MINIREVIEW

Next generation sequencing of microbial transcriptomes: challenges and opportunities
Arnoud H.M. van Vliet
Institute of Food Research, Norwich, UK

Correspondence: Arnoud H.M. van Vliet, Institute of Food Research, Colney Lane, Norwich NR4 7UA, UK. Tel.: +44 1603 255250; fax: +44 1603 255288; e-mail: arnoud.vanvliet@bbsrc.ac.uk

Received 10 June 2009; accepted 15 August 2009. Final version published online 4 September 2009.
DOI:10.1111/j.1574-6968.2009.01767.x

Editor: Ian Henderson

Keywords
microbial transcriptomes; high-throughput sequencing; cDNA libraries; RNA-seq.

Abstract
Over the past 15 years, microbial functional genomics has been made possible by the combined power of genome sequencing and microarray technology. However, we are now approaching the technical limits of microarray technology, and microarrays are now being superseded by transcriptomics based on high-throughput (next generation) DNA-sequencing technologies. The term RNA-seq has been coined to represent transcriptomics by next-generation sequencing. Although pioneered on eukaryotic organisms due to the relative ease of working with eukaryotic mRNA, the RNA-seq technology is now being ported to microbial systems. This review will discuss the opportunities of RNA-seq transcriptome sequencing for microorganisms, and also aims to identify challenges and pitfalls of the use of this new technology in microorganisms.

Introduction
Since the dawn of molecular biology, researchers have always had a particular interest in understanding the mechanics and control of the process of transcription in cells (Seshasayee et al., 2006). Changing levels of transcription is one of the primary mechanisms initiating adaptive processes in a cell, as, via the coupled process of translation, it can lead to production of new proteins, changes in membrane composition and all kinds of other changes in the cellular machinery. The challenge has always been to get as much information as possible about the ‘transcriptome’, which represents the complete collection of transcribed sequences in a cell. This is usually a combination of coding RNA (mRNA) and noncoding RNA (rRNA, tRNA, structural RNA, regulatory RNA and other RNA species). Within these classes of RNA species, it is also of importance to separate de novo synthesized RNA (primary transcripts) and post- transcriptionally modified (secondary) transcripts.

Microarrays: opportunities with limitations
The advent of functional genomics with its availability of the different ‘omics’ technologies has revolutionized our understanding of the process of transcription, as it couples the power of complete genome sequencing with the miniaturization of cDNA and oligonucleotide arrays (jointly known as microarrays), allowing the generation of information about the total cellular responses (Hinton et al., 2004). Annotated genome sequences have been used to construct microarrays representing the majority or all of the predicted genes in a genome, and conversion of RNA into labelled cDNA used for hybridization has allowed the high-throughput detection of relative transcript levels, by either competitive hybridization comparing two RNA samples directly, or by cohybridization to genomic DNA as a common standard for normalization (Hinton et al., 2004). The explosive growth of publications using microarrays prompted the development of the MIAME guidelines (Brazma et al., 2001) to ensure minimal standards for microarray data, and subsequent technological advances in array production allowed for more sophisticated techniques like ChIP-on-chip technologies for the genome-wide detection of binding sites of DNA-binding proteins (Wade et al., 2007). Because of the advances in the technologies, high-density oligonucleotide arrays have become widely available and the subsequent drop in cost has made them applicable in many laboratories worldwide. Recently, high-density tiling arrays...
with short oligonucleotide overlaps have already allowed a more detailed study of transcription in bacteria like Escherichia coli, Caulobacter crescentus, Listeria monocytogenes and Bacillus subtilis (Selinger et al., 2000; McGrath et al., 2007; Rasmussen et al., 2009; Toledo-Arana et al., 2009), and we now know that the microbial transcriptome is much more complicated than previously thought, and includes long antisense RNAs and many more noncoding RNAs than identified previously (Rasmussen et al., 2009; Toledo-Arana et al., 2009).

While microarrays have been instrumental in our understanding of transcription, we have started to reach limitations in their applicability (Bloom et al., 2009). Microarray technology (like other hybridization techniques) has a relatively limited dynamic range for the detection of transcript levels due to background, saturation and spot density and quality. Microarrays need to include sequences covering multiple strains, as mismatches can significantly affect hybridization efficiency and hence oligonucleotide probes designed for a single strain may not be optimal for other strains. This may lead to a high background due to nonspecific or cross-hybridization. In addition, comparison of transcription levels between experiments is challenging and usually requires complex normalization methods (Hinton et al., 2004). Hybridization technologies such as microarrays measure a response in terms of a position on a spectrum, whereas cDNA sequencing scores in number of hits for each transcript, which is a census-based method. The census-based method used in sequencing has major advantages in terms of quantitation and the dynamic range achievable, although it also raises complex statistical issues in data analysis (Jiang & Wong, 2009; Oshlack & Wakefield, 2009). Finally, microarray technology only measures the relative level of RNA, but does not allow distinction between de novo synthesized transcripts and modified transcripts, nor does it allow accurate determination of the promoter used in the case of de novo transcription. Many of these issues can be resolved by using high-throughput sequencing of cDNA libraries (Hoen et al., 2008), and jointly tiling microarrays and cDNA sequencing can be expected to lead to a rapid increase in data on full microbial transcriptomes, as outlined in this article.

**Next-generation (NextGen) high-throughput sequencing technologies**

This review is not meant as an in-depth discussion of sequencing technologies, as there are several excellent recent reviews available (Hall, 2007; Shendure & Ji, 2008; MacLean et al., 2009). It is, however, important to discuss the consequences of the selection of a specific NextGen sequencing technology for the purpose of transcriptome determination. All three commercially available technologies (Roche 454, Illumina and ABI SOLiD) have their pros and cons, and in many cases, access or local facilities will influence the final choice of sequencing technology. All the discussed NextGen sequencing technologies allow for the determination of paired end sequences, and hence can potentially be used for paired end tag (PET) sequencing applications (Fullwood et al., 2009). However, these applications are commonly used in eukaryotic systems for identification of exon domains, and have not been ported to microbial systems. There is currently no direct need for PET applications in microbial transcriptome sequencing.

The **Roche 454 sequencing technology** is based on pyrosequencing in microreactors on a picotiter plate (Margulies et al., 2005), and its strongest features are the generation of long sequence reads and the relative speed of the sequencing run (measured in hours). Its disadvantages lie in the smaller amount of data generated (approximately 0.25–1 Gbp sequence information per plate using the 454 GS FLX and Titanium systems) and hence the relatively high cost, and its difficulty in handling homopolymeric DNA sequences. The **Illumina GA technology** is based on adapter ligation, followed by anchoring to a prepared substrate, followed by local in situ PCR amplification and sequencing using fluorophore-labelled chain terminators (Bennett et al., 2005). Sequences obtained by Illumina sequencing are usually 35–75-nt long, but advances in the technology are expected to result in longer readlengths (up to 125 nt) soon. Advantages of the Illumina technique are the large amounts generated (5–10 Gbp total per run), its sequencing accuracy and the relatively low price per Gbp. However, runtimes are measured in days, and increasing the readlength will increase runtimes significantly, and the images require very large storage space. Because shorter reads may be more difficult to accurately map on genomes (especially those with repeated sequences), operators will have to select the right balance between read length and running time/cost. Finally, the **ABI SOLiD technology** uses amplified DNA on beads, which are bound to glass slides. The amplified DNA is sequentially hybridized with short defined oligonucleotides, which contain known 3‘ dinucleotides and a specific 5‘ fluorophore. The oligonucleotide complementary to the template at its 3‘ dinucleotide is ligated to the 5‘ end of the 5‘-elongating complementary strand, and after fluorophore identification, the 5‘ remainder of the oligonucleotide is cleaved to prepare for the next cycle of oligonucleotide annealing and ligation. Repeated cycles of DNA synthesis and melting allow for colour-recognition of each base in the DNA sequence (Shendure et al., 2005). The SOLiD technology generates reads of 35–50 nucleotides. The advantages are the high fidelity of the sequences obtained, which makes the technology excellently suited for SNP analysis, and the generation of large datasets (6–15 Gbp total per run). The disadvantages are similar to those of Illumina sequencing.
It needs to be noted that the RNA-seq technology needs the availability of a reference genome sequence, similar to microarray technology. If the genome sequence of the specific strain is not available, it may be possible to utilize a reference sequence from another strain in the same species, although this will invariably result in the loss of sequence information and incomplete representation of the genome in the RNA-seq output. Overall, all three technologies can be used for genome and transcriptome sequencing. Other applications aimed at RNA-seq of single cells (Tang et al., 2009) are eagerly awaited, but not yet described for bacteria and are not commercially available.

**Use of NextGen sequencing for analysis of microbial transcriptomes**

As indicated previously, high-throughput sequencing of cDNA libraries has the potential to study transcription at the single nucleotide level and hence yield much more detail on RNA transcripts present in a population of microbial cells. However, when compared with eukaryotic RNA, working with bacterial RNA has always been a challenge. Unlike eukaryotic mRNA, most bacterial mRNAs do not have a poly-A tail (Deutscher, 2003), and hence cannot be isolated from other RNA sources by hybridization to immobilized poly-T. Furthermore, bacterial RNA preparations usually contain up to 80% rRNA and tRNA (Condon, 2007), and to add insult to injury, bacterial mRNA often has a very short half-life and hence can be highly unstable (Deutscher, 2003; Condon, 2007). Hence, it is not surprising that high-throughput sequencing of the transcriptome of a cell (RNA-seq or mRNA-seq) was first described for eukaryotic cells, including the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* (Nagalakshmi et al., 2008; Wilhelm et al., 2008), mouse organs and embryonic stem cells (Cloonan et al., 2008; Mortazavi et al., 2008), human cell lines (Sultan et al., 2008) and the plant *Arabidopsis thaliana* (Lister et al., 2008). In these studies, transcriptome sequencing was highly informative, and allowed for investigation of levels of transcripts as well as (alternative) splicing events. More information on RNA-seq in eukaryotic organisms can be found in recent reviews (Wang et al., 2009; Wilhelm & Landry, 2009).

Figure 1 outlines the basic steps involved in generating cDNA libraries for high-throughput sequencing of microbial transcriptomes, and the subsequent analysis of these. So far, all papers describing the use of high-throughput sequencing for bacterial transcriptomics have specified using the optional enrichment methods, usually based on depletion of tRNA and/or rRNA (Passalacqua et al., 2009; Perkins et al., 2009; Yoder-Himes et al., 2009). Size selection has also been used for the removal of mRNA and rRNA (Liu et al., 2009), although this is a potentially risky approach because this could remove long noncoding or antisense RNA species, as reported in *Listeria* and *Bacillus* (Rasmussen et al., 2009; Toledo-Arana et al., 2009). After sequence reads are mapped onto the genome sequence, these are usually visualized by generating histograms of reads on the annotated genome sequence, using a freely available software like ARTEMIS (Carver et al., 2008) or the Affymetrix Integrated Genome Browser (http://www.affymetrix.com) (Sittka et al., 2008). Figure 2 gives schematic examples of potential histograms for mono- and multicistronic mRNAs, noncoding RNAs and cis-acting RNA species.

The challenges set by bacterial transcriptome sequencing were first met in a study where two different isolates of *Burkholderia cenocepacia* were investigated (Yoder-Himes et al., 2009). The authors compared two strains, one isolated...
from soil and one from a cystic fibrosis patient, and used Illumina sequencing of cDNA libraries to define the responses of these two strains under conditions mimicking soil and cystic fibrosis. Interestingly, the authors reported the identification of 13 previously unknown noncoding RNA species [ncRNA, also often called small RNA (sRNA)], and also indicated that despite genomic similarity, the two B. cenocepacia strains displayed a significant difference in regulatory responses, which may explain their different habitats and pathogenic potential (Yoder-Himes et al., 2009).

A third study on microbial RNA-seq focused on Salmonella enterica serovar Typhi (S. Typhi) (Perkins et al., 2009). Illumina sequencing was used to sequence cDNA derived from RNA depleted of 16S and 23S rRNA genes. These authors demonstrated the importance of genomic DNA removal by DNase treatment of the RNA fraction, and used RNA-seq information to correct the annotation of the genome sequence, to identify transcriptionally active prophage genes, and to identify new members of the OmpR regulon. The information released also included 40 novel noncoding RNA sequences (Perkins et al., 2009).

An alternative use of high-throughput sequencing has been in the sequencing of immunoprecipitated RNA or DNA (IP-seq), which is an alternative to ChIP-on-chip experiments (Wade et al., 2007). A recent example of such an approach has been the simultaneous identification of sRNA and mRNA of S. enterica serovar Typhimurium, which were bound to the RNA chaperone Hfq (Sittka et al., 2008).

Opportunities, challenges and pitfalls

The rapid developments in sequencing technologies allow one to obtain very high-definition transcription snapshots, and these will, undoubtedly, significantly increase our insights in transcriptional and post-transcriptional events in microorganisms. Besides the increased insight into the process of transcription, it will also help in improving or

Fig. 2. Schematic representation of histograms of data that may be obtained using transcriptome sequencing. Examples are shown of monocistronic and polycistronic mRNAs, noncoding sRNA, cis-acting RNA species and antisense RNA. The transcriptional orientation is represented by the arrow at the baseline; black filled arrows represent annotated ORFs.
correcting the annotation of genome sequences (Denoeud et al., 2008). Identification of the 5′ and 3′ boundaries of mRNA species will inform us of the most likely translation initiation codon, especially in those cases where a ribosome-binding site is not apparent (Moll et al., 2002).

Next to technical challenges, the rapid increases in knowledge will be accompanied by new problems, as with previous breakthroughs in functional genomics (like genome sequencing and microarrays). Several issues may require action from the scientific community, and some of these are highlighted below.

1. Differentiation of transcriptional and post-transcriptional events. The sequencing-based approaches used for determining the bacterial transcriptomics to date are not able to distinguish between de novo transcription and post-transcriptional events, as they only record the levels of RNA (cDNA) present. This is a weakness shared with microarray technology. Alternative approaches such as those used for genome-wide determination of transcription start sites by 5′-rapid amplification of cDNA ends (RACE) and 5′-serial analysis of gene expression approaches (Hashimoto et al., 2004, 2009). These approaches use techniques distinguishing between primary (capped) RNA species, which result from de novo transcription, and processed (uncapped) RNA species. The combination with standard RNA-seq allows for specific identification of primary transcripts, and could be coupled to the use of rifampicin to inhibit transcription for the study of RNA stability (Mosteller & Yanofsky, 1970).

2. Standardization and database access. The rapid explosion in availability of microarray datasets prompted the release of the MIAME guidelines (Brazma et al., 2001), which established the minimal requirements for publication of microarray datasets in journals and online databases. Similar guidelines have been proposed for proteomics [minimum information about a proteomics experiment (MIAPE)] (Taylor, 2006) and genome sequences [minimum information about a genome sequence (MIGS)] (Field et al., 2008), and will likely have a positive effect on sequencing-based transcriptomics. Such guidelines should include instructions on the availability of datasets, statistical evaluation and deposition of sequence reads and annotation into online databases.

3. Removal of genomic DNA, rRNA and tRNA. One of the problems when working with bacterial RNA is that 50–80% of the total RNA content of bacteria is thought to be rRNA and tRNA. All the studies on sequencing-based microbial transcriptomics published to date have (partially) removed these rRNA and/or tRNA sequences (Liu et al., 2009; Passalacqua et al., 2009; Perkins et al., 2009; Yoder-Himes et al., 2009), but it is currently unknown as to what effect this has on the composition of the RNA fraction. With the advances in the number of reads and read length, it may, in the future, not be necessary to remove rRNA and tRNA and use unbiased cDNA libraries. Furthermore, improving the quality of the RNA preparation by removal of contaminating genomic DNA with DNase treatment improves the sequencing results, as was shown recently for S. Typhi (Perkins et al., 2009).

4. Construction of cDNA libraries. Choices for cDNA libraries are the use of reverse transcription using random hexamers, or alternatively the poly-A tailing of RNA (Wang et al., 2009). The subsequent choices for library construction will be dependent on the sequencing technology selected. It needs to be noted that the cDNA library construction may include amplification of cDNA, and hence has the potential to introduce an over-representation of shorter transcripts in the cDNA libraries constructed for sequencing.

5. Bioinformatic challenges. The large datasets produced by the different NextGen sequencing technologies come with their own challenges. Besides storage space, it will be important to have accurate sequence determinations to be able to map cDNA reads onto the genome, and to remove poor-quality sequences (Marioni et al., 2008; Jiang & Wong, 2009; Oshlack & Wakefield, 2009). There may also be issues with repeated sequences and homopolymERIC tracts at the 5′ or 3′ ends of cDNA reads, which can complicate 5′ RACE and 3′ RACE experiments. This is mostly a problem with 454 FLX sequencing as this is known to lack accuracy at homopolymERIC tracts. Like with many applications, larger datasets will allow more accurate determination of transcript levels and associated statistics, but will increase the risk of data deluge. Finally, visualization, analysis and interpretation will require significant levels of expertise, and may also require programming skills. Visualization may be achieved with the aforementioned ARTEMIS (Carver et al., 2008) and integrated genome browser (Affymetrix), but commercial programs like LASERGENE (DNAstar) also offer modules optimized for RNA-seq analyses.

Concluding remarks

Historically, research on microbial transcription focused on protein-based signal transduction and regulatory systems, and mRNA was seen as a relatively inert information carrier. However, the conventional view of RNA has changed in the last decade due to the discovery of regulatory and catalytic RNA activity (Waters & Storz, 2009). The significance of the discovery and application of microRNAs in eukaryotic and plant cells has been recognized by many recent awards, such as the 2006 Nobel Prize for Medicine for the discovery of RNA interference in eukaryotes, and the 2008 Lasker Award for Basic Medical Research for the discovery of microRNA regulation in plants. Similarly, bacteria express a variety of regulatory RNA species ranging from trans-acting RNAs (sRNA), cis-acting RNAs (riboswitches), antisense RNAs and protein-interacting RNAs (6S RNA, CsrB-like RNAs)
(Waters & Storz, 2009), and while our knowledge on these species is currently mostly based on E. coli, this is likely to change with the advent of sequencing-based transcriptomics. When combined with the latest developments in microarray technologies, like high-density tiling microarrays (Rasmussen et al., 2009; Toledo-Arana et al., 2009), we now have the ability to investigate transcription at single-nucleotide resolution. This is likely to enrich our knowledge of microbial diversity, and will undoubtedly show us the many different approaches used by bacteria to solve the problems encountered in their respective niches.

Acknowledgements

The author thanks the members of the research group and the collaborators, as well as three anonymous reviewers for helpful comments and suggestions. Research at the author’s laboratory is supported by the BBSRC Institute Strategic Programme Grant to the IFR.

References


