Chapter 7

RNA-seq – Revealing Biological Insights in Bacteria

Mariana P. Santana, Flavia F. Aburjaile, Mariana T.D. Parise, Sandeep Tiwari, Artur Silva, Vasco Azevedo and Anne Cybele Pinto

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/61669

Abstract

New technologies are constantly being released and the improvements therein bring advances not only to transcriptome, the focus of this chapter, but also to diverse areas of biological research. Since the announcement and application of the RNA-seq approach, discoveries are being made in this field, but when we consider bacterial species, this progress proceeded a few years behind. However, with the application of RNA-seq derivative approaches, we can gain biological insights into the bacterial world and aspire to uncover the mysteries involving gene expression, organization and other functional genomic features.

Keywords: RNA-seq, bacteria, transcriptomics, bioinformatics analysis workflow

1. Introduction

RNA-seq technology has driven advances in gene expression analysis through new-generation sequencing platforms, as they are versatile, powerful and ensure quality results with accuracy and reproducibility never reached before. This technology generates information that provides meaning to the set of transcripts (transcriptome), opening up possibilities for understanding cell behavior in different environments. RNA is an important component within the cell, since it plays different roles as a messenger regulatory molecule and carrier; and, it is also essential for the maintenance of housekeeping genes [1].

In 2005, the first new generation of sequencing technology was released and has been evolving rapidly [2]. After starting the process of gene expression analysis in bacteria [3, 4] at a more accessible cost, shorter experimental time and without probes, the technology took off and today overlaps other tools used for this purpose, such as microarray technology, until now extremely useful for this type of analysis.

2. Applications of RNA-seq

Understanding the transcriptome is essential to knowledge of the functional genomics of an organism. The development of next-generation sequencing (NGS) impacts different areas, such as medical and industrial, and has gone through a revolutionary process. Different approaches, among them the RNA-seq technique, have emerged in the fields of microbiology and molecular biology in order to aid in understanding and bring solutions to bacterial domain investigations. In this section, we will detail some applications that are part of our current context.

2.1. The medical field

The applications of these NGS technologies in medicine have allowed expansion in the fields of diagnosis, treatment and prevention, especially concerning bacterial diseases. One of their major applications has been the quantification of expression levels of each transcript under different conditions that simulate the intracellular environment. Such work has been done by Pinto et al. (2014) to understand the host–pathogen relationship [5]. Westermann et al. (2012) demonstrated the validity of this technique, with the transcriptome of the pathogenic bacteria as their host, using the dual RNA-seq that simultaneously analyzed the gene expressions of the pathogen and host [6]. This gives us better understanding of the systems biology involving bacteria and their hosts, helping scientists to develop drugs and vaccines.

Another field that has been explored extensively involves metatranscriptome, as scientists have sought to comprehend the composition and regulation of microbial ecosystems [7, 8]. To pursue this, they have used the RNA-seq technique to generate, and allow the interpretation of, a large volume of very reliable data. Leimena et al. (2013) also validated the RNA-seq technique using the microbiota of a human small intestine with ileostomy. Their aim was to understand the interactions involved in this microbial ecosystem and how these relationships can be associated with disease [8]. Transcriptome analysis pipelines (see Section 5) can be used with different experimental designs and applied to many bacteria in addition to those in the medical field.

2.2. The industrial field

Industrial applications have been developed in recent years, mainly in the probiotic industry, since it benefits the world economy. Bisanz et al. (2014) used the RNA-seq technique [9] to show the metatranscriptome of probiotic yogurt, seeking to understand the metabolic activities that allow the survival of this organism in the products. Their results show the adaptive capacity of this bacterium, as well as the variation in differential gene expression, yielding the taste or storage life of the product [9]. Studies such as these are important because they enrich the knowledge of the industrial field and open new possibilities for an attractive area in the marketplace, which results in improvement in the quality of the product that is ultimately delivered to the consumer.

http://dx.doi.org/10.5772/61669

In addition to the probiotic market, another important area is the bacterial production and synthesis of biomolecules. Wiegand et al. (2013) used the RNA-seq technique to understand the regulatory RNAs in the fermentation of *Bacillus licheniformis*. Their study identified active genomic regions which, in turn, contribute to the efficiency and optimization of the fermentation process, which can promote the industrial production of exoenzymes and antibiotics [10].

Microorganisms produce antioxidant molecules that can be used in the pharmaceutical and cosmetic industries. They also produce other compounds, such as propionate, that are applicable in the production of chemical aids and are produced by *Propionibacterium freudenreichii ssp. shermanii*, which one is considered valuable in the food industry [11]. In this area, the RNA-seq technology is very promising and its application can bring advances in these studies.

3. RNA-seq and derivative techniques

3.1. RNA-seq

The RNA-seq technology is able to identify all RNAs directly and quantitatively: coding and non-coding, rare and abundant, smaller and larger. This method provides information about the transcription start site (TSS), untranslated regions (UTRs), detection of unknown open reading frames (ORFs), improved quality in genomic annotation [12], and also allows the distinction between primary and processed transcripts (dRNA-seq) [13].

The major constraint is to ensure representatives for rare transcripts. In this case, the recommendation is either to increase the representation of reads per library [14] or to enhance these transcripts, eliminating the ribosomal (rRNA) and transfer (tRNA) RNAs that are in abundance in the cells representing about 95% of total RNA [15].

Despite RNA-seq generally being considered the gold standard for gene expression analysis, some researchers nevertheless find it complicated to define this technology as the gold standard. It is a method that is available in different platforms and address different strategies, showing advantages and disadvantages. However, the superiority of this technology, compared to others in the past, is not questioned [16].

Despite the technological superiority, the need for biological replicates and depth of sequencing remains. Hence, the results may achieve greater reliability and reproducibility [17]. Differentially expressed genes are better appraised when there are samples with more biological replicates, as compared to enhanced depth with fewer replicates [18].

Transcriptomics studies have contributed a revolution in the study of the bacterial environment. Different bacterial species have been targeted for RNA-seq studies [5, 13, 19, 20], and gene expression-based discovery has transformed the scientific paradigm of these organisms. The detection of an unexpected amount of coding genes in *Helicobacter pylori* has demonstrated that, despite having a small compact genome, the transcriptome of this bacterium is extremely complex [13].

A surprising result was the detection of a large number of transcription start sites (TSS). This has never been achieved before using any technology aside from derivative RNA-seq technology, like the differential RNA-seq (dRNA-seq), which differentiated primary transcripts that exhibit triphosphate ends from processed transcripts that present monophosphate ends, such as rRNAs and tRNAs. In this case, to enrich mRNA, the strategy was to treat all the RNA samples with exonuclease enzymes that degrade nucleotide monophosphate. This strategy identified 5'UTR ends, operons and antisense transcription, thus providing a new perception of the organization of the bacterial transcriptome and a new model for the analysis of individual genes [13].

The results obtained allow the inference of a role of 5'UTR regions. A correlation between size and cell function was proposed by the researchers, who found that larger size is related to pathogenicity [13]. These results show how little knowledge there is regarding microorganisms, believed to be the simplest form of life, yet which nevertheless prove to be more complex than previously anticipated. This leaves a lot to be discovered.

An RNA-seq application that has been widely used in bacterial genomes is found in studies focused on identifying small RNAs (sRNA). These elements are regulators of various biological processes and were initially studied primarily in *Escherichia coli* [21]. However, with the advances in technology, it has been possible to identify and characterize small RNAs in a variety of bacterial species [13, 22, 23]. Yan et al. (2013) identified an expression profile of sRNA in the *Yersinia pestis*, both *in vitro* and *in vivo*. This has allowed the identification of new sRNAs and the recognition of gene expression modulation during the infection process, thus improving the understanding of the transcription regulation mechanisms of this organism [24]. The importance of studies involving sRNA also includes assistance in research related to antibiotics therapies, a study in initial development despite a lot of knowledge to be better exploited [25].

RNA-seq has been used in different areas and situations. Advanced studies using this technology can detect details in cell expression [26]. Even with the difficulties in separating eukaryotic and prokaryotic materials, it was possible to distinguish the simultaneous expression profiles between the host–pathogen responses through dual transcriptome studies. This work allowed to disclosure the host response against the bacterial infection and virulence factors, enabling the infectious process determination [27]. These studies contribute to the research in the field of biological infection by examining diverse pathogens with different life cycles and methods of infection and providing crucial knowledge for studies of diagnostics and vaccines, such as metatranscriptomics study.

After a relatively short time on the market, RNA-seq can accurately reveal structural and functional elements of bacteria. The mapping of transcripts in the genome can refine the annotation or even identify new regions, improve the quality of the studied genome compared to regions previously annotated by predictors or assembled using an *ab initio* approach [28, 29], and can even check the abundance of transcript expression.

Data coming from a quality genome tends to provide more promising results, responding to the biological question being investigated by researchers. In search of a quality genome, *ab initio* transcripts assembly or even a hybrid approach, which uses both the reference genome

and *ab initio* assembly, become an auspicious endeavour to solve many problems encountered in the genome and complicated to adjust [28].

Pinto et al. (2012) conducted a study of *Corynebacterium pseudotuberculosis* adopting *ab initio* assembly and, therefore, were able to identify differences in the expression of active genes under different environmental conditions. This allowed them to detect new possible virulence factors involved in pathogenicity, making them targets for vaccine development, diagnosis or treatment against caseous lymphadenitis disease caused by this bacterium [30].

These results suggest the importance of this technology and the possibility of going further with a tool that aims to improve, and probably will expand, the field of analysis. This could bring the results increasingly closer to bacterial molecular reality.

3.2. tagRNA-seq

Bacterial RNA can be divided in two groups: primary and processed transcripts. Primary transcripts are represented by the presence of 5′-triphosphate (5′PPP), which includes messenger RNA (mRNA) and small RNAs (sRNA). Processed transcripts are those carrying 5′-monophosphate (5′P), such as mature ribosomal RNA (rRNA) and transfer RNA (tRNA).

Transcriptome represents approximately 95% of the total bacterial transcriptome [15]. A recently developed approach called dRNA-seq [13] revolutionized the study of the primary transcripts by considering the 5′ difference between the primary and the processed groups, as mentioned previously (see Section 3.1).

RNAs are very stable and during preparation, considering the "wet-lab" experiments, some transcripts are partially or totally degraded. 5′PPP and 5′P are two of the mechanisms of protection against exonucleases and the first degraded portion of the transcripts. During that process, information is lost and some primary transcripts end up with 5′P and are treated as processed transcripts. Consequently, they are eliminated by the dRNA-seq technique. A new methodology was created to overcome this problem by tagging and clustering the two groups together in an RNA-seq-derived approach named tagRNA-seq [31]. This technique also considers the difference between processed and primary transcripts, but instead of degrading the processed ones, two different ligation reactions are implemented with two different markers: PSS-tag (processed start site) and TSS-tag (transcription start site). They differ in their nucleotide sequence. Figure 1 exhibits briefly the methodology, considering the three main steps: (1) the first reaction tags (PSS-tag) on the processed transcripts; (2) treatment with tobacco alkaline phosphatase (TAP), where the 5′PPP loses two phosphates, which allows the third step; (3) the second ligation reaction (TSS-tag) on the primary transcripts. After those steps are completed, the transcripts are sequenced and, due to the different markers, they can be distinguished and compared [31].

This methodology was first described for *Enterococcus faecalis* [31] and was based on another technique, 5′tagRACE [32], a 5′RACE derived method. The results provided by tagRNA-seq improved the annotation of the *E. faecalis* genome by having identified or corrected several genome portions, including both non-coding and coding regions. This study also compared different libraries to prove the effectiveness of this innovative approach. With this, it provided

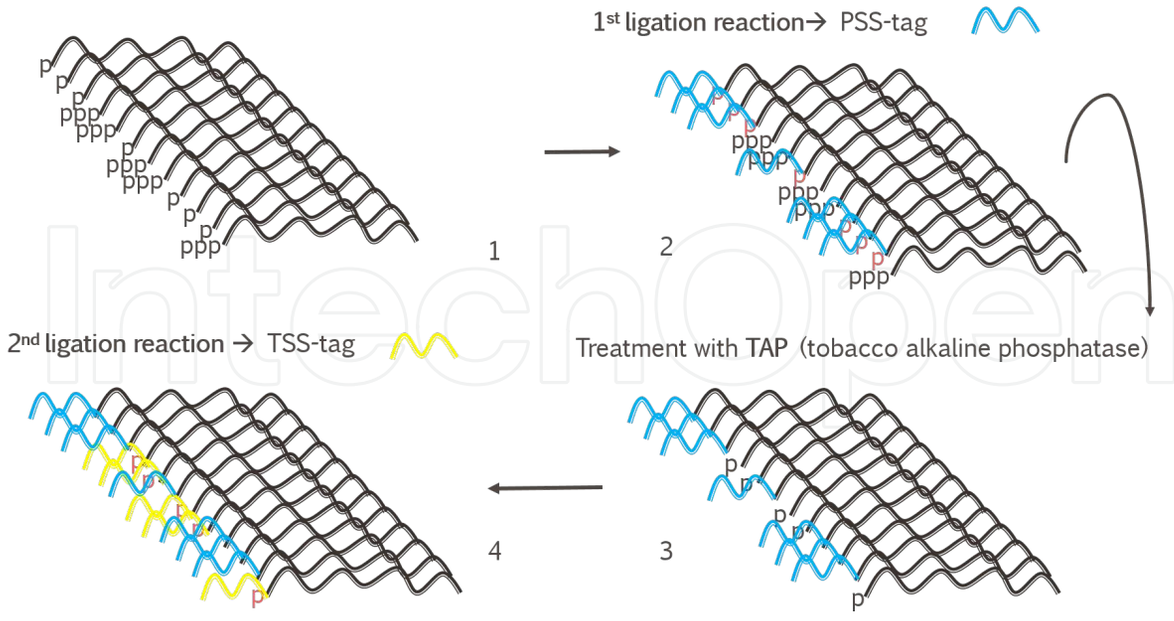

Figure 1. The three main steps of the tagRNA-seq approach. (1) The first ligation reaction, during which the attachment of the PSS-tag (blue) to the processed transcripts (5'P) occurs. (2) Treatment with tobacco alkaline phosphatase (TAP), turning triphosphate to monophosphate groups. (3) The second ligation, corresponding to the TSS-tag (yellow) marker on the previously 5'PPP group (primary transcripts). The different markers allow the differentiation of the triphosphate and monophosphate groups after sequencing.

a new method capable of differentiating primary and processed RNAs and was suited to better comprehending of the genetic information of bacteria as other groups [31].

dRNA-seq and tagRNA-seq are approaches that enable a new view of the transcriptome by selecting the primary transcripts for sequencing or by differentiating the primary from the processed transcripts, for a broader insight into the transcriptome. These state-of-the-art techniques promise a better understanding of RNA structures like TSS, 5'UTR, promoters, among others, besides the knowledge of non-annotated genes and small RNAs.

3.3. FRT-seq (flowcell reverse transcription sequencing)

Flowcell reverse transcription sequencing (FRT-seq) is a new and improved methodology, derived from the RNA-seq technology that was created for Illumina sequencers. Unlike RNA-seq, FRT-seq does not require amplification by PCR, a step that usually introduces bias into the results by displaying an erroneous view of the quantity of some RNA species [33]. Other important features of the Illumina sequencing methodology are the ability to generate strand-specific information, the use of pair-end libraries and the need for a considerable initial amount of RNA template. PCR-free amplification is a major step towards a more comprehensive library, akin to the original one, but without the formation of intermolecular priming artefacts among other errors. It will probably become a fairly useful technique in the near future [33, 34]. Third-generation sequencing platforms, like Nanopore and PacBio, also use amplification-

http://dx.doi.org/10.5772/61669

free approaches. However, neither is currently being broadly used since they still exhibit sequencing errors.

FRT-seq comprises the fragmentation of the template (e.g., mRNA) followed by ligation of adapters in both the 3′ and the 5′ ends, which are responsible for the hybridization of the template with oligonucleotides on the flowcell surface. The next steps performed are quantification, reverse transcription and then sequence reaction [33, 34].

This approach can be applied to both eukaryotes and prokaryotes, although the number of published papers involving eukaryotes is more substantial. From the bacterial world, we can quote papers involving *Salmonella enterica* [23] and *Shigella fleneri* [35] in which FRT-seq was applied as a complementary approach to describe the transcriptional landscape of the species. In both cases, FRT-seq showed greater sensitivity and excellent concordance when compared to other approaches and replicates.

The *S. enterica* paper [23] shows that FRT-seq is as efficient as the RNA-seq and dRNA-seq techniques (Figure 2) (Table 1). Figure 2 compares nine different RNA libraries: TEX (1, 2, 3), RNA-seq (1, 2, 3, *) and FRT-seq (depleted and not depleted). TEX (libraries treated with terminator exonuclease) is a dRNA-seq methodology (see Sections 3.1 and 3.2) that, together with the first three RNA-seq biological replicates, was sequenced using a 454 (1 and 2) or an Illumina GAII (3 and FRT-seq) sequencer and the RNA-seq* (library enriched for small RNA species) was sequenced using Illumina HiSeq. The charts relate the percentages of different RNA species and show that the FRT-seq libraries provide similar or better results than the other approaches. The data presented in Table 2 also support this claim, especially considering both the total number of reads and the uniquely mapped reads achieved using the FRT-seq libraries.

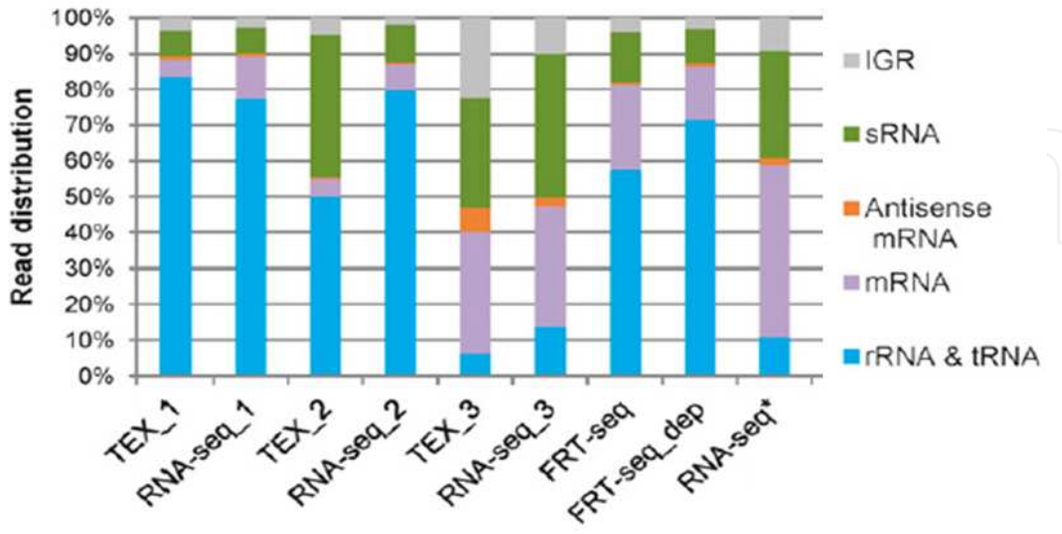

Figure 2. Sequencing methodology comparison. Adapted from [23]. IGR – Intergenic region; TEX – libraries treated with terminator exonuclease; RNA-seq* – library enriched for small RNA species (sRNA).

Library	Sequencing technology	Description	Total number	Number of reads (not mapped)	Number of reads (uniquely mapped)	Percent uniquely mapped reads [%]	Minimum fold coverage[#]
TEX_1	454	dRNA-seq library biological replicate 1	161,031	72,623	88,408	54.90	1.11
RNA-seq_1	454	RNA-seq library biological replicate 1	248,993	83,030	165,963	66.65	2.03
TEX_2	454	dRNA-seq library biological replicate 2	111,462	10,785	100,677	90.32	2.16
RNA-seq_2	454	RNA-seq library biological replicate 2	93,337	38,577	54,760	58.67	0.61
TEX_3	Illumina GAII	dRNA-seq library biological replicate 3	1,738,867	122,058	1,211,426	69.67	20.99
RNA-seq_3	Illumina GAII	RNA-seq library biological replicate 3	2,148,563	136,871	1,360,113	63.30	21.16
RNA-seq*	Illumina HiSeq	RNA-seq library biological replicate 4	3,750,797	164,658	2,596,010	69.21	25.11
FRT-seq	**Illumina GAII**	**FRT-seq library biological replicate 5**	**18,563,218**	**4,203,715**	**2,456,792**	**13.23**	**16.42**
FRT-seq dep	**Illumina GAII**	**FRT-seq library biological replicate 5 rRNA depleted**	**24,585,564**	**9,652,397**	**4,093,744**	**16.65**	**27.77**

Table 1. Sequencing statistics. Adapted from [23]

The *S. fleneri* paper [35] also reports a favourable result concerning FRT-seq. In fact, this approach revealed a larger gene repertoire than the RNA-seq (Table 2).

	RNA-seq		FRT-seq	
	Condition A	**Condition B**	**Condition A**	**Condition B**
Total number of mapped reads	20,099,597	22,736,494	49,925,286	47,605,241
Total number of reads mapping to genes	1,525,782	2,271,423	3,037,954	2,585,600
Reads mapping genes in sense	1,195,446	1,958,533	2,469,828	2,129,951
Reads mapping genes in antisense	330,336	312,890	568,126	455,649

Table 2. Sequencing statistics. Adapted from [31].

The data presented in this topic demonstrate the quality of this recently published methodology and, according to the authors [33, 34], new updates are still being developed. This will probably provide an even better approach for users. The fact that this technique is only applicable for Illumina sequencers is a drawback; but, since this sequencing platform is available worldwide, this disadvantage can easily be fixed. Perhaps, in the near future, it can be extended to work in other sequencing platforms. Another particularity of this technique is its efficiency with AT-rich genomes, which does not constrain its application with AT-poor genomes. This is due to the PCR-free amplification, which raises a question for other sequencers like Nanopore and PacBio. Despite these issues, this technology has a bright future and is a great advance over the conventional RNA-seq.

3.4. Chromatin immunoprecipitation followed by sequencing (ChIP-seq)

Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) is a technique for the genome-wide profiling of DNA-binding proteins, histone modifications or nucleosomes [36]. ChIP-Seq has become an essential tool for studying gene regulation and epigenetic mechanisms. It offers higher resolution, less noise and greater coverage than its array-based predecessor, the ChIP-chip [37, 38]. This approach has six main steps: (1) it is initiated with cell cultures that are grown under defined conditions; and, when the cultures reach the desired stage of development, they are treated with formaldehyde for the cross-linking of proteins and DNA; (2) the chromatin is sheared by sonication into small fragments (200–600 bp); (3) an antibody specific to the protein is used to immunoprecipitate the DNA–protein complex; (4) the cross-links are reversed by heating; (5) the released DNA is subjected to high-throughput sequencing and (6) in silico analysis is carried out in which the resulting sequencing reads are studied for quality and then cropped, based on the quality of the reads [38–40]. The cropped reads are then aligned to a reference genome. Afterwards, areas of enrichment in the ChIP-seq data are identified and those areas, usually called peaks, represent where the transcription factors (TF) bind throughout the genome. CisGenome, MOSAiCs and MACS are some known algorithms that have been utilized in bacterial ChIP-seq analysis [38, 41]. After peaks are associated with genes downstream, a number of bioinformatics analyses can be carried out,

including identification and analysis of motifs, differential analysis and association with expression data for deep understanding of bacterial regulon. This is shown in Figure 3 [36].

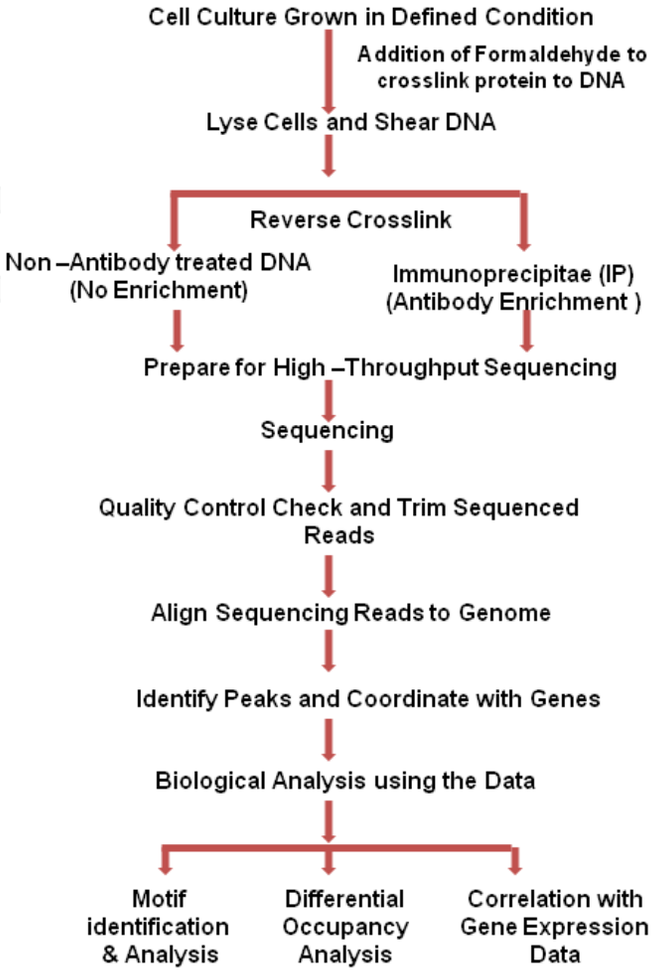

Figure 3. ChIP-seq sample preparation and analysis. Adapted from [36].

As whole-genome transcription profiling cannot reveal whether the influence of the transcription factors (TF) on RNA levels is direct or indirect, this requires identification of transcription factors binding within the appropriate promoter region. ChIP-seq provides information about where the TF are bound. Thus, by integrating ChIP methods and transcription profiling, it is possible to identify all direct regulatory targets of a TF for a given condition. For example, work carried out by Stringer et al. (2014) on the *araC* gene of *Escherichia coli* and *Salmonella enterica* has identified direct regulatory targets of AraC, including five novel target genes: *ytfQ*, *ydeN*, *ydeM*, *ygeA* and *polB* [42]. Although ChIP-seq has been used only in moderation to study bacterial systems in a few bacterial species, such as *Vibrio harveyi*, *V. cholerae*, *Rhodobacter sphaeroides*, *Mycobacterium tuberculosis*, *S. enterica* and *Caulobacter crescentus* [36, 37, 43–45], it is used to identify novel regulatory interactions, even for well-studied proteins [46, 47].

ChIP-seq, in combination with RNA-seq, could be an efficient tool to get detailed information about bacterial transcription regulation and how bacteria respond to different external conditions.

3.5. RNA immunoprecipitation sequencing (RIP-seq)

RNA immunoprecipitation (RIP) is the study of intracellular RNA and protein binding; it is a tool for understanding the dynamic process of post-transcriptional regulatory networks. With this technique, an antibody is used against a protein of interest to recover the RNA species bound to the protein. Since the sequence information of the RNA species bound to a specific protein is often desired, an approach combining RNA immunoprecipitation with sequencing technology (RIP-seq) was created [48]. The main challenge of RIP-seq is the cross-linking step, which is relatively inefficient and only a small amount of RNA is available to construct the library [48, 49]. After that step, treatment with endonuclease elucidates the specific binding sites within the RNA, as they will be protected from digestion. This is followed by purification of the RNA–protein complexes using electrophoresis and high-throughput sequencing [48, 50]. Finally, the data obtained from the sequencer are analyzed using bioinformatics tools. The first study using the RIP-seq-based technique was carried out on *Salmonella* by Sittka et al. (2008) [51]. They used the RNA-binding property of the Hfq protein in their analysis and, as a result, many new sRNA were discovered [52]. Thus, RIP-Seq could be an efficient tool for the identification of bacterial non-coding RNAs.

3.6. LEA-seq (low error amplicon sequencing)

The LEA-seq technique (low error amplicon sequencing) emerged in 2013 and was developed and patented by Gordon and Faith (2014) [53]. This method was created to improve the quality and depth of sequencing runs, since the massive amount of data produced by NGS has caused a high error rate in the sequencing, due to problems with the algorithms or platform reading lengths [53].

LEA-seq is a nucleic acid sequencing technique that identifies events that occur at low frequency, seeking to understand mutation events. The three basic steps for implementing this technique are: (1) linear PCR, (2) exponential PCR and (3) sequencing. This technique is performed based on bacterial 16S sequencing in which PCR carries numerous times and each amplified PCR uses specific primers for each linear molecule [53].

The LEA-seq technique is a quantitative method that has the advantages of generating and reading. This permits the formation of a consensus and the elimination of errors for each molecule. Currently, the available techniques do not support error detection in sequencing or identification of whether there is a real variation in the sequence of that microorganism. The multiple sequencing, using the LEA-seq technique, supports better quality and precision about the organism.

The study by Faith et al. (2013) aimed to identify the composition of the faecal microbiota of adults and to understand the role of these bacterial species and their therapeutic potential for intestinal diseases. This technique allowed them to work with a large number of samples (over 500 isolates), as well as to achieve a fast and accurate analysis of the data [54].

Researchers have a continuing interest in improving this technique, since it can be used for clinical investigation due to its high accuracy: for example, in patients with genetic mutations or somatic mutations. LEA-seq can assist in the search for knowledge about intestinal microbiota, as it may reveal their composition, opening up prospects for the diagnosis, treatment and prevention of gastrointestinal tract diseases.

3.7. CRISPR (clustered regularly interspaced short palindromic repeats)

Ishino et al. (1987) were the first to describe CRISPR [55]. This system has been identified in 40% of bacterial genomes so far [56] and they are defined as short repetitions of grouped bases. The determination of the CRISPR locus and the characterization of adjacent genes, known as *cas* genes, responsible for the function of CRISPR, only occurred in 2002 [57]. The CRISPR/Cas system uses small non-coding RNAs in association with Cas proteins. Cas9 is a nuclease which cleaves DNA in the selected region, so that the CRISPR system/Cas9 can be used to edit genomes.

CRISPR/Cas activity involves three main mechanisms: (1) acquisition, the step in which the DNA fragment is inserted into the CRISPR locus in the genome of interest; (2) transcription, in which the CRISPR locus is transcribed and processed; (3) interference, in which the ejection of nucleic acids occurs. All those mechanisms contribute to bacterial persistence in the environment [58, 59]. Furthermore, CRISPR provides mechanisms to limit the spread of antibiotic resistance or virulence factors. However, Gophna et al. (2015) demonstrated that, even though there are different measurements to evaluate horizontal gene transfer, it is not possible to identify a correlation between the CRISPR/Cas system and the evolution of the species. Changes occur only at the population level [60].

RNA-seq helped in the annotation transcription of regions, mainly non-coding, and also enabled the identification of CRISPR elements in prokaryotes [61]. The CRISPR system can also be used as a tool in studies centered on gene regulation, since this system is able to activate or repress genes.

Zoephel and Randau (2013) discuss how the structure of CRISPR can affect the maturation of RNA and, thus, influence the functionality of the CRISPR/Cas system [62]. The RNA-seq approach was used to evaluate differential gene expression in *S. aureus*, a pathogen of major importance. It was able to identify the CRISPR in these strains and helped in investigating their possible role, since these regions show an adaptive response to infection [63]. Thus, we see the importance of the use of the RNA-seq approach in the magnification of knowledge about function in prokaryotes.

4. RNA Sequencing Platforms

The RNA-seq approach can be applied to different next-generation sequencing platforms and the results obtained by them are proportional to the machine capability. In Table 3, a comparison is made with some of the platforms currently most employed [64].

http://dx.doi.org/10.5772/61669

Company Name	Instrument	Version	Run Time (Hours)	Read Lengths (Mean)	Reads Per Run (Millions)	Applications
Illumina	HiSeq 2000	High Output	132	50	6,000	Gene expression, Splice junction detection, variant calling, fusion
Illumina	HiSeq 2500	High Output	132	50	6,000	Gene expression, Splice junction detection, variant calling, fusion
Illumina	MiSeq	v2 kit	39	250	30	Splice junction detection, variant calling,
Life Technologies	PGM	318 Chip	7.3	176	6	Splice junction detection, variant calling
Life Technologies	Proton	Proton I chip	2-4	81	70	Gene expression, Splice junction detection, variant calling
Pacific Biosciences	RS	RS	0.5-2	1,289	0.03	Splice junction detection, variant calling, full-length gene coverage
Roche	454	GS FLX+	20	686	1	Splice junction detection

Table 3. Different Next Generations sequencing platforms in the study of RNA-seq. Adopted and modified from [64].

5. Bioinformatics Analysis

Experimental investigations in prokaryotes have been facilitated, extended and complemented using computational approaches [65]. Large amounts of data have been generated from RNA-seq experiments which need to be stored and analyzed using computational techniques and tools [66]. This amount has become a bottleneck to bioinformatics analysis and to biologists, since today's transcriptome analysis consists of experiments and data evaluation [65]. Extracting biological information from RNA-seq datasets requires bioinformatics knowledge and tools, making the software choice an important issue for successful RNA-seq analysis [65, 67].

According to Chierico et al. (2015) [68] and Pinto et al. (2011) [67], RNA-seq can be understood as a five-step process: (1) isolation of the total RNA of the organism; (2) mRNA enrichment; (3) synthesis of cDNA; (4) NGS sequencing, which returns raw data to the (5) bioinformatics analysis [67]. A flowchart of this process can be seen in Figure 4.

Figure 4. RNA-seq five-step process.

This session focuses on bioinformatics analysis and the computational tools available. Based on a literature review [29, 65, 67–69], bioinformatics analysis can be comprehended as the extraction and classification/division of biological information gleaned from the sequencing of raw data (Figure 5).

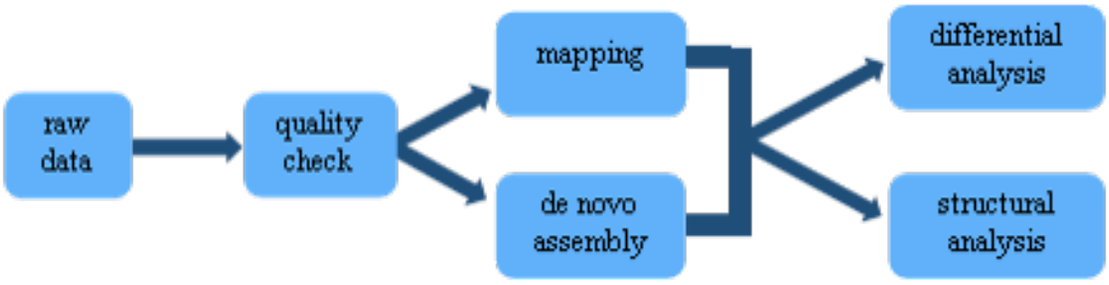

Figure 5. Bioinformatics analysis workflow

5.1. Bioinformatics workflow

The quality check step aims to increase the accuracy of the results by removing sequences that may contain errors [70]; trimming sequences introduced in the library preparation step, such as adapters and poly(A)-tails [71]; and, removing reads with low phred quality. However, in that regard, the use of poor-quality databases can lead to less precise results [72]; considering this, the quality check can affect the next steps drastically.

Some RNA-Seq pipelines, like ReaDemption [71], implement quality checking which performs quality trimming, removes adapters and poly(A) tails and discards reads shorter than a given cut-off (the default cut-off is 12 nucleotides (nt)). Quality assessment [72] evaluates the quality based on quality-graph analysis and estimated coverage. According to Backofen et al. (2014) [65], FastQC (http://www.bioinformatics.babraham.ac.uk/projects/ fastq c/) is a tool commonly used to check read quality and to determine the quality profile of the reads. Software suites can also be used for this purpose, FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) provides tools to remove sequences attached in previous steps and to perform other pre-processing strategies on raw data.

After the quality check, if a reference genome is available, then a mapping step will be done; otherwise, *de novo* assembly. Mapping consists of producing the transcriptome map by aligning reads to a reference genome [67]. This aims to detect the right position of the reads and to distinguish between sequencing errors and genetic variations [73]. Abundant mapping software has been released, differing in their algorithms, memory management, velocity and computational cost [65]. This makes the choice of a mapping tool a challenge. McClure et al.

(2013) [69] made a comparison between SOAP2, BWA, Bowtie and Bowtie2 aligners using 75 RNA-seq experiment data. The comparison of mapping algorithms applied to IonTorrent data can be seen in [73]. After mapping quality is evaluated, ReadXplorer software offers quality classification of read mapping in order to provide information about the quality and quantity of each single read mapping [74]. This approach is recommended when a high-quality genome is available as a reference. If one is unavailable, transcripts should be assembled *de novo* [29].

De novo assembly can be used when investigating poorly studied organisms [14], complex microbial communities or uncultivable organisms [29]. Both DNA and RNA must be assembled, but transcriptome assembly is significantly different than genome assembly [75]; thus, it is important to use RNA assemblers. Tjaden (2015) [29] affirms that assemblers should be specifically designed to prokaryotes, owing to the different challenges of eukaryotic and prokaryotic transcriptomes. Bacterial genomes are often denser than eukaryotic genomes, considering the proximity of the genes. Neighbouring bacterial transcripts can overlap, making it difficult to identify transcript boundaries appropriately. Non-coding eukaryotic RNA models are not appropriate for detecting bacterial small regulatory RNAs [29]. An assembly comparison of three different software titles (Trinity, SOAPdenovo2 and Rockhopper 2), using data from nine different bacteria, can be seen in [29].

When reference mapping or *de novo* assembly is done, data can be analyzed structurally and differentially. The main purpose of differential analysis is to determine the differences in expression among different growth conditions or treatments [76]. Several software titles have been released for this purpose, but there is no consensus about best practices, which makes it difficult to select a tool or method. Seyednasrollah et al. (2013) [76] compared eight differential expression software packages using two real, publicly available datasets. Software that analyzes differential expression can be based on the Poisson method (DEGseq and Myrna), negative binomial method (edgeR and DEseq) or other methods [67, 76]. Pinto et al. (2011) [67] recommends using DEseq or edgeR when analyzing replicates.

Transcriptome annotation and classification can be based on structural analysis, evaluating transcripts regarding the genomic region with which they have been associated and in which they have been classified: protein-coding, non-coding and intergenic regions [65]. Aiming to predict ncRNA transcripts, several computational methods have been developed. Herbig and Nieselt (2011) [77] highlight the SIPHT, sRNAFinder, sRNAscanner, NOCORNAr and sRNAPredict software. NOCORNAr distinguishes itself as it is useful for predicting and characterizing ncRNAs in bacteria [77].

Assessing transcripts concerning genomic regions rely on transcript annotation. The computational approach is convenient to use due to its velocity and precision, compared to manual annotation. However, human supervision of the results is considered important in order to avoid false-positives or missing features [1]. With this technique, some main structures must be detected: 5' transcript ends, 3' transcript ends, TSS and operon [1, 65].

a. Transcript boundaries identification

Annotation of transcript boundaries is important for operon identification and regulatory analyses [1]. Identifying 5' UTR is not always possible; a significant number of transcripts

lacking 5' UTR were found in bacteria and called leaderless transcripts. In this situation, the transcript translation start site and the transcription start site remain in almost the same position [65]. Annotation of 3' UTR is important in order to obtain the entire analytical value of the RNA-seq data. Creecy and Conway (2014) [1] affirm that the current best method for detecting 3' ends is to search for correlations between replicates data. They highlight that the software package TransTermHP can find intrinsic terminators successfully.

b. TSS identification

TSS annotation can assist in ncRNA annotation and polycistronic transcripts [65]. According to Creecy and Conway (2013) [1], it is essential to discover unknown transcripts and to analyze operon, 5' UTR and promoters architecture. Although there are no well-established strategies for TSS identification, owing to scarce knowledge about transcription start sites in bacteria, with computational developments in both computational analyses and "wet-lab" experiments, TSS annotation has become more feasible [65]. TSSAR is a dRNA-seq data-based tool for rapid annotation of TSS that considers dRNA-seq library statistics [78]. According to Backofen et al. (2014) [65], the main advantage is in the statistical analysis presented as an easy-to-use web service. The TSSpredator tool provides automated TSS detection and classification from RNA-seq data, performing a genome-wide comparative prediction of TSS [79]. A comparison among manual annotation, TSSpredator and TSSAR annotation can be seen in [78].

c. Operon identification

The operon represents clusters of co-transcribed genes regulated by the same regulatory sequence and co-transcribed into a single mRNA. This structure has immense biological importance, improving functional gene annotation and giving important information to studies of drug targeting, functional analyses and antibiotic resistance [80]. To handle operon occurrence complexity, the occurrence should be detected using operon architecture (i.e., 5' ends and 3' ends) and have sufficient read coverage to connect promoters and terminators. A strong indication that an operon is real is that at least 90% of the bases of the reads is covered [1]. Chuang et al. (2012) [80] classify computational methods to predict operons and they evaluate 15 algorithms with respect to accuracy, specificity and sensitivity.

5.2. RNA-seq pipeline tools

Not all pipeline tools feature the complete RNA-seq workflow described earlier. To help with tool selection, a software functionalities comparison was developed and is shown in Table 4. To provide additional support, important issues about each software are described, below.

Rockhopper is a system designed specifically for bacterial transcriptome RNA-seq data analysis. A novel approach to mapping transcripts is implemented in this software (similar to the Bowtie2 approach). Mapping normalization is performed followed by transcripts assembly, identification of transcript boundaries, quantification of transcript abundance, testing for differential gene expression and operon prediction. Analysis results are presented using Integrative Genome Viewer, which allows different experiments to be viewed simultaneously [69].

http://dx.doi.org/10.5772/61669

Tool	Quality Check	Mapping	*De novo*assembly	Differential analyses
Rockhopper [69]	-	x	x	x
Rockhopper 2 [29]	-	-	x	x
RNA-Rocket [81]	x	x	-	x
READemption [71]	x	x	-	x
ReadXplorer [74]	x	-	-	x

Table 4. Software comparison.

Rockhopper 2 is a comprehensive system focused on *de novo* assembly that supports differential analysis and transcripts abundance quantification. According to Tjaden (2015) [29], it does not require high-performance computers and can run on personal computers. Rockhopper 2 implements a novel *de novo* assembly algorithm for bacterial transcriptomes. The algorithm works in two stages: (1) candidate transcripts are assembled using a found k-mer and (2) sequencing reads are mapped to candidate transcripts aimed at filtering candidate transcripts to high-quality final transcripts. Concerning differential analysis, Rockhopper 2 first normalizes each RNA-seq dataset, enabling it to compare different experiments or samples [29].

RNA-Rocket aims to simplify the process of aligning RNA-seq data to a reference genome and to generate quantitative transcript profiles. It is built on Galaxy, to provide the tools and services necessary to process RNA-seq data. Some of its benefits are: the possibility of sharing results across research groups; the support of batch analysis for multiple samples; and, the integration of tools and projects, integrating data from the PATRIC platform [81].

READemption pipeline aims to integrate individual RNA-seq analysis tasks and provides a user-friendly tool with a command line interface. This tool was primarily developed to analyze bacterial transcriptome. In order to use the full capacity of modern computers and reduce run time, READemption offers parallel data processing. First, it performs quality trimming of polyA and adapters followed by mapping, coverage calculation, gene expression quantification, differential gene expression analysis and plotting. The software is able to analyze RNA-seq data from Illumina and 454 platforms.

ReadXplorer offers straightforward visualization and analysis functions built around its unique read mapping classification. Analyses such as TSS and operon detection, differential expression, RPKM value and read count calculations are available in ReadXplorer and can be exported to Microsoft Excel files. Read mapping classification sorts read mappings into three different classes: perfect match, best match and common match. These classifications are incorporated in all analyses functions.

5.3. Bioinformatics challenges

Through bibliographic research [29, 66, 69, 71, 82, 83], it has been concluded that bioinformatics has many challenges related to computational issues. RNA-seq experiments generate large amounts of data that must be computationally processed, analyzed, stored and retrieved using a great deal of computational power. In addition to the computational issues, it is important to take into account that not all bioinformatic researchers have extensive computational experience: this makes the lack of user-friendly tools a problem for some users and an important issue for developers. However, great computers, excellent bioinformatic researchers and user-friendly tools do not guarantee successful analysis. The software selected must be appropriate to each biological question and to the organisms studied. Even with all questions presented here, RNA-seq analysis has been very successful in recent years. This success can lead us to imagine the wonderful possibilities for RNA-seq bioinformatic analyses in the future.

Author details

Mariana P. Santana[1], Flavia F. Aburjaile[1], Mariana T.D. Parise[1], Sandeep Tiwari[1], Artur Silva[2], Vasco Azevedo[1*] and Anne Cybele Pinto[1]

*Address all correspondence to: vasco@icb.ufmg.br

1 Instituto de Ciências Biológicas-ICB/UFMG, Departamento de Biologia Geral, Pampulha, Belo Horizonte, Minas Gerais, Brasil

2 Centro de Ciências Biológicas, Departamento de Genética. Universidade Federal do Pará, Campus do Guamá, Guamá. Belém, Pará, Brasil

References

[1] Creecy JP, Conway T. Quantitative bacterial transcriptomics with RNA-seq. Curr Opin Microbiol 2015;23:133–40.

[2] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005 Sep 15;437(7057):376–80.

[3] Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, et al. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi. PLoS Genet 2009 Jul;5(7):e1000569.

[4] Passalacqua KD, Varadarajan A, Ondov BD, Okou DT, Zwick ME, Bergman NH. Structure and complexity of a bacterial transcriptome. J Bacteriol 2009 May 15;191(10):3203–11.

[5] Pinto AC, Sá PHCG de, Ramos RTJ, Barbosa S, Barbosa HPM, Ribeiro AC, et al. Differential transcriptional profile of Corynebacterium pseudotuberculosis in response to abiotic stresses. BMC Genomics 2014 Jan 9;15(1):14.

[6] Westermann AJ, Gorski, SA and Vogel J. Dual RNA-seq of pathogen and host. Nat Rev Microbiol 2012;10:618–30.

[7] Macklaim JM, Fernandes AD, Bella JMD, Hammond J-A, Reid G, Gloor GB. Comparative meta-RNA-seq of the vaginal microbiota and differential expression by Lactobacillus iners in health and dysbiosis. Microbiome 2013 Apr 12;1(1):12.

[8] Leimena MM, Ramiro-Garcia J, Davids M, van den Bogert B, Smidt H, Smid EJ, et al. A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. BMC Genomics 2013;14:530.

[9] Bisanz JE, Macklaim JM, Gloor GB, Reid G. Bacterial metatranscriptome analysis of a probiotic yogurt using an RNA-Seq approach. Int Dairy J 2014;39(2):284–92.

[10] Wiegand S, Dietrich S, Hertel R, Bongaerts J, Evers S, Volland S, et al. RNA-Seq of Bacillus licheniformis: active regulatory RNA features expressed within a productive fermentation. BMC Genomics 2013;14(1):667.

[11] Wang Z, Yang S-T. Propionic acid production in glycerol/glucose co-fermentation by Propionibacterium freudenreichii subsp. shermanii. Bioresour Technol 2013;137:116–23.

[12] Sorek R, Cossart P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nat Rev Genet 2010 Jan;11(1):9–16.

[13] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiß S, Sittka A, et al. The primary transcriptome of the major human pathogen Helicobacter pylori. Nature. 2010 Mar 11;464(7286):250–5.

[14] Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC Genomics 2012 Dec 27;13(1): 734.

[15] Bischler T, Siew Tan H, Nieselt K, Sharma CM. Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in Helicobacter pylori. Methods [Internet]. 2015 Jul 6 [cited 2015 Jul 6]; Available from: http://www.sciencedirect.com/science/article/pii/S1046202315002546

[16] Kratz A, Carninci P. The devil in the details of RNA-seq. Nat Biotechnol 2014 Sep; 32(9):882–4.

[17] Sendler E, Johnson GD, Krawetz SA. Local and global factors affecting RNA sequencing analysis. Anal Biochem 2011;419(2):317–22.

[18] Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics 2014 Feb 1;30(3):301–4.

[19] Isabella VM, Clark VL. Deep sequencing-based analysis of the anaerobic stimulon in Neisseria gonorrhoeae. BMC Genomics 2011 Jan 20;12(1):51.

[20] Patenge N, Pappesch R, Khani A, Kreikemeyer B. Genome-wide analyses of small non-coding RNAs in streptococci. Front Genet 2015;6:189.

[21] Gottesman S, Storz G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. Cold Spring Harb Perspect Biol [Internet]. 2011 Dec [cited 2015 Jul 2];3(12). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3225950/

[22] Papenfort K, Vogel J. Regulatory RNA in bacterial pathogens. Cell Host Microbe 2010 Jul 22;8(1):116–27.

[23] Kröger C, Dillon SC, Cameron ADS, Papenfort K, Sivasankaran SK, Hokamp K, et al. The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. Proc Natl Acad Sci 2012 May 15;109(20):E1277–86.

[24] Yan Y, Su S, Meng X, Ji X, Qu Y, Liu Z, et al. Determination of sRNA expressions by RNA-seq in Yersinia pestis grown in vitro and during infection. PLoS ONE 2013 Sep 11;8(9):e74495.

[25] Storz G, Vogel J, Wassarman KM. Regulation by small RNAs in bacteria: expanding frontiers. Mol Cell 2011 Sep 16;43(6):880–91.

[26] Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet 2014;30(9):418–26.

[27] Humphrys MS, Creasy T, Sun Y, Shetty AC, Chibucos MC, Drabek EF, et al. Simultaneous transcriptional profiling of bacteria and their host cells. PLoS ONE 2013 Dec 4;8(12):e80597.

[28] Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet 2011 Sep 7;12(10):671–82.

[29] Tjaden B. De novo assembly of bacterial transcriptomes from RNA-seq data. Genome Biol 2015;16(1):1.

[30] Pinto AC, Ramos RTJ, Silva WM, Rocha FS, Barbosa S, Miyoshi A, et al. The core stimulon of Corynebacterium pseudotuberculosis strain 1002 identified using ab initio methodologies. Integr Biol 2012;4(7):789.

[31] Innocenti N, Golumbeanu M, d' Hérouël AF, Lacoux C, Bonnin RA, Kennedy SP, et al. Whole genome mapping of 5'RNA ends in bacteria by tagged sequencing: a comprehensive view in Enterococcus faecalis. ArXiv Prepr ArXiv14101925 [Internet]. 2014 [cited 2014 Dec 15]; Available from: http://arxiv.org/abs/1410.1925

[32] Fouquier d'Herouel A, Wessner F, Halpern D, Ly-Vu J, Kennedy SP, Serror P, et al. A simple and efficient method to search for selected primary transcripts: non-coding and antisense RNAs in the human pathogen Enterococcus faecalis. Nucleic Acids Res 2011 Apr 1;39(7):e46–e46.

[33] Mamanova L, Turner DJ. Low-bias, strand-specific transcriptome Illumina sequencing by on-flowcell reverse transcription (FRT-seq). Nat Protoc 2011 Nov;6(11):1736–47.

[34] Mamanova L, Andrews RM, James KD, Sheridan EM, Ellis PD, Langford CF, et al. FRT-seq: amplification-free, strand-specific transcriptome sequencing. Nat Methods 2010 Feb;7(2):130–2.

[35] Vergara-Irigaray M, Fookes MC, Thomson NR, Tang CM. RNA-seq analysis of the influence of anaerobiosis and FNR on Shigella flexneri. BMC Genomics 2014 Jun 6;15(1):438.

[36] Myers KS, Park DM, Beauchene NA, Kiley PJ. Defining bacterial regulons using ChIP-seq methods. Methods [Internet]. 2015 [cited 2015 Jul 17]; Available from: http://www.sciencedirect.com/science/article/pii/S1046202315002285

[37] Myers KS, Yan H, Ong IM, Chung D, Liang K, Tran F, et al. Genome-scale analysis of Escherichia coli FNR reveals complex features of transcription factor binding. PLoS Genet 2013;9(6):e1003565.

[38] Park PJ. ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet 2009;10(10):669–80.

[39] Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science 2007;316(5830):1497–502.

[40] Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 2007;4(8):651–7.

[41] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5(10):R80.

[42] Stringer AM, Currenti S, Bonocora RP, Baranowski C, Petrone BL, Palumbo MJ, et al. Genome-scale analyses of Escherichia coli and Salmonella enterica AraC reveal noncanonical targets and an expanded core regulon. J Bacteriol 2014;196(3):660–71.

[43] Haycocks JRJ, Sharma P, Stringer AM, Wade JT, Grainger DC. The molecular basis for control of ETEC enterotoxin expression in response to environment and host. PLoS Pathog 2015 Jan 8;11(1):e1004605.

[44] Singh SS, Singh N, Bonocora RP, Fitzgerald DM, Wade JT, Grainger DC. Widespread suppression of intragenic transcription initiation by H-NS. Genes Dev 2014 Feb 1;28(3):214–9.

[45] Kahramanoglou C, Seshasayee ASN, Prieto AI, Ibberson D, Schmidt S, Zimmermann J, et al. Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. Nucleic Acids Res 2011 Mar 1;39(6):2073–91.

[46] Wade JT, Struhl K, Busby SJ, Grainger DC. Genomic analysis of protein–DNA interactions in bacteria: insights into transcription and chromosome organization. Mol Microbiol 2007;65(1):21–6.

[47] Grainger DC, Aiba H, Hurd D, Browning DF, Busby SJW. Transcription factor distribution in Escherichia coli: studies with FNR protein. Nucleic Acids Res 2007 Jan 1;35(1):269–78.

[48] Chu Y, Corey DR. RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic Acid Ther 2012;22(4):271–4.

[49] Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 2010;141(1):129–41.

[50] Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. Nature 2009;460(7254):479–86.

[51] Sittka A, Lucchini S, Papenfort K, Sharma CM, Rolle K, Binnewies TT, et al. Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. PLoS Genet 2008;4(8):e1000163.

[52] Cho S, Cho Y, Lee S, Kim J, Yum H, Kim SC, et al. Current challenges in bacterial transcriptomics. Genomics Inform 2013;11(2):76.

[53] Gordon JI, Faith JJ. Methods of low error amplicon sequencing (LEA-Seq) and the use thereof [Internet]. Google Patents; 2014 [cited 2015 Jul 14]. Available from: https://www.google.com/patents/US20140357499

[54] Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, et al. The long-term stability of the human gut microbiota. Science 2013;341(6141):1237439.

[55] Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. J Bacteriol 1987;169(12):5429–33.

[56] Kunin V, Sorek R, Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol 2007;8(4):R61.

[57] Jansen R, Embden J, Gaastra W, Schouls L, others. Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol 2002;43(6):1565–75.

[58] Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci 2012;109(39):E2579–86.

[59] Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science 2012;337(6096):816–21.

[60] Gophna U, Ron EZ. Virulence and the heat shock response. Int J Med Microbiol IJMM 2003 Feb;292(7-8):453–61.

[61] Heidrich N, Dugar G, Vogel J, Sharma CM. Investigating CRISPR RNA biogenesis and function using RNA-seq. CRISPR Methods Protoc 2015;1–21.

[62] Zoephel J, Randau L. RNA-Seq analyses reveal CRISPR RNA processing and regulation patterns. Biochem Soc Trans 2013 Dec;41(6):1459–63.

[63] Osmundson J, Dewell S, Darst SA. RNA-Seq reveals differential gene expression in Staphylococcus aureus with single-nucleotide resolution. PLoS ONE 2013 Oct 7;8(10):e76572.

[64] Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, et al. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nat Biotechnol 2014 Aug 24;32(9):915–25.

[65] Backofen R, Amman F, Costa F, Findei S, Richter AS, Stadler PF. Bioinformatics of prokaryotic RNAs. RNA Biol 2014;11(5):470–83.

[66] McGettigan PA. Transcriptomics in the RNA-seq era. Curr Opin Chem Biol 2013 Feb; 17(1):4–11.

[67] Pinto AC, Melo-Barbosa HP, Miyoshi A, Silva A, Azevedo V. Review application of RNA-seq to reveal the transcript profile in bacteria. Genet Mol Res 2011;10(3):1707–18.

[68] Del Chierico F, Ancora M, Marcacci M, Camma C, Putignani L, Conti S. Choice of next-generation sequencing pipelines. Bacterial Pangenomics [Internet]. Springer; 2015 [cited 2015 Jul 14]. p. 31–47. Available from: http://link.springer.com/protocol/10.1007/978-1-4939-1720-4_3

[69] McClure R, Balasubramanian D, Sun Y, Bobrovskyy M, Sumby P, Genco CA, et al. Computational analysis of bacterial RNA-Seq data. Nucleic Acids Res 2013 Aug 1;41(14):e140–e140.

[70] De Sá PH, Veras AA, Carneiro AR, Pinheiro KC, Pinto AC, Soares SC, et al. The impact of quality filter for RNA-Seq. Gene 2015;563(2):165–71.

[71] Förstner KU, Vogel J, Sharma CM. READemption–A tool for the computational analysis of deep-sequencing-based transcriptome data. Bioinformatics 2014;btu533.

[72] Ramos RT, Carneiro AR, Baumbach J, Azevedo V, Schneider MP, Silva A. Analysis of quality raw data of second generation sequencers with Quality Assessment Software. BMC Res Notes 2011 Apr 18;4(1):130.

[73] Caboche S, Audebert C, Lemoine Y, Hot D. Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. BMC Genomics 2014;15(1):264.

[74] Hilker R, Stadermann KB, Doppmeier D, Kalinowski J, Stoye J, Straube J, et al. ReadXplorer—visualization and analysis of mapped sequences. Bioinformatics 2014;btu205.

[75] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 2013;8(8):1494–512.

[76] Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform [Internet]. 2013 Dec 2 [cited 2014 Apr 30]; Available from: http://bib.oxfordjournals.org/cgi/doi/10.1093/bib/bbt086

[77] Herbig A, Nieselt K. nocoRNAc: characterization of non-coding RNAs in prokaryotes. BMC Bioinformatics 2011;12(1):40.

[78] Amman F, Wolfinger MT, Lorenz R, Hofacker IL, Stadler PF, Findei S. TSSAR: TSS annotation regime for dRNA-seq data. BMC Bioinformatics 2014;15(1):89.

[79] Dugar G, Herbig A, Förstner KU, Heidrich N, Reinhardt R, Nieselt K, et al. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple Campylobacter jejuni isolates. 2013 [cited 2015 Jul 14]; Available from: http://dx.plos.org/10.1371/journal.pgen.1003495

[80] Chuang L-Y, Chang H-W, Tsai J-H, Yang C-H. Features for computational operon prediction in prokaryotes. Brief Funct Genomics 2012;els024.

[81] Warren AS, Aurrecoechea C, Brunk B, Desai P, Emrich S, Giraldo-Calderón GI, et al. RNA-Rocket: an RNA-Seq analysis resource for infectious disease research. Bioinformatics 2015;btv002.

[82] Van Verk MC, Hickman R, Pieterse CM, Van Wees SC. RNA-Seq: revelation of the messengers. Trends Plant Sci 2013;18(4):175–9.

[83] Dai L, Gao X, Guo Y, Xiao J, Zhang Z, others. Bioinformatics clouds for big data manipulation. Biol Direct 2012;7(1):43.