GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Review

Application of RNA-seq to reveal the transcript profile in bacteria

A.C. Pinto[1], H.P. Melo-Barbosa[2], A. Miyoshi[1], A. Silva[2] and V. Azevedo[1]

[1] Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[2] Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, PA, Brasil

Corresponding author: V. Azevedo
E-mail: vasco@icb.ufmg.br

Genet. Mol. Res. 10 (3): 1707-1718 (2011)
Received July 7, 2011
Accepted August 4, 2011
Published August 16, 2011
DOI http://dx.doi.org/10.4238/vol10-3gmr1554

ABSTRACT. The large number of microbial genomes deposited in databanks has opened the door for in-depth studies of organisms, including post-genomics investigations. Thanks to new generation sequencing technology, these studies have made advances that have lead to extraordinary discoveries in bacterial transcriptomics. In this review, we describe bacterial RNA sequencing studies that use these new techniques. We also examined the advantages and biases of these new generation technologies; advances in bioinformatics make it possible to overcome the biases, providing interesting and surprising results.

Key words: RNA-seq; Transcriptome; Bacteria

INTRODUCTION

From one gene to millions

With the progress in sequencing technologies, the number of microbial genomes deposited in databanks has grown in an accelerated manner; to date, 1548 complete bacterial genomes have been included (NCBI, 2011; http://www.ncbi.nlm.nih.gov/genomes). As more genomes are added, it becomes easier to elucidate relevant biological processes. Since sequencing on its own does not give us information about these processes, the next step is investigation based on post-genomics processes, such as transcriptomics. A transcriptome is a collection of all the transcripts (RNAs) present in a given cell, evaluated qualitatively and quantitatively at a particular moment of cell development or during a specific physiological condition (Wang et al., 2009).

During the last decades, the techniques for evaluating gene expression have advanced greatly in the volume of data obtained, which has progressed from one or a few genes, analyzed for example, through Northern blotting, quantitative real-time polymerase chain reaction (RT-PCR), and nuclease protection assay, to analyzing a large number of genes, through subtractive hybridization, differential display, serial analysis of gene expression (SAGE), and microarray (Moody, 2001). However, we are approaching the technical limits of microarray technology; it is being substituted by transcriptomics, using new generation sequencing (NGS) of RNA on an "ultra-large-scale".

Hybridization methodology (microarrays), which is considered to be a large-scale method, is relatively inexpensive, with the exception of high-resolution arrays (tiling arrays); these are used to investigate large genomes. However, this technology has limitations, including a high rate of noise due to cross-hybridization, limitations in the range of detection due to self-generated noise and signal saturation, inability to detect transcripts with a low copy number per cell, detection restricted to transcripts represented in glass slides, and no guarantee of total coverage of the transcripts. Additionally, comparisons of expression data from different experiments are difficult and require a complicated method of normalization based on very complex statistical calculations (Marioni et al., 2008).

Tiling array technology, which investigates the whole genome, requires hundreds of thousands of probes, which substantially increases the cost. Consequently, transcriptome maps produced by tiling array techniques can be low in resolution compared to those developed with RNA-seq. One advantage of tiling arrays is that they do not require enrichment of mRNA, which is required for sequencing (Sorek and Cossart, 2010).

The term RNA-seq has been utilized to represent the transcriptome revealed by sequencing cDNA through NGS; this methodology was developed and initially utilized for identifying the transcriptional map of yeasts (Nagalakshmi et al., 2008). This system provides researchers with a revolutionary method that has high sensitivity and can be used for characterizing the entire transcriptome of an organism. It is useful for discovering new transcripts, identifying mutations, deletions and insertions, and splicing alternatives; it provides excellent coverage and can generate more than 600 million reads in a single run (Anonymous, 2011; http://www.appliedbiosystems.com.br).

In this review, our objective was not to detail the technologies involved, as excellent reviews are available (Hall, 2007; Shendure and Ji, 2008; MacLean et al., 2009), but to show how fast the results of transcriptome analyses can be obtained by sequencing RNA in bacteria. New publications are being added constantly, providing very interesting data concerning the transcriptional profile of these organisms. We also describe the advantages and disadvantages of this technology.

ADVANTAGES AND BIASES OF RNA-seq

Although the pioneering study was done with eukaryotic organisms, because their mRNAs with poly-A tails are easier to isolate, RNA sequencing technology has also been applied to prokaryotes (van Vliet, 2010), mainly in bacteria (Table 1). Although new generation sequencers were introduced relatively recently, RNA-seq has already become quite popular for all areas of genomic research, including transcriptomics, where it has been replacing microarray technology (Teng and Xiao, 2009). Among the advantages of RNA-seq, there is almost no noise, and it allows unequivocal mapping of the sequences in a single region of the genome. It also provides high coverage and has a large range in the detection of transcripts, as it is able to detect from one to numerous copies of RNA per cell. The costs are also considerably lower in comparison with previous methods (Mane et al., 2009). RNA-seq utilizes NGS technology, and thus minimizes errors and simplifies sample preparation by eliminating the cloning step (Morozova and Marra, 2008). RNA-seq technology has been shown to be highly precise in the quantification of transcription levels, giving results similar to those provided by quantitative PCR (qPCR) (Wang et al., 2009). Until now, quantitative RT-PCR has been the reference and is the most precise means to measure expression; according to Roberts et al. (2011), although it is not a perfect assay, it is the best option except for RNA-seq.

Table 1. Representatives of the domain bacteria that have had their transcriptomes studied by RNA-seq to date.

Species	Phylum	Sequencing platform	Reference
Bacillus anthracis	Firmicutes	Illumina SOLiD™	Passalacqua et al., 2009
Burkholderia cenocepacia	Betaproteobacteria	Illumina	Yoder-Himes et al., 2009
Listeria monocytogenes	Firmicutes	Illumina	Oliver et al., 2009
Mycoplasma pneumoniae	Firmicutes	Illumina	Guell et al., 2009
Salmonella typhi	Gammaproteobacteria	Illumina	Perkins et al., 2009
Acinetobacter baumannii	Gammaproteobacteria	Illumina	Camarena et al., 2010
Bacillus anthracis	Firmicutes	SOLiD™	Martin et al., 2010
Chlamydia trachomatis	Verrucomicrobia	Roche GS-FLX	Albrecht et al., 2010
Helicobacter pylori	Epsilonproteobacteria	Roche FLX Illumina	Sharma et al., 2010
Pseudomonas syringae	Gammaproteobacteria	Illumina	Filiatrault et al., 2010
Staphylococcus aureus	Firmicutes	Illumina	Beaume et al., 2011
Neisseria gonorrhoeae	Betaproteobacteria	SOLiD™	Isabela and Clark, 2011
Streptococcus pneumoniae	Firmicutes	Illumina	Croucher et al., 2011

An ideal transcriptome analysis method should be capable of identifying directly and quantitatively all RNAs (coding and non-coding), independent of their being rare or abundant, small or large; new generation sequencers allow such analyses. RNA-seq is the first method that allows quantitative and precise examination of the whole transcriptome in a rapid manner and at a much lower cost than arrays or large-scale Sanger sequencing of expressed sequence tags (ESTs) (Wang et al., 2009).

RNA sequencing data are highly reproducible, with few differences between technical replicates, according to research carried out with data from Illumina, provided that it is sequenced from the same library (Marioni et al., 2008); thus, it is necessary to sequence only once. However, RNA-seq is not free of biases that are inherent to the technology and can introduce problems in the analyses, if not noticed and adjustments made, such as the size of the transcripts, the fragmentation step, synthesis of cDNA, and the mRNA enrichment step. This last step has been removed in many recent studies, because the volume of data obtained from the sequencing of cDNA is sufficient, even though mRNA corresponds to only a small percentage of the total RNA (Siezen et al., 2010). Considering the size of the transcripts, the largest are preferentially detected when compared to the smallest, because the probability of having more representatives (fragments) in the sample is greater. In other words, the total number of reads mapping a particular transcript is directly related to the level of expression, multiplied by the size of the transcript. Therefore, the larger the transcript the more reads are mapped, when compared to smaller transcripts with similar expression levels. The analysis utilized for detection of differential expression between samples is influenced by this bias; therefore, one should be careful in choosing the correct statistical test to resolve this question before initiating the analyses; otherwise, one can overestimate or underestimate a gene. For more detail, the following references can be consulted: Oshlack and Wakefield, 2009; Philippe et al., 2009; Bullard et al., 2010; Roberts et al., 2011.

SEQUENCING STEPS

Most of the processes for obtaining the transcripts follow the steps in the flowchart in Figure 1. Total RNA is isolated and the mRNA enriched, fragmented (or not), and converted to cDNA, which is amplified and sequenced on the NGS platform (Wang et al., 2009). The process generates about a million short or large reads (50 up to 400 bases), depending on the platform chosen (Oshlack et al., 2010).

After sequencing to produce the transcriptome map, the reads are mapped to a reference genome; if there is none, *ab initio* assembly of the transcripts is performed. The expressed regions are determined by the coverage of the reads.

AVAILABLE SOFTWARE AND BIAS OF THE TECHNOLOGY

Despite the advantages offered, sequencing data involve gigabytes of information; it is complex and requires robust analyses. Consequently, analysis of this type of data requires a profound understanding of bioinformatics and an intense computational process, using powerful servers. New methods are constantly being developed to facilitate the interpretation of this type of data (Gonçalves et al., 2011), as summarized in Table 2.

Wang et al. (2009) believed that because of the quality of the data generated by RNA-seq, a complicated normalization, which is necessary for microarray data, would not be needed. However, this step was found to be necessary; although it is not simple; it is of extreme importance to guarantee veracity of the data (Robinson and Oshlack, 2010).

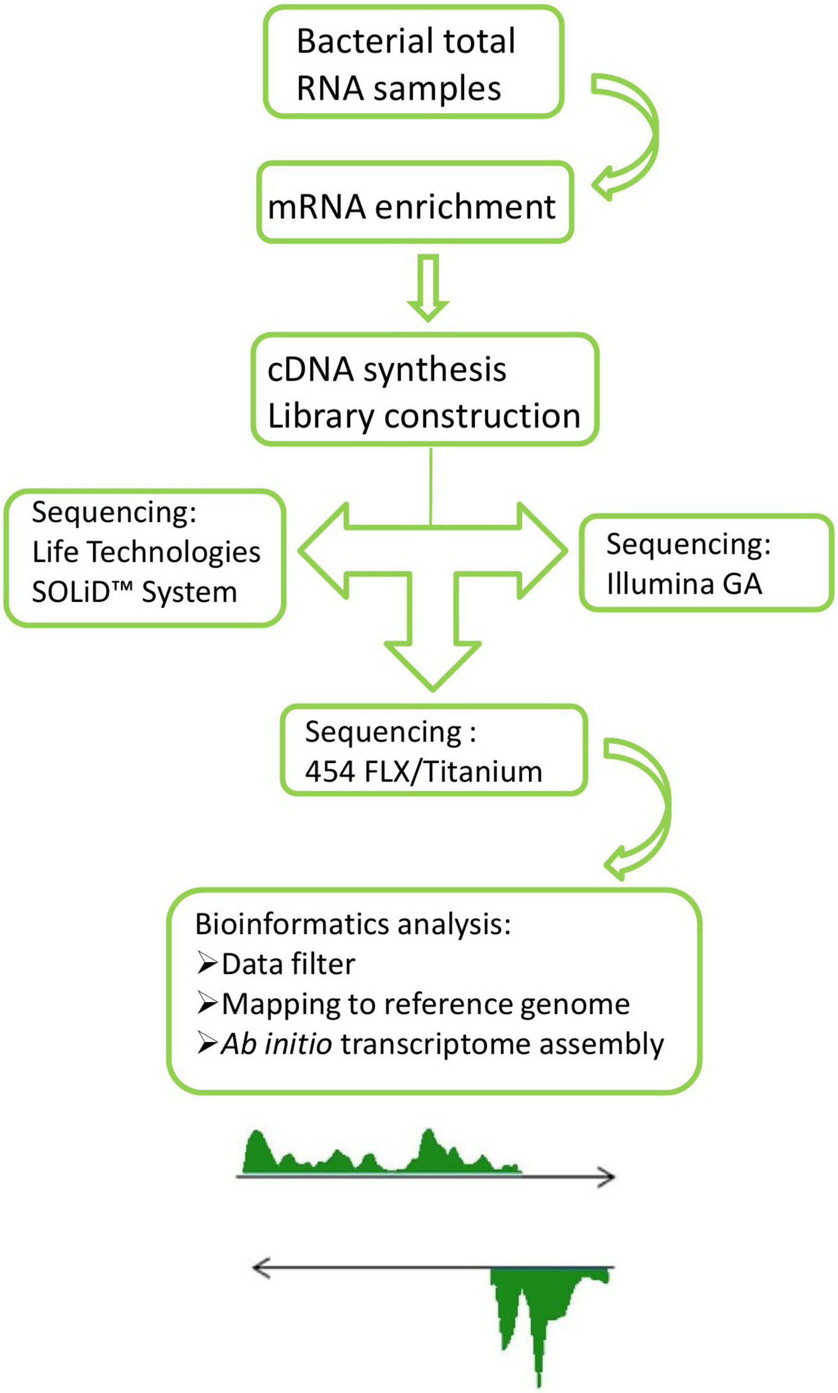

Figure 1. Basic flowchart for obtaining a bacterial transcriptome using the NGS platform.

Table 2. Type of analyses and tools utilized for RNA-seq data.

Analysis	Software	Reference
Genomic mapping	BFAST	Homer et al., 2009
	BOWTIE	Langmead et al., 2009
	CloudBurst	Schatz, 2009
	GNUmap	Clement et al., 2010
	GSNAP	Wu and Nacu, 2010
	MAQ/BWA	Li and Durbin, 2009
	PerM	Chen et al., 2009
	RazerS	Weese et al., 2009
	SOAP/SOAP2 SHRiMP	Li et al., 2008; Li et al., 2009; Rumble et al., 2009
Assembly *ab initio*	Oases	http://www.ebi.ac.uk/~zerbino/oases
	Mira	Chevreux et al., 2004
Normalization	edgeR	Robinson and Oshlack, 2010
	Myrna	Langmead et al., 2010
	Erange	Mortazavi et al., 2008
Differential expression	DEseq	Anders and Huber, 2010
	edgeR	Robinson and Oshlack, 2010
	DEGseq	Wang et al., 2009
	Myrna	Langmead et al., 2010

Programs are available for analysis of differential expression, such as DEGseq and Myrna, which are based on the Poisson method, or edgeR and Deseq, which use the negative binomial method (Gonçalves et al., 2011). When analyzing replicates, DEseq or edgeR is recommended.

TRANSCRIPT PROFILE IN BACTERIA - DISCOVERIES BY SEQUENCING

Studies based on RNA-seq have resolved interesting questions about biological processes in the bacterial cell; they have also contributed to genomics, allowing better quality annotation. For example, mapping the transcript to the reference genome reveals the true start codon of the protein.

An interesting discovery from a transcriptome study was obtained from the genome of *Mycoplasma pneumoniae.* This study contradicted the old concept that bacteria function in a rudimentary manner when compared with complex organisms, such as plants or animals. Guell et al. (2009) utilized two technologies, tiling arrays and direct strand-specific sequencing (DSSS) by means of Illumina, for analyzing bacteria. Using a combination of these techniques, they were able to observe the expression of all the genes. The analysis revealed the versatility of operons in response to different conditions (173 different conditions were tested); that is, one gene coded as polycistronic under one condition, can be transcribed as monocistronic in another. This versatility of the operon was observed in more than 40% of the transcriptions of *M. pneumoniae* and has already been documented in another transcriptome study in Archaea (Koide et al., 2009). These reports reinforce the notion of the operons as non-static structures, which increase the regulatory capacity of bacterial transcriptomes, so that they are functionally analogous to alternative promotors or alternative splicing in eukaryote transcriptomes. The availability of experimental tools for determining operons based on the transcriptome has increased our understanding of the versatile regulation and evolution of polycistronic transcripts in prokaryotes even more.

Although the field of transcriptomics is still in its infancy, it is already clear that it has furnished means for the comprehension of RNA-based regulatory mechanisms. Study of the transcriptome brings us close to reality through the discovery of new genes, defining the structure of the annotated gene, correcting the first methionine, detecting untranslated regions (UTRs), including riboswitches and binding sites of regulatory small RNAs (sRNAs), and by demonstrating the involvement of operons in the physiology, adaptation, and pathogenicity of prokaryotes (Sorek and Cossart, 2010).

Among the discoveries based on NGS technology, sRNAs are particularly outstanding; these studies have shown their role in prokaryote physiology. These elements had been little studied, because of their small size, between 50 and 500 bases, making them difficult to detect. sRNAs regulate important biological processes, such as virulence, response to stress and "quorum sensing" (Bejerano-Sagie and Xavier, 2007; Toledo-Arana et al., 2007). They function in a way that is analogous to miRNAs (microRNAs) in eukaryotes; the abundance of these elements in prokaryotes was reported recently (Sorek and Cossart, 2010).

Cis-antisense transcripts in prokaryotes were believed to be extremely rare. However, they have been found to participate in important regulatory processes, such as replication, response to stress and iron transport (Brantl, 2007). Thanks to the revolutionary analysis of the total transcriptome, utilizing new ultra-large-scale technologies, the abundance of these transcripts in prokaryotes is now known. An RNA transcript from the sense strand can interact with the antisense RNA, providing a common form of regulation for prokaryotes (Sorek and Cossart, 2010).

Utilizing strand specific sequencing of cDNA, through the use of Illumina, reads of 36 bases were mapped in the genome *Salmonella typhi*; this allowed identification of annotation errors of some genes that are actively transcribed regions within prophages. Thus, it was possible to find a longer 5'UTR end in 25 genes. Normally, 5'UTRs in bacteria are less than 30 bp; two of these shorter genes were found in genomic regions with a high concentration of virulence genes, which suggests a role for this UTR in the regulation of virulence (Perkins et al., 2009). This demonstrates that RNA-seq is a powerful method for obtaining useful information.

Although in prokaryotes, attention has been mainly directed towards studies involving 5'UTR or the transcription start site (TSS), it is clear that in the future, transcriptome studies will also reveal the regulatory role of the 3'UTR end (Sorek and Cossart, 2010). A long 3'UTR can affect the expression of genes that are located on the opposite strand (Toledo-Arana et al., 2009).

Oliver et al. (2009) showed how RNA-seq methodology can quantitatively characterize prokaryote transcriptomes with precision, providing new means to explore the regulatory transcriptional network in bacteria. They utilized Illumina technology to characterize the transcriptome of *Listeria monocytogenes* 10403S in the stationary phase and of another strain of this species devoid of factor σ^B, which is involved in response to stress. They found that 83% of the genes are transcribed in the stationary phase and that 42% of the genes annotated in the genome had a moderate or high level of transcription under these conditions. Ninety-six genes had significantly elevated transcription levels compared to the mutant, demonstrating that transcription of these genes is dependent on σ^B. RNA-seq analyses indicated 67 non-coding RNAs, including seven that were previously unknown. The RNA-seq data was also able to improve the annotation of probable operons, as well as visualization of the 5' and 3'UTRs.

In an analysis of the transcriptome of *Pseudomonas syringae* utilizing Illumina, gene expression was validated on a large-scale by precise mapping of 14,951,333 single reads in the genome. This will help identify possible transcription start sites, confirm genes coding for hypothetical proteins, discover new genes that have not been annotated in the genome, and find non-coding RNAs (Filiatrault et al., 2010).

Another important contribution of the RNA-seq strategy, using Illumina and SOLiD™, was the study of Passalacqua et al. (2009). This was the first project to use the SOLiD™ system. These researchers examined the total transcriptome of *Bacillus anthracis* under a variety of growth conditions; they furnished an accurate and high-resolution map of single transcripts and operon structure along the genome. Using short reads of 35 nucleotides, they mapped about 39 million sequences, employing the SOCS program to analyze SOLiD™ data and SOAP for data from Illumina. They identified sequences that had not been annotated in the genome, demonstrating considerable transcriptional activity, including 37 UTR at the 5' end (5'UTR) that were longer than 100 bp. That kind of information can help determine the functional capacity of a specific region, guaranteeing better quality in the annotation of the genome of a species. They also achieved accurate quantification of the transcripts, and demonstrated that there can be significant transcriptional heterogeneity within a clone.

Yoder-Himes et al. (2009), working with two strains of the opportunistic bacterium *Burkholderia cenocepacia*, obtained interesting results when cDNA samples were sequenced using Illumina. Comparing isolates from two different niches (one from soil and the other from a patient with cystic fibrosis), where this bacterium can be found, they observed a large number of regulatory differences between the two strains. They detected 13 ncRNAs apparently involved in habitat adaptation of the strains; these had not been identified in an earlier microarray study by Drevinek et al. (2008), who worked with this same species. The number of reads varied between 1.7 and 4.5 million per sample; based on the results, they came to the conclusion that RNA-seq is a powerful tool for qualitatively and quantitatively examining the transcriptome of bacteria.

An innovation in RNA-seq was achieved with the pathogen *Helicobacter pylori* (Sharma et al., 2010). Two libraries were constructed; one was treated with exonuclease, which eliminates molecules with 5'monophosphates (rRNA and tRNA), and is therefore enriched with primary transcripts that show the 5'triphosphate end. The other library included total RNA without treatment with exonuclease. They then produced a resolution map of nucleotides of this human pathogen by sequencing with Roche FLX. Approximately 217 million reads of the cDNA were mapped in the genome. To help with the prediction of operons, samples of cDNA obtained from randomly fragmented RNAs were also sequenced using Illumina. A large number of previously unknown TSS were found. These investigators observed that the 5'UTR, extending from the TSS to the start codon, generally ranged from 20 to 40 nucleotides. A correlation was found between the size of the 5'UTR and cellular function. For example, larger size was related to pathogenicity. Other recent discoveries assured a new perspective of the organization of transcriptome of *H. pylori* and provided data for improved analysis of individual genes. The transcriptome of this bacterium was found be highly complex, even though the genome is small and compact.

Study of the transcripts of the Gram-negative bacterium *Acinetobacter baumannii* brought new insight into bacterial molecular genetics (Camarena et al., 2010). The effect of ethanol on the transcriptional profile was tested by sequencing RNA using Illumina; 3,596,474 reads were obtained by mapping only in the genome, which allowed detection of the expression of 49 induced genes belonging to different functional categories, and 21 repressed genes. Ethanol was found to be efficiently assimilated and influenced the virulence and growth of the bacteria. These results demonstrated that RNA-seq is a powerful tool for identifying the main modulators involved in bacterial pathogenicity. Even though the samples were contaminated with a large amount of rRNA, despite being filtered, the reads that align guarantee the depth of the sequencing and the quality of the results.

Isabella and Clark (2011) detected a large number of genes that were differentially expressed in the genome of *Neisseria gonorrhoeae*, when grown in an anaerobic environment, a condition that is uncommon for this organism. The sensitivity of the technology of RNA sequencing using the SOLiD^{TM} system gave an average of 1,490,403 reads in four experiments. These data revealed transcriptional regulators involved in anaerobic growth, demonstrating expression and regulation of sRNAs and many genes involved in adaptation and response to anaerobic stress. Many hypothetical proteins were induced under this condition; they could be targets of future studies. This study also identified induced genes, which were expressed from plasmid pJD1 of *N. gonorrhoeae*.

Another more recent investigation, also concerning the bacterium *B. anthracis* (Martin et al., 2010), investigated the transcriptional profile of this pathogen under various stress conditions, including cold shock, osmotic shock with 0.75 M NaCl and shock with 6% ethanol. The authors sought to establish limits for identifying regions as transcribed or not, as well as determine their levels of expression, using data obtained from the different cell growth conditions. Reads obtained by RNA-seq on the SOLiD^{TM} sequencing platform, mapped to the reference genome, were utilized as input for an algorithm based on the Markov chain (HMM), where the outputs of the algorithm were the predicted positions of TSS and transcription termination sites (TES), as well as the level of expression of each predicted transcript. The results extended existing gene annotation, showing evidence of transcription for the majority of the regions already annotated as active genes, identifying hundreds of genes not identified by *ab initio* prediction, and many pseudogenes that were possibly transcriptionally active. Additionally, previously unknown correlations with levels of expression were found, including the average translation speed (ATS), codon adaptation index (CAI), and ribosome binding site (RBS) correlated with genes downstream from those that would be the first genes of an operon.

The total transcriptome inventory of the ubiquitous bacterium *Staphylococcus aureus* N315 was revealed by RNA-seq using Illumina (Beaume et al., 2011). Samples of four points on the bacterial growth curve were taken, for the purpose of defining the temporal expression of each transcript. Total RNAs were extracted with two different methods, RNeasy (Qiagen) and MirVana (Ambion); in both, the RNA was later treated with the MICROBExpress kit (Ambion) to reduce the percentage of rRNAs. Two RNA fragmentation methods were also tested, but no significant differences in results were observed with any of these method variations. RNA sequencing allowed identification of all the transcripts expressed at each different point, revealing a complete pattern of sRNAs for specific conditions. Besides the transcriptional units already annotated in the genome of the strain in question, 195 more small transcribed regions were identified in its chromosome and plasmid, namely 160 sRNAs and 35 antisense RNAs. As a whole, it was discovered that about 10% of the transcripts expressed in this organism were previously uninvestigated, which allows us now to assess their roles in the virulence and/or metabolism of *S. aureus*.

The transcriptomes of the elementary bodies (EB) and the reticulate bodies (RB) of

Chlamydia trachomatis, a pathogenic obligate intracellular bacterium, were sequenced using Roche-FLX (Albrecht et al., 2010). The RNA-seq approach revealed 363 TSS of annotated genes and identified various non-coding RNAs; 42 were from the chromosome and one was from the bacterial plasmid. There was differential expression of 84 genes between EB and RB; these can be directly implicated in the development of the life cycles of this organism.

Analysis of the genome of *Streptococcus pneumoniae* ATCC 700669 showed three families of repeats, conserved or not with other species of this genus. This finding, together with the localization of these sequences in the bacterial chromosome, led us to conclude that these sequences may not have an essential role in the organism but could be parasitic elements. Meanwhile, analysis of the transcriptome of this organism demonstrated the functional importance of these repetitive regions, because it was revealed by RNA-seq utilizing Illumina and RT-PCR that many of these regions were expressed, while others were shown forming structures of riboswitches (Croucher et al., 2011).

CONCLUSIONS

Even today, after many years of studies of bacteria, new discoveries continue to surprise us. Through RNA-seq, it can be seen how the microbial transcriptome is more complex than initially thought and how it approximates that of eukaryotes in various aspects. Perhaps it is because of this transcriptomic versatility that bacteria are able to adapt to diverse environments with such agility.

RNA-seq technology was developed for research for the purpose of guaranteeing results that are improved, closer to reality and at a lower cost, compared to previously employed technologies. Based on the results currently obtained through RNA sequencing, and on the reliability of the data that is guaranteed by the coverage and accuracy of the reads, it is clear that this technology will increasingly lead to extraordinary discoveries.

ACKNOWLEDGMENTS

The authors acknowledge support from the development agencies CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), and FAPESPA (Fundação de Amparo à Pesquisa do Estado do Pará).

REFERENCES

Albrecht M, Sharma CM, Reinhardt R, Vogel J, et al. (2010). Deep sequencing-based discovery of the *Chlamydia trachomatis* transcriptome. *Nucleic Acids Res.* 38: 868-877.

Anders S and Huber W (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11: R106.

Anonymous (2011). Applied Biosystems by Life Technologies. Available at [http://www.appliedbiosystems.com.br]. Accessed May 26, 2011.

Beaume M, Hernandez D, Docquier M, Delucinge-Vivier C, et al. (2011). Orientation and expression of methicillin-resistant *Staphylococcus aureus* small RNAs by direct multiplexed measurements using the nCounter of NanoString technology. *J. Microbiol. Methods* 84: 327-334.

Bejerano-Sagie M and Xavier KB (2007). The role of small RNAs in quorum sensing. *Curr. Opin. Microbiol.* 10: 189-198.

Bentley SD (2011). Identification, variation and transcription of pneumococcal repeat sequences. *BMC Genom.* 12: 120.

Brantl S (2007). Regulatory mechanisms employed by cis-encoded antisense RNAs. *Curr. Opin. Microbiol.* 10: 102-109.

Bullard JH, Purdom E, Hansen KD and Dudoit S (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11: 94.

Camarena L, Bruno V, Euskirchen G, Poggio S, et al. (2010). Molecular mechanisms of ethanol-induced pathogenesis revealed by RNA-sequencing. *PLoS Pathog.* 6: e1000834.

Chen Y, Souaiaia T and Chen T (2009). PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. *Bioinformatics* 25: 2514-2521.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, et al. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* 14: 1147-1159.

Clement NL, Snell Q, Clement MJ, Hollenhorst PC, et al. (2010). The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics* 26: 38-45.

Croucher NJ, Vernikos GS, Parkhill J and Bentley SD (2011). Identification, variation and transcription of pneumococcal repeat sequences. *BMC Genom.* 12: 120.

Drevinek P, Holden MT, Ge Z, Jones AM, et al. (2008). Gene expression changes linked to antimicrobial resistance, oxidative stress, iron depletion and retained motility are observed when *Burkholderia cenocepacia* grows in cystic fibrosis sputum. *BMC Infect. Dis.* 8: 121.

Filiatrault MJ, Stodghill PV, Bronstein PA, Moll S, et al. (2010). Transcriptome analysis of *Pseudomonas syringae* identifies new genes, noncoding RNAs, and antisense activity. *J. Bacteriol.* 192: 2359-2372.

Goncalves A, Tikhonov A, Brazma A and Kapushesky M (2011). A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics* 27: 867-869.

Guell M, van Noort V, Yus E, Chen WH, et al. (2009). Transcriptome complexity in a genome-reduced bacterium. *Science* 326: 1268-1271.

Hall N (2007). Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.* 210: 1518-1525.

Homer N, Merriman B and Nelson SF (2009). BFAST: an alignment tool for large scale genome resequencing. *PLoS One* 4: e7767.

Isabella VM and Clark VL (2011). Deep sequencing-based analysis of the anaerobic stimulon in *Neisseria gonorrhoeae*. *BMC Genom.* 12: 51.

Li H and Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.

Li R, Li Y, Kristiansen K and Wang J (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics* 24: 713-714.

Li R, Yu C, Li Y, Lam TW, et al. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25: 1966-1967.

Koide T, Reiss DJ, Bare JC, Pang WL, et al. (2009). Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol. Syst. Biol.* 5: 285.

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

Langmead B, Hansen KD and Leek JT (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol.* 11: R83.

MacLean D, Jones JD and Studholme DJ (2009). Application of ‘next-generation’ sequencing technologies to microbial genetics. *Nat. Rev. Microbiol.* 7: 287-296.

Mane SP, Evans C, Cooper KL, Crasta OR, et al. (2009). Transcriptome sequencing of the Microarray Quality Control (MAQC) RNA reference samples using next generation sequencing. *BMC Genom.* 10: 264.

Marioni JC, Mason CE, Mane SM, Stephens M, et al. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18: 1509-1517.

Martin J, Zhu W, Passalacqua KD, Bergman N, et al. (2010). *Bacillus anthracis* genome organization in light of whole transcriptome sequencing. *BMC Bioinformatics* 11 (Suppl 3): S10.

Moody DE (2001). Genomics techniques: an overview of methods for the study of gene expression. *J. Anim. Sci.* 79 (Suppl E): E128-E135.

Morozova O and Marra MA (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92: 255-264.

Mortazavi A, Williams BA, McCue K, Schaeffer L, et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5: 621-628.

Nagalakshmi U, Wang Z, Waern K, Shou C, et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320: 1344-1349.

NCBI (2011). National Center for Biotechnology Information. Available at [http://www.ncbi.nlm.nih.gov/genomes]. Accessed June 30, 2011.

Oliver HF, Orsi RH, Ponnala L, Keich U, et al. (2009). Deep RNA sequencing of L. monocytogenes reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs. *BMC Genom.* 10: 641.

Oshlack A and Wakefield MJ (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct.* 4: 14.

Oshlack A, Robinson MD and Young MD (2010). From RNA-seq reads to differential expression results. *Genome Biol.* 11: 220.

Passalacqua KD, Varadarajan A, Ondov BD, Okou DT, et al. (2009). Structure and complexity of a bacterial transcriptome. *J. Bacteriol.* 191: 3203-3211.

Perkins TT, Kingsley RA, Fookes MC, Gardner PP, et al. (2009). A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS Genet.* 5: e1000569.

Philippe N, Boureux A, Brehelin L, Tarhio J, et al. (2009). Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity. *Nucleic Acids Res.* 37: e104.

Roberts A, Trapnell C, Donaghey J, Rinn JL, et al. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* 12: R22.

Robinson MD and Oshlack A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 11: R25.

Rumble SM, Lacroute P, Dalca AV, Fiume M, et al. (2009). SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.* 5: e1000386.

Schatz MC (2009). CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics* 25: 1363-1369.

Sharma CM, Hoffmann S, Darfeuille F, Reignier J, et al. (2010). The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* 464: 250-255.

Shendure J and Ji H (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* 26: 1135-1145.

Siezen RJ, Wilson G and Todt T (2010). Prokaryotic whole-transcriptome analysis: deep sequencing and tiling arrays. *Microb. Biotechnol.* 3: 125-130.

Sorek R and Cossart P (2010). Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat. Rev. Genet.* 11: 9-16.

Teng X and Xiao H (2009). Perspectives of DNA microarray and next-generation DNA sequencing technologies. *Sci. China C Life Sci.* 52: 7-16.

Toledo-Arana A, Repoila F and Cossart P (2007). Small noncoding RNAs controlling pathogenesis. *Curr. Opin. Microbiol.* 10: 182-188.

Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, et al. (2009). The *Listeria* transcriptional landscape from saprophytism to virulence. *Nature* 459: 950-956.

van Vliet AH (2010). Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS Microbiol. Lett.* 302: 1-7.

Wang Z, Gerstein M and Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10: 57-63.

Weese D, Emde AK, Rausch T, Döring A, et al. (2009). RazerS - fast read mapping with sensitivity control. *Genome Res.* 19: 1646-1654.

Wu TD and Nacu S (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26: 873-881.

Yoder-Himes DR, Chain PS, Zhu Y, Wurtzel O, et al. (2009). Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 106: 3976-3981.