

Chapter 2

Transcriptome Analysis Throughout RNA-seq

Tainá Raiol, Daniel Paiva Agostinho, Kelly Cristina Rodrigues Simi, Calliandra Maria de Souza Silva, Maria Emilia Walter, Ildinete Silva-Pereira and Marcelo Macedo Brígido

Abstract Differential gene expression profile is a powerful tool to identify changes in cell or tissue transcriptomes, which allows to understanding complex biological process such as oncogenesis, cell differentiation and host immunological response to pathogens, among others. To date, the gold standard technique to compare gene expression profile is micro-array hybridization of a RNA preparation. In recent years technological advances led to a new generation of sequencing methods, which can be explored to uncover the complete content of a cell transcriptome. Such a deep sequencing of a RNA preparation, named RNA-seq, allows to virtually detect the complete RNA content, including low abundant isoforms. The RNA-seq quantitative aspect may be further explored to detect gene differential expression based on a reference genome and gene model. In contrast to micro-arrays, RNA-seq may find a broader range of RNA isoforms as well as novel RNA molecules, and has been gradually substituting micro-arrays to differential gene expression profile. In this chapter we describe how deep sequencing may be used to describe changes in the gene expression profile, its advantages and limitations.

2.1 High-Throughput Sequencing Techniques

Since the development of Sanger's technology in the 70's, DNA sequencing has been continuously improved regarding to both throughput and low cost. Next generation sequencing (NGS), also called high-throughput or deep sequencing, constitutes a new breakthrough of increasingly research power, a revolutionary advance

M. M. Brígido (✉) · T. Raiol · K. C. R. Simi · C. Maria de Souza Silva · I. Silva-Pereira
Laboratório de Biologia Molecular, CEL/IB (Pós-graduação em Biologia Molecular/CEL/IB),
Universidade de Brasília, Brasília, DF 70910-900, Brazil
e-mail: brigido@unb.br

D. P. Agostinho
Laboratório de Biologia Molecular, CEL/IB (Pós-graduação em Patologia Molecular/FM/UnB),
Universidade de Brasília, Brasília, DF 70910-900, Brazil

M. E. Walter
Departamento de Ciência da Computação, Instituto de Ciências Exatas, Universidade de Brasília,
Brasília, DF 70910-900, Brazil

in molecular biology knowledge. An increasing number of biological questions may be addressed by NGS technologies, which provides a much larger comprehensive survey compared to the Sanger method, and under a system biology perspective. In particular, transcriptomics has been strongly benefited by the use of these new technologies, also called RNA-seq, allowing a complete characterization of whole transcriptome at both gene (Kvam et al. 2012) and exon (Anders et al. 2012) levels, and with an additional ability to identify rare transcripts, new genes, novel splicing junctions and gene fusions (Katz et al. 2010; Wang et al. 2009; Van Verk et al. 2013).

In this chapter, first we address a brief overview on sequencing techniques and the most common next-generation platforms, as well as computational methods for RNA-seq data analysis. After, we discuss two case studies to assess the capabilities of RNA-seq in addressing important biological issues.

2.1.1 Sanger's Sequencing Technology

In 1977, Frederick Sanger and colleagues (Sanger et al. 1977) developed the DNA sequencing method, which in 2001 allowed the first Human genome draft (Lander et al. 2001). This method called dideoxy chain-termination or simply Sanger method is based on the use of special nucleotide molecules (called ddTNP), lacking a 3'-OH at the deoxyribose, which blocks the DNA elongation. These special nucleotides are mixed in lower concentrations to the regular nucleotides and used as reagents for DNA polymerase reaction. Therefore, with the polymer synthesis stopped by the inclusion of a ddNTP, the last nucleotide can be determined. Each of the four ddNTPs was added separately in four different reactions. At the beginning, one of the regular nucleotides, most commonly dATP or dCTP, was radioactively labeled (e.g., ^{32}P or ^{35}S) in order to achieve the radioactive signal. Usually, polyacrylamide gel electrophoresis was used for separation of the DNA molecules, which diverged in length by a single nucleotide. Then the gel was dried and exposed to X-ray film.

An important modification of the method was the substitution of the radioactive label by a fluorescent dye (Smith et al. 1986). Each distinct wave length produced by the fluorescent dyes linked to dideoxynucleotides corresponds to a different nucleotide, with the four sequencing reactions performed in the same tube. With the automation of the Sanger sequencing method, the performance reached up to 96 different reactions running in parallel capillary gel electrophoresis (Marsh et al. 1997), which is considered the first-generation technology. In the top of the technology 384 samples could be sequenced at once in a single multi-well plate. The main sequencing devices using Sanger method are ABI (Applied Biosystems) and MegaBACE (GE Healthcare Life Sciences).

The main advantage of Sanger sequencing is the length of the produced sequences, about 1000 kb, which is still unreachable by the main NGS technologies nowadays. However, deep sequencing has the advantage of high coverage, i.e., a large amount of redundant data, further treated through bioinformatics analysis, generating much more informative data in a single run.

Table 2.1 Comparison of next generation sequencing technologies

	Sanger ABI 3730xl	454 GS FLX	HiSeq 2000	SOLiDv4	Ion Torrent PGM (318 chip)
Read length (bp)	900	700	150	85	100
Cost (US\$/Mb)	500	12.56	0.02	0.04	0.63
Output data/run	2.88 Mb	0.7 Gb	600 Gb	30 Gb	1 Gb
Time run	3 h	1 day	8 days	7 days	3 h

2.1.2 Next Generation Sequencing

Regulatory mechanisms and gene expression profiles have been widely investigated towards elucidation of several essential cellular processes. Hybridization-based technology, e.g., microarray, has been very useful for determining global gene expression. However, the high background levels due to cross-hybridization, a limited range of quantification and a restricted detection of known genes are bottlenecks for large scale use of this technique (Shendure 2008). RNA-seq allows a genome-scale transcriptome analysis, including novel genes and splice variants, with a large range of quantification and reduced sequencing costs (Wang et al. 2009; Soon et al. 2013). These advantages make RNA-seq a better and attractive solution for whole-genome transcriptome analysis of several organisms, even for those with no sequenced reference genomes.

Nowadays, the most commonly used NGS platform for RNA-seq research is the Illumina HiSeq. A comparison of NGS technologies is shown in Table 2.1 based on data from Liu et al. (2012).

The enormous amounts of data generated by NGS create new challenges to the downstream bioinformatics analysis, which has to handle with large sequence files while searching for comprehensive and useful biological information, discussed later in this chapter.

2.1.3 454 Sequencing

In 2005, the 454 sequencing platform was formally announced by Roche as a new massive parallelized sequencer (Margulies et al. 2005). It was the first technology, among several others, considered as next generation sequencing. Since 454 produces the largest sequences among the NGS platforms, it is mainly used for transcriptome studies concerning organisms without a reference genome.

The pyrosequencing method used by 454 sequencing is based on the detection of pyrophosphate released during the nucleotide incorporation promoted by DNA polymerase (Harrington et al. 2013; Mardis 2013). In contrast to the Sanger sequencing, pyrosequencing is designated as a sequence-by-synthesis technique because DNA synthesis is monitored in real time. Single-stranded DNA library is generated after fragmentation and addition of adaptors to both fragment ends

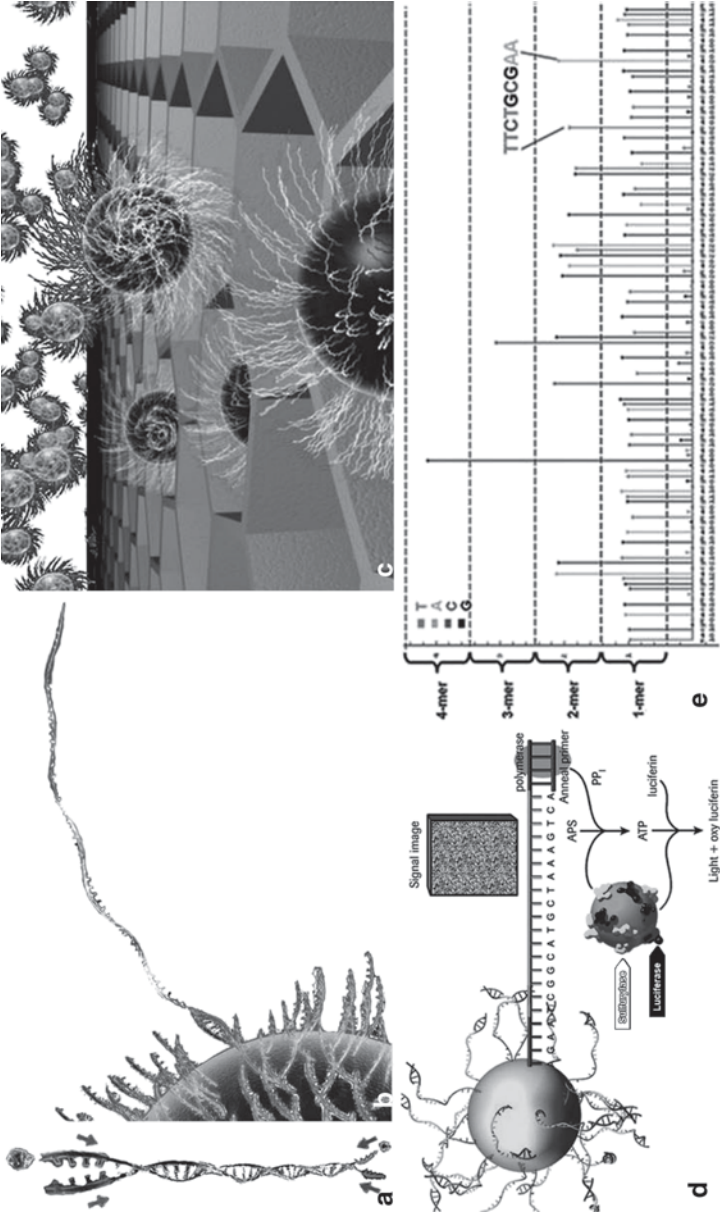


Fig. 2.1 The 454 sequencing technology. **a** Adaptors are added to the DNA fragments to be used in subsequent sequencing steps. **b** The library is attached to beads. Each bead carries a unique single-stranded library fragment. **c** The beads are loaded onto the PicoTiterPlate device, where the surface design allows for only one bead per well. **d** Pyrosequencing reaction of 454 sequencing systems. When a nucleotide complementary to the template strand is incorporated, a chemiluminescent light signal is recorded by the camera. **e** Flowgram or pictogram corresponds to the signal intensity record of each incorporation event at each well position, determining the sequence of all the reads in parallel. (Source: <http://454.com/>)

(Fig. 2.1a and b). One single fragment is ligated to beads covered by adaptors to proceed to the clonal amplification by emulsion PCR. Bead-ligated sequence is added, along with amplification reagents, in a water-in-oil mixture to trap individual beads in amplification micro-reactors. Next, the bead-ligated amplified sequences are added to the PicoTiterPlate device containing millions of 28 μm wells,

the precise size for a single bead (Fig. 2.1c). To these wells, enzyme beads (containing sulfurylase and luciferase) are added to the device. Each nucleotide is added to the system separately during the sequencing rounds. With the incorporation of one nucleotide, a pyrophosphate is released and used by sulfurylase to convert ADP into ATP, the substrate of luciferase (Fig. 2.1d). ATP and luciferin are used by luciferase to produce luminescence, which is detected by a visible-light high-sensitivity CCD camera. Apyrase is subsequently added to remove any non-incorporated nucleotide, and, then, the next round is initiated. The signal strength is proportional to the number of added nucleotides, recorded as a pyrogram (Fig. 2.1e). Sequences are stored as standard flowgram format (SFF), a binary format that is further converted in the FASTQ format, used in the bioinformatics analysis.

2.1.4 *Illumina Sequencing*

Illumina sequencing uses reversible dye-terminator technique that adds a single nucleotide to the DNA template in each cycle (Bentley et al. 2008). This system was initially developed in 2007 by Solexa and was subsequently acquired by Illumina, Inc. Illumina is widely used in several whole transcriptome studies since it reaches the deepest depth among NGS technologies. However, the small sequence size (around 100 bp) hampers the assembly into contigs as normally used for Sanger and 454 sequencing. Therefore, a reference genome is usually necessary for Illumina data analysis.

As 454, Illumina sequencing is based on sequencing-by-synthesis, however, instead of clonal amplification using beads, Illumina sequencing is performed in a solid slide covered by adaptors complementary to those added to the fragmented DNA sequences (Metzker 2010). This procedure, called bridge PCR, consists in amplification of bended DNA sequences, attached by both ends to the solid surface (Fig. 2.2a). By the end of the clonal amplification, clusters of identical DNA sequences will be formed in order to amplify the fluorescence signals. In each round, one single nucleotide is added to the single-strand template sequences followed by fluorescence detection by a high-sensitivity CCD camera (Fig. 2.2b and d). As in Sanger's technology, different fluorophore molecules are attached to each nucleotide, however, these molecules hamper the polymerase to add new ones. The fluorescence emission releases the 3'OH of the recent added nucleotide allowing it to receive new monomers in the next sequencing round.

Single-end sequencing, i.e., reads generated from a single end adaptor, is being replaced by the paired-end sequencing, since the accuracy in downstream analysis is greater with a fairly price. Paired-end reads are produced from the adaptor priming sites in both template sequence ends, being the second adaptor primer used in a subsequent sequencing run (Fig. 2.2c).

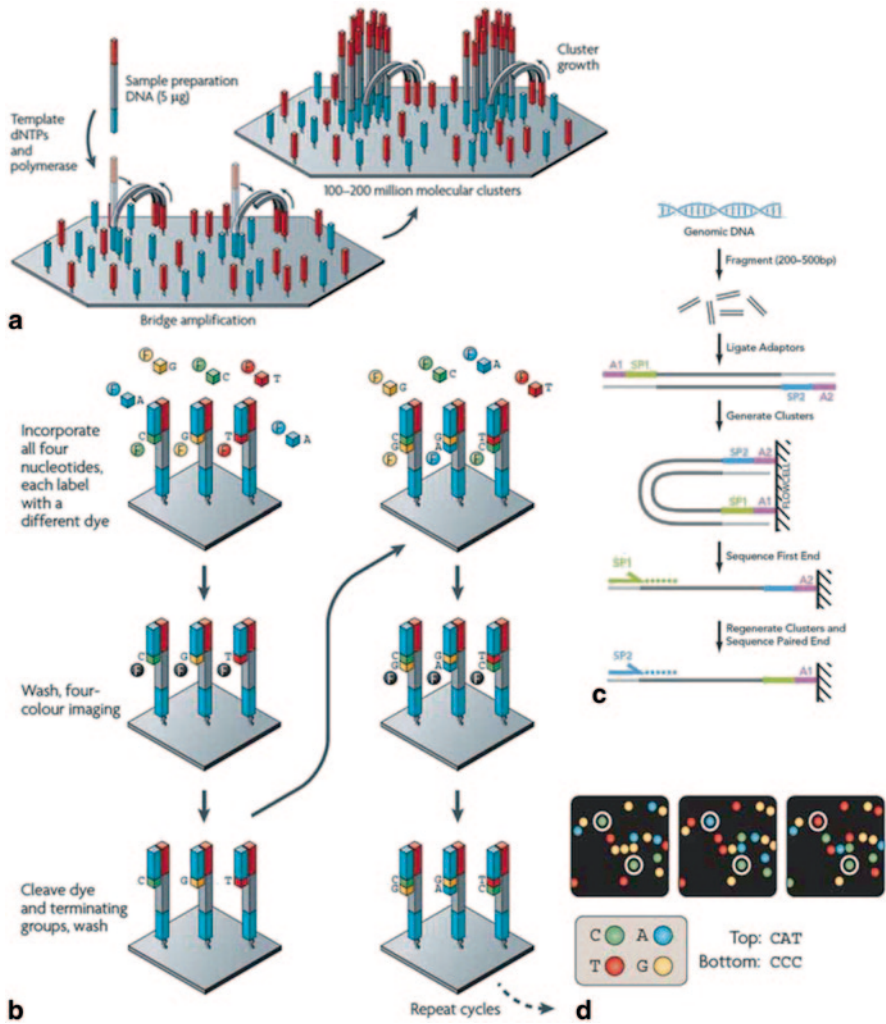


Fig. 2.2 The Illumina sequencing technology. **a** Two basic steps encompass an initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template in a solid device with immediately adjacent primers to form clusters. **b** The four-color cyclic reversible termination (CRT) method uses terminator chemistry. A cleavage step removes the fluorescent dyes and regenerates the 3'-OH group. **c** Paired-end sequencing by which reads are generated from both template strand. "A" block indicates the device-ligation adaptors and "SP", sequencing primers. **d** In the images, the sequencing data is highlighted from two sequence clusters. (Source: Metzker 2010 and <http://www.illumina.com/>)

2.2 Bioinformatics Pipelines for Transcriptome Projects

As described previously, Illumina sequencing has been commonly used in transcriptome projects, since the volumes of sequenced reads (named *raw data*) allow to finding virtually the total of the expressed genes (transcripts). Due to the short

lengths of the Illumina reads, they are usually mapped in a reference genome, the mapped regions indicating the expressed genes of the RNA-seq sample. If the organism genome is not yet sequenced, new specially developed methods to handle short reads have been used to reconstruct the transcript sequences, e.g., 454 sequencing produces sequences four times larger than those produced by Illumina. Also in this case, the original reads are usually assembled in larger sequences, in order to rebuild each (fragment of) transcript. In both cases, the metaphor for reconstructing the transcripts is like mounting a puzzle, where the pieces (the reads) have to be assembled (relative to a reference genome or not) to obtain the picture (the transcripts of the transcriptome). After this, different analyses can be performed on these reconstructed transcripts, e.g., quantitative analysis and differential expression. In transcriptome projects, the tasks of reconstructing transcripts and performing biological analyses are performed by bioinformatics pipelines, discussed next.

2.2.1 Pipelines

A bioinformatics pipeline is a computational system composed of a sequence of softwares (computer programs), sequentially executed, in which the output data from one software is the input data for the following software.

In general, transcriptome bioinformatics pipelines have the following phases, which can be combined according to the input raw data and the objectives of each project:

- *filter (or clean)* raw data for quality assessment: this is usually performed in two steps as follows. In the *clipping step*, a fragment (or the whole read) containing adapters is removed, while in the *trimming step*, reads are filtered to remove low quality sequences. This filtering phase guarantees a reliable dataset of quality short reads, to be used in the following phases of the pipeline;
- *map* short reads to reference genomes: the filtered reads are aligned to a reference genome, in order to find the genomic regions presenting matches with these reads;
- *assembly (or group)* reads: each group of reads (called *contig*), composed of reads having similar extremities (the end of a read is similar to the beginning of another read), allows to construct one larger sequence (called *consensus*), which is a predicted (fragment of) transcript;
- *analysis* of the set of (fragments of) transcripts obtained from the mapping or the assembling phase: allows to obtain relevant biological information, e.g.,
 - *quantitative analysis*: among others, coverage analysis shows the abundance of genes expressed in one RNA-seq sample, more precisely, the number of reads mapped in a certain region of the chromosome
 - *differential expression*: allows to analyze the variability of genetic expression between samples
 - *annotation*: assigns a biological function to each transcript

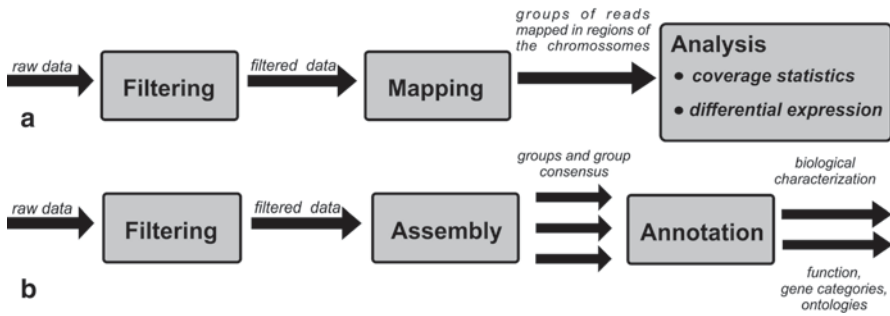


Fig. 2.3 **a** Pipelines for short reads, with a well-characterized reference genome, and two types of analyses—coverage statistics and differential expression. **b** Pipeline for longer reads, with no reference genome, and annotation (biological function, gene categories and ontologies)

Two generic bioinformatics pipelines for transcriptomes are discussed next, although the design of a particular pipeline depends on the objectives of the transcriptome project and other information, e.g., the sequencer (since the sequencing techniques may cause specific errors in the raw data, which have to be treated), and availability of information to be used in the analysis phase (e.g., quantitative and differential expression softwares and availability of reference genome).

Pipeline 1 The organism of interest has at least one reference genome already sequenced, with well-annotated genes and other biological characteristics, and the reads are short (about 100 bp, e.g., short reads produced by Illumina). A pipeline with three phases can be designed (Fig. 2.3a): filtering, mapping and quantitative analysis.

Pipeline 2 The organism of interest has not been sequenced before, and the reads are longer (from 400 bp to 800 bp, e.g., reads produced by 454). A pipeline can be designed with three phases (Fig. 2.3b): filtering, assembly and annotation. The assembly phase construct one consensus sequence for each group of reads presenting similar extremities. The annotation phase assigns biological functions to the consensus sequences.

A pipeline is usually implemented using a programming language (e.g., Java or Perl) that controls the execution of the softwares, which use files organized in file directories, or a database management system (e.g., MySQL (MySQL 1995) or PostgreSQL [PosGres]) that stores, retrieves and manages data. Each pipeline phase uses public (open source) or private softwares, and some of the most commonly used public ones are described next.

2.2.2 Bioinformatics Softwares

2.2.2.1 Filtering

As can be seen in Part 4.1, the high-throughput sequencers use different techniques, which may cause specific errors in the reads. These errors have to be treated to guarantee quality to the reads used in the next pipeline phases. Therefore, the *filtering*

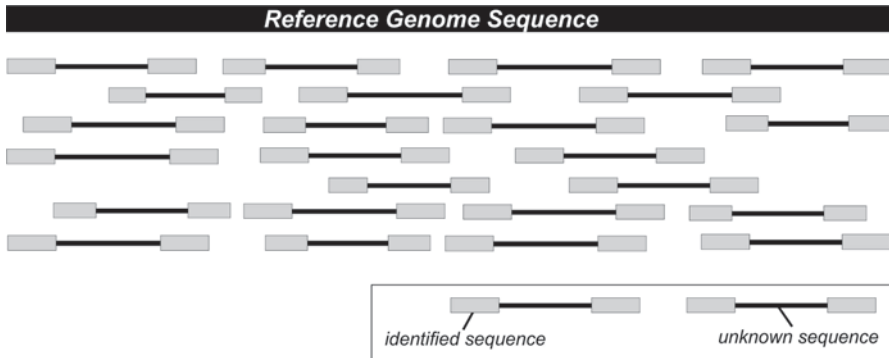


Fig. 2.4 Short reads mapped to a reference genome. (Source: <http://readtiger.com/wkp/en/Genomics>)

(or *cleaning*) phase performs clipping and trimming, as described before. The reads are stored using format FASTQ, which stores the nucleotides of each read together with their corresponding quality scores.

Some tools are used to assess the qualities as well as other information about the input sequences. FastQC (Andrews 2010) allows to verifying quality of raw data. FASTX-Toolkit (Gordon and Hannon 2010) provides options for performing both clipping and trimming. Other commonly used tools are Cutadapt (Martin 2011) for clipping and PRINSEQ (Schmieder and Edwards 2011) for trimming. All of them present several options such as minimum size of one read, minimum quality score and polyadenylation removal.

2.2.2.2 Softwares for Mapping

The objective of the mapping phase is to find where each filtered short read is located in a reference genome (Fig. 2.4).

There are many softwares capable to performing the mapping process. In general, these softwares are computational intensive (to process and store data), and mapping techniques use indices to accelerate the search procedure and to reduce the memory cost associated to finding the location of the short reads to the reference genome.

Bowtie (Langmead et al. 2009) is a fast short aligner that tolerates a small number of mismatches. Bowtie first concatenates all the reference genome in one single string, and performs the Burrows-Wheeler transformation to generate one index to this reference genome. Next, one character of each sequence is mapped until the entire sequence is aligned. If the sequence cannot be aligned, the program backtracks one step, substituting one character, and repeating the process. The maximum number of character substitutions is a parameter in Bowtie.

Transcriptomics in Health and Disease

Passos, G.A. (Ed.)

2014, XVII, 344 p. 44 illus., 34 illus. in color., Hardcover

ISBN: 978-3-319-11984-7