (Han et al. 2015), etc. The fragments are read and sequenced to obtain the raw sequence data in the prescribed formats. These raw data sequences are then analysed to extract meaningful results from the sequences using various tools and pipelines.

The workflow for data analysis involves quality check and pre-processing of the raw reads, assembly to a reference genome, quantification of transcripts and identification of differentially expressed transcripts. The transcripts of interest are then annotated to different databases for functional enrichment, gene ontology analysis and pathway enrichment, etc. (Garber et al. 2011). A schematic workflow of the steps involved in data analysis has been shown in Fig. 12.1.

To explore deep into the genome or transcriptome (Sharma et al. 2020), other RNAseq technologies such as single cell RNAseq, small RNA sequencing etc. were developed. The development of these modified versions of RNAseq technologies also led to certain variabilities during sample processing, technical noise, normalization processes, etc. The challenges in data analyses for these processes accounted for advancements in development of computational tools and bioinformatics applications with certain modifications.

The present chapter provides an overview of workflow for analysis of RNAseq data on different sequencing platforms using bioinformatics approaches. Also, a brief outlook of different tools and softwares, based on different algorithms, can provide an understanding of using them in multiple dimensions depending upon the type of analysis to be performed (Table 12.1).

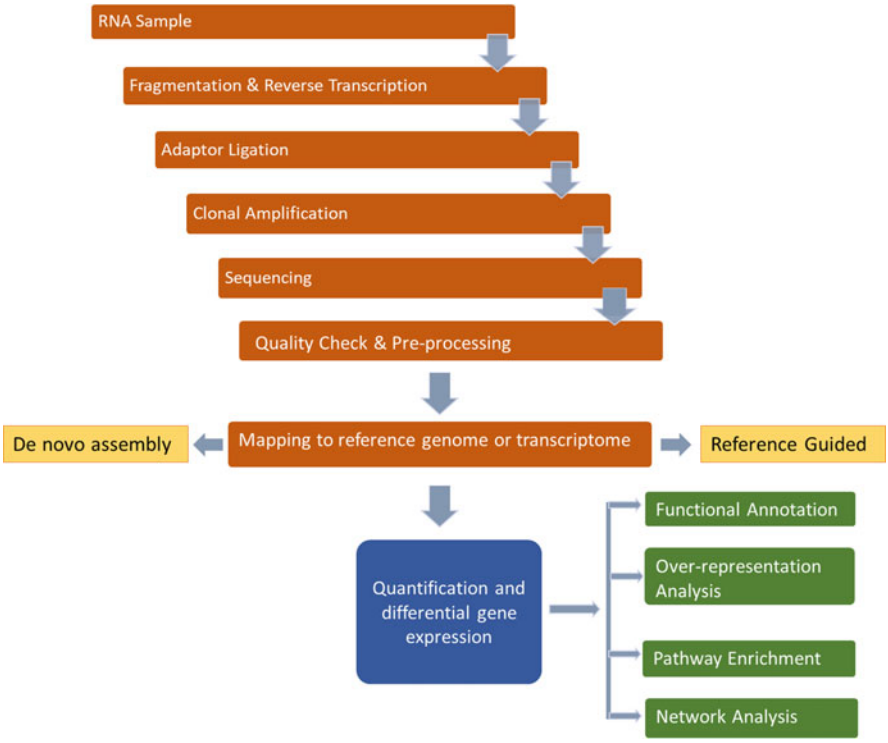

Fig. 12.1 Schematic Workflow showing steps in RNAseq data analysis

12.2 Platforms Available for Sequencing

Since the commencement of sequencing technologies various platforms have been developed which are based on different chemistries. The differences in the sequencing platforms also lie in the data output, performance and data quality. Some of the sequencing platforms and chemistries are discussed below:

12.2.1 SOLiD

SOLiD stands for Sequencing by Oligo Ligation and Detection and the technique was developed in 2005 (Hedges et al. 2011). It is based on oligonucleotide ligation to ligate dsDNA strands with the help of enzyme DNA ligase. A primer-binding adapter is bound to the target sequence on a bead, which is then amplified using emulsion PCR. A universal primer is hybridized to the adapter, followed by exposure of beads to a library of 8-nucleotide probes tagged with four different fluorescent dyes at 5'end and a hydroxyl group at 3'end. Based on the complementarity of

Table 12.1 List of tools available for different analytical processes of RNAseq data analysis

S. no.	Process	Tool	Link
1.	Quality check	FastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
		Kraken	https://github.com/DerrickWood/kraken2
		HTSeq	https://htseq.readthedocs.io/en/master/
		NGS QC Toolkit	http://www.nipgr.res.in/ngsqctoolkit.html
		RNASeQC	https://github.com/getzlab/rnaseqc
2.	Pre-processing	BBDuk	https://github.com/BioInfoTools/BBMap/blob/master/sh/bbduk.sh
		Cutadapt	https://bioinformaticshome.com/tools/rna-seq/descriptions/cutadapt.html
		FASTX Toolkit	http://hannonlab.cshl.edu/fastx_toolkit/
		SortMeRNA	https://bioinfo.lifl.fr/RNA/sortmerna/
		Trimmomatic	https://github.com/timflutre/trimmomatic
3.	Alignment of reads		
	Reference guided	Bowtie	http://bowtie-bio.sourceforge.net/index.shtml
		Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
		Burrows-Wheeler Aligner (BWA)	http://bio-bwa.sourceforge.net/
		Bayesembler	https://github.com/bioinformatics-centre/bayesembler
		Cufflinks	http://cole-trapnell-lab.github.io/cufflinks/
		IsoLasso	http://alumni.cs.ucr.edu/~liw/isolasso.html
	De novo assemblers	CLC Genomics Workbench	https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-clc-genomics-workbench/
		Oases	https://github.com/dzerbino/oases
		rnaSPAdes	https://cab.spbu.ru/software/rnaspades/
		Rnnotator	https://www.osti.gov/biblio/1231732-rnnotator
		SOAPdenovo-trans	http://sourceforge.net/projects/soapdenovotrans/
		Trans-ABySS	https://github.com/bcgsc/transabyss
		Trinity	https://github.com/trinityrnaseq/trinityrnaseq/wiki
		Velvet	https://www.ebi.ac.uk/~zerbino/velvet/
4.	Assembly evaluation tools	Busco	https://busco.ezlab.org/
		Detonate	http://deweylab.biostat.wisc.edu/detonate/
		rnaQUAST	https://github.com/ablab/rnaquast
		TransRate	https://hibberdlab.com/transrate/

(continued)

Table 12.1 (continued)

S. no.	Process	Tool	Link
	Co-expression networks		http://gnw.sourceforge.net/
		WGCNA	http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA.
5.	Functional, network and pathway analysis tools	BioCyc	https://biocyc.org/
		FunRich	http://www.funrich.org/
		GeneSCF	http://genescf.kandurilab.org/
		GOexpress	http://bioconductor.org/packages/release/bioc/html/GOexpress.html
		PathwaySeq	https://rna-seqblog.com/pathwayseq-pathway-analysis-for-rna-seq-data/
		ToPASeq	https://www.bioconductor.org/packages/release/bioc/html/ToPASeq.html
		RNA-Enrich	http://lrpath.ncibi.org
6.	miRNA prediction and analysis	miRDeep2	https://www.mdc-berlin.de/content/mirdeep2-documentation
		miRExpress	http://mirexpress.mbc.nctu.edu.tw/
		miR-PREFeR	https://github.com/hangelwen/miR-PREFeR
		miRDeep-P	http://faculty.virginia.edu/lilab/miRDP/
		miRPlant	http://www.australianprostatecentre.org/research/software/mirplant
		ShortStack	https://github.com/MikeAxtell/ShortStack
		mireap	https://github.com/liqb/mireap

first two bases, the probes get attached to the target sequence with the help of the enzyme DNA ligase. The fluorescent tag is then cleaved from the fragment at 5th and 6th base of the probe which is joined by phosphorothioate linkage. The fluorescence of the dyes generated due to cleavage is measured at different spectra. After the completion of first round of sequencing, the second-round sequencing starts with primer of length N-1, and so on. The sequencing of the target is ensured by measuring the fluorescence signals at each round of sequencing. However, the technique was low-cost and provided results with high accuracy due to the two-base sequencing, the main disadvantages were the time-consumption and shorter read lengths (Wyrzykiewicz and Cole 1994).

12.2.2 Ion Torrent Semiconductor Sequencing

The Ion Torrent sequencing is well-versed as 'semiconductor sequencing', where the target is sequenced by measuring changes in the pH variation due to release of hydrogen ion after incorporation of a specific nucleotide (Quail et al. 2012). A cDNA library is prepared here by fragmenting the RNA using enzymatic

degradation. The fragmented libraries are then ligated with complementary probes embedded on beads and mixed with PCR reagents and oil to perform emulsion PCR. Here, each microsphere of emulsion, specifically known as Ion Sphere Particles (ISPs), is covered with multiple copies of same DNA fragment for clonal amplification. After amplification the ISPs with template fragment are enriched from the mixture using biotin labelled magnetic beads and the rest are melted off. The positive templates are then prepared for sequencing and loaded onto Ion chips which contain millions of microwells with many copies of single-stranded DNA template and other sequencing reagents such as DNA polymerase, dNTPs in each well. The incorporation of the complementary nucleotide results into the change in pH level and is converted to digital signals to obtain the sequence of the target sequence. The technology is not based on fluorescence signals and does not require optical reading for detection so the sequencing is rapid and number of bases gets incorporated in less time. The technology limits in reading of homopolymer sequences in the template, such as 'TTTTTT', and becomes challenging to distinguish between the multiple oligomers, resulting into an increase in the error rate (Merriman et al. 2012).

12.2.3 Illumina Sequencing Technology

Illumina sequencing also known as 'sequencing by synthesis' approach (Ansorge 2009). Here, the target sequence is cleaved into smaller fragments of 100–150 bp to form a library and is ligated to customized adapters followed by generation of multiple copies of the same read using PCR. The adapter ligated templates are then washed onto a flow cell where millions of clusters are formed by the process of 'bridge amplification' PCR. The amplification process is carried out with DNA polymerase and modified dNTPs with a terminator tagged with a fluorescent label corresponding to each base. This terminator blocks the addition of another nucleotide and only one base is added by the polymerase at a time. The fluorescence is detected by imaging the signals, indicating a base that has been added to the sequence. With the addition of four nucleotides, the terminators are removed preparing the slide for next cycle of sequencing. The signals are then converted to construct the entire sequence. As the sequencing takes place in fixed cycles and of uniform read length, the sequences generated are also of uniform length (Meyer and Kircher 2010).

12.3 Quality Check and Pre-Processing of Reads

12.3.1 Formats Available for Storage of Raw Data

The sequences, can be referred to as raw reads, generated by sequencing on different platforms are stored in multiple files of short reads. After sequencing, the raw data is generated and can be stored in different file formats such as FASTQ, FASTA, SAM/BAM, etc.

- FASTQ is the most commonly used file format. It allows storing of data with corresponding quality values known as Phred scores. The files in fastq format are with extension '.fq' or '.fastq'. A FASTQ file contains four lines of textual information. The first line starts with a sign '@', generally known as a sequence identifier. The second line consists of a sequence of nucleotides, i.e. A, T, G, C. The third line consists of a '+' sign which is usually a separator and indicates the end of the sequence. The fourth line provides a quality score corresponding to the sequence in the second line (Deorowicz and Grabowski 2011).
- FASTA format is also one of the data storing formats and is available with extension '.fa' and '.fasta'. The sequences are recognized by a '>' sign in the beginning followed by a descriptive information about the sequence. This format is generally used while alignment or reference genome mapping by different tools and softwares. The sequence consists of nucleotides A, T, G, C and N (for undetermined base) (Gilbert 2003). The sequence can be viewed using text editor tools or LINUX/UNIX environment.
- BAM/SAM—The raw sequence data generated from the sequencer have no genomic information and are need to be aligned to a reference genome. After mapping or aligning to a reference genome, the output is generated in SAM/BAM format. SAM is Sequence Alignment/Map format which stores the sequences in an aligned format against the reference genome. A SAM file is a tab-delimited file, recognized by a '.sam' extension and can be viewed using text editor tools (Li et al. 2009). A BAM file is binary version of SAM file and is often found with '.bam' extension (Niemenmaa et al. 2012).

12.3.2 Quality Check Using Available Softwares and Tools

The data generated after sequencing often contains contaminants such as poor-quality reads, PCR artefacts, adapter sequences, over-represented sequences, etc. which interferes in downstream analytical operations of the data. Hence, the data needs to be quality checked to obtain clean and filtered high quality reads. For this, many softwares are available to assess the quality of the reads. These softwares perform a quality check (QC) on the data and provide a QC report depicting low-quality sequencing reads impeding the quality of the data. FASTQC is a commonly used tool for assessing the quality of the data. It measures scores associated with data such as read length, quality score, GC percentage, k-mers, etc. and produces results in different modules (Andrews 2010).

The *per base sequence quality* module assesses the overall quality of the bases at each position of the read which is represented by a box whisker plot. A higher score determines better quality of the base call. Likewise, *per sequence quality score* report presents a subset of overall sequences having low-quality scores. This constitutes a small fraction of the total sequences; however, a large subset possessing bad quality scores indicates some systematic errors.

The *per base GC content* shows the GC content of each base in the sequence. A shift in the graph of GC content with the underlying genome indicates presence of

over-represented sequences creating a sequence bias. Further in this, *per sequence GC content* marks for GC content across whole length of sequences comparable to normal distribution plot of GC content. A shift of the plot from the normal distribution on the graph indicates some systematic bias which is independent of base position. Some other modules such as *per base N content*, *sequence length distribution, duplicate sequences, over-represented sequences* and *over-represented k-mers,* etc. also provide report for the quality of the data.

12.3.3 Pre-Processing of Data

Before using the data for functional annotation and differential expression, etc. it is required to be pre-processed for removal of contaminated reads. For this, various tools are available such as Fastx-toolkit (Gordon and Hannon 2010), NGStoolkit (Mulcare 2004), Trimmomatic (Bolger et al. 2014), etc. Fastx-toolkit is most commonly used tool to filter out the good data from the bad quality data. During the course of filtration, the data is processed for removal of low-quality bases, adapter sequences, and other such reads interfering with the quality of the data.

The sequencing data is often contaminated with adapter sequences which are synthetically designed fragments of DNA added to the target sequences. These sequences are generally removed by the sequencers after the completion of sequencing process. But less often they remain attached to the sequenced read and are responsible for background noise in the data. Various tools such as Cutadapt (Martin 2011), Trimmomatic (Bolger et al. 2014), etc. are most frequently used tools for removal of adapter sequences.

Other contaminants are bases with low-quality, i.e. those with high error rate of being incorrect. The quality of base is assigned by a phred score (Q score) value, which is commonly used to measure the accuracy of the base call while sequencing the read by the sequencer. A quality score of <20 is generally considered of poor quality with high chances of inaccuracy. Fastx-toolkit is the most commonly used tool to trim off the reads with phred score <20.

Few other sequences such as rRNA sequences also act as contaminants in case of whole transcriptome sequencing. To remove the rRNA reads, rRNAFilter (Wang et al. 2017), SortMeRNA (Kopylova et al. 2012) and RiboPicker (Schmieder and Edwards 2011) are commonly used tools for the process.

12.4 Assembling Reads to Reference Genome/Transcriptome

12.4.1 Alignment of Reads

The raw reads generated after sequencing are then mapped onto a reference genome or transcriptome of the same species or the nearest relative, whichever available. (Roberts et al. 2011; Trapnell et al. 2010). The mapping of reads is affected by complexities of the genome, polymorphisms, gene isoforms, alternative splicing, etc.

leading to reduced percentage of mapped reads. The percentage of reads assembled indicates the accuracy of the results and presence of contaminated sequences (Conesa et al. 2016). The mapping can be done uniquely to one position or can also be mapped to multiple reads due to presence of repetitive sequences. In case of reference transcriptome multiple reads are found more often because of the presence of all isoforms of genes in the transcriptome.

12.4.2 Reference Guided/de Novo Assembly

In reference guided assembly, the reads are mapped onto a reference genome or transcriptome, whichever available, to assemble them into transcripts. The reads to be mapped are split into parts where one part maps to the exonic part and the other one to the intronic region. Reads mapping on the reference genome minimizes the complexities in the assemblies as they are mapped specifically to their genomic locations (Voshall and Moriyama 2018). Several assemblers are available for reference guided assemblies, such as Bayesembler (Maretty et al. 2014), Cufflinks (Ghosh and Chan 2016), Stringtie (Pertea et al. 2015), etc. Different assemblers use different strategies to assemble reads with highest percentage of read coverage, such as Cufflinks uses few numbers of transcripts to assemble large number reads to the genome or transcriptome, whereas Bayesembler uses Bayesian likelihood to estimate the most likely combination of transcripts constructed for each splice junction. Other assemblers such as IsoLasso (Li et al. 2011) and iReckon (Mezlini et al. 2013) use L-1 norm and specific sparse constraints, respectively, to obtain possible transcripts combinations.

The reference guided assemblers use reference genomes to align the reads and assemble them into transcripts, where graphs are prepared and isoforms are considered as paths of graphs (Li and Xuejun 2016). The accuracy of the assembly depends on the availability of complete and good quality reference genome which are usually available for the model organisms such as human, mouse, rat, Arabidopsis, Oryza, etc., but not for non-model species.

Therefore, for species with no reference genome de novo or reference-independent method is used to construct the transcripts. The de novo assembly is based on generation of short fragments of reads known as k-mers which overlaps to form a de Bruijn graph structure (Martin and Wang 2011). The assemblage of contigs using different algorithms depends on the varying lengths of the k-mers. Shorter k-mers generally cover the reference sequences completely but also provides ambiguity because of the presence of multiple reads from different transcripts. In case of longer k-mers, ambiguity is resolved but also does not cover the entire region of the reference genome/transcriptome.

Various assemblers are available based on optimization of k-mer lengths for assemblage of contigs using different algorithms. SOAPdenovo-Trans (Xie et al. 2014) and Trinity (Freedman 2016) use the preferred k-mer lengths for producing the de Bruijn graph. Trinity is a package of three independent softwares: Inchworm, Chrysalis and Butterfly, where Inchworm assembles the transcripts, Chrysalis forms

the de Bruijn graph by clustering those transcripts and finally Butterfly evaluates the graphs and produces the full-length assembly (Grabherr et al. 2011). rnaSPAdes (Bushmanova et al. 2019) identifies the k-mer lengths based on the read data. rnaSPAdes is the optimized version of SPAdes (Bankevich et al. 2012), where three assemblies are produced and one can choose any of them depending upon the downstream analyses. The three assemblies contain, one assembled with all transcripts, assembly with long and highly expressing transcripts, and assembly with short and lowly expressing transcripts (Geniza and Jaiswal 2017). Another assembler Velvet/Oases assembles the contigs based on de Bruijn graph using short reads. Velvet assembles the contigs using the short reads which are then clustered into loci using Oases program (Schulz et al. 2012).

12.4.3 Quality Check (QC) of Assembled Reads

Before processing the data for further downstream analysis the assembled reads are checked for their quality. The quality metrics of the assembled reads can be evaluated using two different criteria, either by calculating number and length of contigs or by mapping the assembled reads to coded proteins for similarity search. Softwares such as rnaQUAST (Bushmanova et al. 2016), CD-HIT (Li and Godzik 2006), TransRate (Smith-Unna et al. 2016) and Bowtie (Langmead 2010), etc. can be used to measure the quality of the assembly by measuring the lengths of the contigs and N50 value of the assemblies (T O'Neil and Emrich 2013). N50 value is defined as the minimum contig length required to cover fifty percent of the genome. While N50 value is more suitable quality of a genome assembly, transcriptome assembly is checked by measuring their ExN50 value which is dynamic and real time estimation of the assembled reads (Geniza and Jaiswal 2017).

ExN50 calculates the highly expressing transcripts which accounts for half of the overall transcriptome data. Another criterion based on mapping of the assembled reads to the coded proteins provides more probable notion of completeness of the assembled transcripts. The similarity searches are generally done by aligning the assembled reads against well-annotated databases containing non-protein sequences, conserved domains of proteins with functional annotation or lineage dependent protein databases (Nakasugi et al. 2014). These include BLAST (Altschul et al. 1990), Pfam (Finn et al. 2014), UniProt/Swiss-Prot (Apweiler et al. 2004), BUSCO (Waterhouse et al. 2018), etc. However, the protein-coded similarity search is a more plausible metric of QC of an assembly, the performance is limited by the relatedness of the biological entity in question to the sequences present in the databases. The more the divergence of the organism, more will be the possibility of lower percentage of assembled reads and gaps in the assembly.

12.5 Expression Quantification and Differential Expression

The first approach for transcriptome quantification is done by quantifying the expression of number of reads of specific transcripts. The most likely used method is maximizing likelihood (Glebova et al. 2016), based on different variants of expectation maximization (EM) (Li and Dewey 2011; Li and Jiang 2012), min-cost flow (Tomescu et al. 2013) and regression (Li et al. 2011), etc. RNAseq by Expectation Maximization (RSEM) quantitates the expression at isoform level and produces the output with 95% confidence interval. Moreover, all approaches use sequence specific transcripts to assess the expression level of each transcript. RSEM processing requires transcript sequences produced by the assembler as reference transcript sequences for RNAseq analysis for species with only transcript sequences available (Li and Dewey 2011). The mapped reads on multiple isoforms can be used to quantitate the expression in terms of prospective measures such as counting Fragments Per Kilobase of transcript per Million (FPKM) (Trapnell et al. 2010).

Another most widely used tool Cufflinks-Cuffdiff (Trapnell 2013) upgraded to Cuffdiff2 provides more determined method for differential expression analysis at transcript level. The newer version Cuffdiff2 uses negative binomial model and provides FPKM reads after normalization using relative log expression and inter-sample normalization method Q (Trapnell 2013).

Normalization of read counts is one of the critical steps in differential analysis of RNAseq data. The primary step in this process is to equate the total read counts from different libraries, as the variation caused by sequencing depths and size of the library are not comparable directly. In association to the number of expressing reads and gene length, the expression analysis also depends on the sample RNA that is being processed. For instance, genes with high expression shares a large percentage of the total reads of the sample compared to the left-over reads. This could be compared to the samples where reads are distributed evenly, in which case these lowly expressed genes show false positive result of differential expression for those genes (Zyprych-Walczak et al. 2015).

12.6 Annotation

12.6.1 Functional Annotation

The output of differential gene analysis provides information for the altered expression level of particular set of genes, now the next step is to explore the biological function of the genes. This is done by analysing the functional aspects, interaction network, pathway analysis and gene ontology, etc. of the genes involved in different processes of the biological system.

For functional annotation of the genes, various databases such as PANTHER classification system (Mi et al. 2016), DAVID Gene Functional Classification Tool (Sherman et al. 2007), etc. are available which assign particular function to genes and categorize them into different protein classes and biological pathways based on

their over-representation analysis (ORA) in the data (Khatri et al. 2012). Based on similar biological functions, cellular localization and pathway annotation these genes are classified into different functional categories. The genes are analysed for their over-representation in the particular category by calculating their occurrence in the specific category compared to the proportion of genes accommodated in the same category. The results can further be evaluated for significant results by applying statistical tools such as Fisher's exact test, Hypergeometric correction, etc.

12.6.2 Pathway Analysis

Annotation of differentially expressed genes to different pathways ensues to offer biological insights of genes based on their functional and structural similarities. Few methods of pathway annotation involve categorization of genes into different pathways irrespective of the mechanistic model of the pathway (Zhao et al. 2016). Another method involves analysis of certain genes enriched more than the expected count. This is known as pathway enrichment analysis which provides more functional understanding to the gene sets obtained from sequencing data. Here, the over-represented pathways are identified with strong statistical significance, such as FDR (False Discovery Rate) and p-value, relative to the expected chance of occurrence, using ranking score, overlapping genes over the size of the pathway and pathway topology, etc.

Some databases identify the enriched genes by assigning a scoring system based on their position and interaction amongst other genes in the network. Resultantly, interacting genes obtain higher weightage compared to the non-interacting ones, showing the functional relatedness of few sets of genes (Zhao et al. 2016). The analysis involves identification of set of genes from the sequencing data, selection of statistically significant enriched pathways and visualization and graphical representation of the results.

12.6.3 Gene Ontology (GO) Analysis

Gene Ontology analysis is a method to distribute genes into hierarchical classification and their representation in graphical structure. GO classification is distributed into different terms in which the genes or gene products get distributed into Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). These GO terms can be defined as:

- *Biological Process*—defines the role of the genes in the biological processes of an organism, such as, transcription, translation, signalling, apoptosis, etc.
- *Molecular Function*—provides the information related to functional activity of the gene in molecular terms. These activities include protein binding, nuclease activity, protease activity, etc.

- *Cellular Component*—provides information for cellular localization of the gene product. This includes components such as nucleus, lysosome, plasma membrane, etc.

The GO terms are said to be loosely hierarchical based on the available information regarding their biological functions and localizations. Based on this information they can be arranged in terms of 'parent terms' or more specific 'child terms'.

GO analysis also provides information for genes that are over- or under-regulated under specific conditions. This is done by calculating the enrichment analysis for the over-representation of certain set of genes in a particular condition (Gene Ontology Consortium@2015). The results are statistically evaluated based on their p-values. Various tools such as WebGeStalt (Wang et al. 2013), Clusterprofiler (Yu et al. 2012), Gorilla (Eden et al. 2009), WEGO (Ye et al. 2006), etc. are widely used.

12.7 Other RNAseq Applications

12.7.1 Single Cell RNAseq

RNAseq provides information for expression profile for a population of millions of cells. But different population of cells behave distinctly in different tissues. Single cell RNAseq is a recently developed technique designed to explore the distinct expression profile of single gene entity. Several tools have been designed to improve the procedural factures in employing this technique, such as dividing and disintegrating the cells to obtain single cell molecule (Zappia et al. 2018).

Since transcriptomic profiles of bulk samples provide a comprehensive outlook of bulk population of cells, single cell RNA sequencing meant to decipher the distinctiveness of cells at individual level. This approach is an addition to identify distinguishing variations in gene expression which are more complex and understanding of biological diversities in cellular context. Different approaches are being used to achieve unbiased, high throughput single cell RNAseq with exhaustive quantitative information at individual scale (Avital et al. 2014). One such approach is droplet based single cell RNAseq, developed independently by Klein et al. (2015) and Macosko et al. (2015). This technology is based on identification of single cells by barcoding individual cells from bulk of cells and analysing them using high throughput sequencing.

Another approach developed recently for single cell RNAseq is based on differential analysis of discrete expression pattern in different biological conditions. The approach developed by Korthauer and his team uses simulated data to detect the variations in the differential patterns under given set of biological conditions using a modelling framework (Korthauer et al. 2015).

12.7.2 Small RNA Sequencing

Small RNAs, such as siRNA (small interfering RNA), miRNA (microRNA), etc. belong to class of non-coding RNAs that plays crucial roles in regulation of gene expression at transcriptional level. The developing technologies in high throughput sequencing opened new prospects to explore the world of the miRNAs (Sharma@2020). Despite their pivotal roles, miRNAs share very less percentage in the genome. In order to obtain a comprehensive profile of miRNAs, deep sequencing is performed which is a modified version of next generation sequencing, sequencing a genomic region hundred or thousand times and allowing to detect molecules present in rare volumes (Motameny et al. 2010).

Currently, only a small number of tools and pipelines are available for analysis of miRNA data which is also a major challenge faced by many researchers. The analysis of miRNA data involves:

(a) Pre-processing of the raw data to filter out low-quality reads and other non-coding RNAs such as rRNA, tRNA, snRNA, snoRNA, etc.
(b) Mapping of reads to miRbase (largest repository of published miRNA sequences and annotations of various organisms) (Griffiths-Jones et al. 2007) to obtain known or conserved miRNAs in an organism.
(c) Prediction of novel miRNAs in an organism based on generation of hairpin loop structure using an RNA folding algorithm.
(d) Quantification of miRNAs for detection of differentially expressing miRNAs.

Further, these miRNAs regulate expression of various genes by binding to the 3'UTR (untranslated region) of their target mRNAs with near specific complementarity. Based on the complementarity between miRNA and target mRNAs various tools have been developed to detect the potential targets of candidate miRNAs using different algorithms. Tools such as microrna.org (Betel et al. 2008) and TargetScan (Lewis et al. 2005) account for detection of target mRNAs by searching for the binding sites for specific miRNAs. Few other tools such as Pictar (Lall et al. 2006), RNAhybrid (Rehmsmeier et al. 2004), miTarget (Kim et al. 2006), miRDB (Wong and Wang 2015), DIANAmicroT (Maragkakis et al. 2009) also predict putative binding mRNAs for given miRNAs using different algorithms in the background.

Identification of target mRNAs also accounts for involvement of these target mRNAs in different molecular processes and significant pathways, which is done by functional annotation, gene ontology and pathway analysis, etc. This could provide information for miRNA-mRNA regulatory network and can further be exploited for disease aetiology and therapeutic interventions.

12.8 Concluding Remarks

The rapid increase in technological expansion in the current times resulted in a tremendous upsurge of NGS technologies such as DNA sequencing, RNA sequencing and other targeted sequencing projects (Sharma et al. 2016). But to translate the

data generated from sequencing, the prime requisite is development of appropriate, specialized and reliable tools and bioinformatics applications. RNA sequencing is an advanced technique of NGS which favours the quantification and presence of RNA content in the biological sample. It also infers the presence of post-transcriptional modifications, SNPs, mutations, alternative spliced transcripts and their association with disease pathogenesis (Conesa et al. 2016). The use of RNAseq technology for various applications on a massive scale also demands for development of computational tools and softwares, with significant and reliable results, to match the pace by analysis and interpretation of data parallelly.

However, RNAseq is a gold standard technique to generate a comprehensive profile of whole transcriptome and other small non-coding RNAs in the sample. It is also highly prone to biasness and discrepancies in the data due to RNA extraction process, fragmentation of RNA, cDNA synthesis, amplification and sequencing, etc. Hence, to avoid these inconsistencies various tools and pipelines have been developed, based on different algorithms, to avoid the artefacts generated at various steps during the process. Data normalization is one such step which is crucial to reduce the biasness in the data. Several researchers deliver different thoughts on using different tools for data normalization and to minimize the noise and obtain best possible results.

Furthermore, different analysis tools offer varied results depending on the algorithms and backend procedures they are based on, hence relying on single tool cannot be recommended to provide substantial results. Therefore, it is always advisable to go through different school of thoughts and use multiple tools to attain comprehensive and comparative values for conclusive considerations.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data

Ansorge WJ (2009) Next-generation DNA sequencing techniques. New Biotechnol 25:195–203

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M (2004) UniProt: the universal protein knowledgebase. Nucleic Acids Res 32:D115–D119

Avital G, Hashimshony T, Yanai I (2014) Seeing is believing: new methods for in situsingle-cell transcriptomics. Genome Biol 15:110

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477

Betel D, Wilson M, Gabow A, Marks DS, Sander C (2008) The microRNA.org resource: targets and expression. Nucleic Acids Res 36:D149–D153

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Bushmanova E, Antipov D, Lapidus A, Suvorov V, Prjibelski AD (2016) rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. Bioinformatics 32:2210–2212

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD (2019) rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience 8:giz100

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X (2016) A survey of best practices for RNA-seq data analysis. Genome Biol 17:13

Deorowicz S, Grabowski S (2011) Compression of DNA sequence reads in FASTQ format. Bioinformatics 27:860–862

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10:1–7

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230

Freedman A (2016) Best practices for de novo transcriptome assembly with trinity

Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 8:469–477

Geniza M, Jaiswal P (2017) Tools for building de novo transcriptome assembly. Curr Plant Biol 11:41–45

Ghosh S, Chan C-KK (2016) Analysis of RNA-Seq data using TopHat and Cufflinks. In: Plant bioinformatics. Springer, New York, pp 339–361

Gilbert D (2003) Sequence file format conversion with command-line Readseq. Curr Protoc Bioinformatics 00(1):A-1E.1–A-1E.4

Glebova O, Temate-Tiagueu Y, Caciula A, Al Seesi S, Artyomenko A, Mangul S, Lindsay J, Măndoiu II, Zelikovsky A (2016) Transcriptome quantification and differential expression from NGS data. In: Computational methods for next generation sequencing data analysis. Wiley, Hoboken, NJ, pp 301–327

Gordon A, Hannon G (2010) Fastx-toolkit. FASTQ/A short-reads pre-processing tools. Unpublished. http://hannonlab.cshl.edu/fastx_toolkit

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al (2011) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol 29:644

Griffiths-Jones S, Saini HK, Van Dongen S, Enright AJ (2007) miRBase: tools for microRNA genomics. Nucleic Acids Res 36:D154–D158

Han Y, Gao S, Muegge K, Zhang W, Zhou B (2015) Advanced applications of RNA sequencing and challenges. Bioinform Biol Insights 9:BBI-S28991

Hedges DJ, Guettouche T, Yang S, Bademci G, Diaz A, Andersen A, Hulme WF, Linker S, Mehta A, Edwards YJ (2011) Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. PLoS One 6:e18595

Khatri P, Sirota M, Butte AJ (2012) Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 8:e1002375

Kim S-K, Nam J-W, Rhee J-K, Lee W-J, Zhang B-T (2006) miTarget: microRNA target gene prediction using a support vector machine. BMC Bioinformatics 7:1–12

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161:1187–1201

Kopylova E, Noé L, Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28:3211–3217

Korthauer KD, Chu L-F, Newton MA, Li Y, Thomson J, Stewart R, Kendziorski C (2015) scDD: a statistical approach for identifying differential distributions in single-cell RNA-seq experiments. bioRxiv 035501

Lall S, Grün D, Krek A, Chen K, Wang Y-L, Dewey CN, Sood P, Colombo T, Bray N, MacMenamin P (2006) A genome-wide map of conserved microRNA targets in C. elegans. Curr Biol 16:460–471

Langmead B (2010) Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics 32:11–17

Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120:15–20

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323
Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659
Li W, Jiang T (2012) Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads. Bioinformatics 28:2914–2921
Li Z, Xuejun L (2016) A comprehensive review on RNA-seq data analysis. Trans Nanjing Univ Aeronaut Astronaut 33(3):339–361
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079
Li W, Feng J, Jiang T (2011) IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. J Comput Biol 18:1693–1707
Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161:1202–1214
Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K (2009) DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res 37:W273–W276
Maretty L, Sibbesen JA, Krogh A (2014) Bayesian transcriptome assembly. Genome Biol 15:501
Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12
Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12:671–682
Merriman B, Ion Torrent R&D Team, Rothberg JM (2012) Progress in ion torrent semiconductor chip based sequencing. Electrophoresis 33:3397–3417
Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc 2010:pdb-prot5448
Mezlini AM, Smith EJ, Fiume M, Buske O, Savich GL, Shah S, Aparicio S, Chiang DY, Goldenberg A, Brudno M (2013) iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. Genome Res 23:519–529
Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD (2016) PANTHER version 11: expanded annotation data from gene ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res 45:D183–D189
Motameny S, Wolters S, Nürnberg P, Schumacher B (2010) Next generation sequencing of miRNAs–strategies, resources and methods. Genes 1:70–84
Mulcare D (2004) NGS toolkit. Part 8: the National Geodetic Survey. NADCON tool. Prof Surv Mag 24(2):120–125
Nakasugi K, Crowhurst R, Bally J, Waterhouse P (2014) Combining transcriptome assemblies from multiple de novo assemblers in the allo-tetraploid plant Nicotiana benthamiana. PLoS One 9: e91776
Niemenmaa M, Kallio A, Schumacher A, Klemelä P, Korpelainen E, Heljanko K (2012) Hadoop-BAM: directly manipulating next generation sequencing data in the cloud. Bioinformatics 28:876–877
Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33:290–295
Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 13:1–13
Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes. RNA 10:1507–1517
Roberts A, Pimentel H, Trapnell C, Pachter L (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 27:2325–2329
Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–1092
Sharma P, Bhunia S, Poojary SS, Tekcham DS, Barbhuiya MA, Gupta S, Shrivastav BR, Tiwari PK (2016) Global methylation profiling to identify epigenetic signature of gallbladder cancer and gallstone disease. Tumor Biol 37:14687–14699
Sharma P, Kumar S, Beriwal S, Sharma P, Bhairappanavar SB, Verma RJ, Das J (2020) Comparative transcriptome profiling and co-expression network analysis reveals functionally coordinated genes associated with metabolic processes of Andrographis paniculata. Plant Gene 23:100234
Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA (2007) The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 8:R183
Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S (2016) TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res 26:1134–1144
T O'Neil S, Emrich SJ (2013) Assessing De Novo transcriptome assembly metrics for consistency and utility. BMC Genomics 14:465
Tomescu AI, Kuosmanen A, Rizzi R, Mäkinen V (2013) A novel min-cost flow method for estimating transcript expression with RNA-Seq. BMC Bioinformatics 14(Suppl 5):S15
Trapnell C (2013) Cufflinks. cuffdiff (v6). Open module on GenePattern public server. GenePattern. https://software.broadinstitute.org/cancer/software/genepattern/modules/docs/Cufflinks.cuffdiff/6
Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515
Voshall A, Moriyama EN (2018) Next-generation transcriptome assembly: strategies and performance analysis. In: Bioinformatics in the era of post genomics and big data. IntechOpen, London, pp 15–36
Wang J, Duncan D, Shi Z, Zhang B (2013) WEB-based gene set analysis toolkit (WebGestalt): update 2013. Nucleic Acids Res 41:W77–W83
Wang Y, Hu H, Li X (2017) rRNAFilter: a fast approach for ribosomal RNA read removal without a reference database. J Comput Biol 24:368–375
Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol 35:543–548
Wong N, Wang X (2015) miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic Acids Res 43:D146–D152
Wyrzykiewicz T, Cole D (1994) Sequencing of oligonucleotide phosphorothioates based on solid-supported desulfurization. Nucleic Acids Res 22:2667–2669
Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30:1660–1666
Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 34:W293–W297
Yu G, Wang L-G, Han Y, He Q-Y (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. Omics 16:284–287
Zappia L, Phipson B, Oshlack A (2018) Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput Biol 14:e1006245
Zhao S, Zhang B, Zhang Y, Gordon W, Du S, Paradis T, Vincent M, von Schack D (2016) Bioinformatics for RNA-seq data analysis. Bioinformatics—updated features and applications. InTechOpen, London, pp 125–149
Zyprych-Walczak J, Szabelska A, Handschuh L, Górczak K, Klamecka K, Figlerowicz M, Siatkowski I (2015) The impact of normalization methods on RNA-Seq data analysis. Biomed Res Int 2015:621690