COMPREHENSIVE ANALYSIS OF TRANSCRIPTOME BY MEANS OF RNA-SEQ NEXT-GENERATION SEQUENCING

Karolina Janitz[1], Natalie Twine[2], Marc Wilkins[2], Michal Janitz[2]*

[1]Ramaciotti Centre for Gene Function Analysis, University of New South Wales, Sydney, NSW 2052

[2]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052

*Corresponding author: m.janitz@unsw.edu.au

Introduction

With the completion of the human genome project and the genome sequencing of many important model organisms with medical significance, re-sequencing applications have gained primary importance. They are powered by a group of conceptually new sequencing technologies, which are jointly called next-generation sequencing (NGS). Four commercially available sequencing technologies – Illumina, Roche/454, Applied Biosystems SOLiD, and recently, Helicos HeliScope – are characterised by higher throughput and lower costs than Sanger sequencing.

The arrival of next-generation sequencing techniques has significantly reduced sequencing costs and has also improved transcript coverage, making transcriptome analysis more easily available and useful for individual laboratories. This technological advancement has challenged the dominance of microarray systems, making it possible to introduce numerous new applications to transcriptome research. A few recent reports, which systematically compare microarrays and next-generation sequencing, have clearly proved the superiority of the latter, both with respect to the low frequency of false-positive signals and the high reproducibility of the method (1,2). A recent report by Hughes and co-workers concerning transcript analysis of intragenic regions unambiguously showed that an analogous method of collecting hybridisation signals from microarrays leads to massive false-positive signals from transcripts of low expression levels (3). In consequence, the results of many gene expression studies conducted to date using microarrays should be treated with caution. Numerous gene expression experiments performed in this way will probably have to be revised by the NGS technique based on digital signal detection directly from the cDNA template (see below).

RNA-Seq: Sequencing of an Entire Transcriptome

A transcriptome is the full set of transcripts in a cell, both in terms of the kind and the quantity. The arrival of highly efficient sequencing methods has dramatically improved the scope and efficiency of transcriptome research. mRNA sequencing (RNA-Seq) performs the analysis of complementary DNA (cDNA) by means of highly efficient, next-generation DNA sequencing methods and subsequent mapping of short sequence fragments (reads) onto the reference genome (**Fig. 1**). The fact that the new technology allows identification of exons and introns, mapping their boundaries and the 5′ and 3′ ends of genes, makes it possible to comprehensively understand the complexity of eukaryotic transcriptomes. Moreover, RNA-Seq enables the identification of the locations of the transcription initiation sites, new splicing variants and a precise quantitative determination of exons and splicing isoform levels (4).

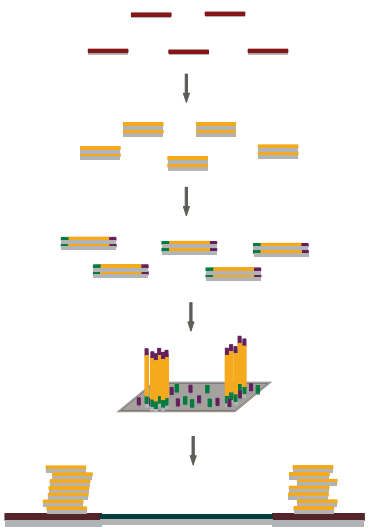

Fig. 1. RNA-Seq protocol flow-chart.
Long RNAs are first converted into a library of cDNA fragments through RNA fragmentation. Sequencing adapters are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing. The resulting sequence reads are aligned with the reference genome or transcriptome and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene, as exemplified in **Fig. 2**.

The Next-generation DNA Sequencing Technology

The high-throughput sequencing technique is the core component of the RNA-Seq application. In the following paragraphs, the next-generation sequencing technology will be described, illustrated by the Illumina Genome Analyzer (GA*IIx*), which we currently use in our research on the brain transcriptome in cooperation with Ramaciotti Centre for Functional Gene Analysis. Regarding other NGS platforms, we refer the reader to a recently published comprehensive monograph on the NGS technologies (5).

cDNA sequencing using the Illumina Genome Analyzer is conducted in two stages. During the first stage, a cluster

station is used to prepare a flow cell. The flow cell contains eight lanes, so that eight experimental samples can be analysed at a time. During the second stage, the DNA amplified in the flow cell is sequenced in the Genome Analyzer.

In the cluster station, the denatured DNA template intended for sequencing is loaded onto the flow cell, where it becomes linked to oligomers covalently bound with the surface of the flow cell. The second DNA strand is then synthesised from these oligomers, which results in formation of a double-stranded DNA molecule covalently attached to the surface of the flow cell. The double-stranded helix subsequently undergoes denaturation; the free ends of the immobilised strands are intercepted by the complementary, adjacent oligomers on the surface of the flow cell. A new second strand is then synthesised, which remains also covalently linked to the flow cell. This process, called bridging PCR, is repeated 35 times in order to generate a covalently bound cluster of identical DNA molecules on the surface of the flow cell. As a result of the bridging PCR process, the DNA cluster reaches a size of approximately 2,000 molecules. The more DNA initially applied onto the flow cell, the more densely packed DNA clusters will be. Up to a certain point, the sequencing throughput increases; however, when the molecule density increases, the analytic capacity of the Genome Analyzer become insufficient to distinguish between neighbouring clusters. Sequencing data generated from such overlapping clusters are useless due to an inability to unambiguously identify the incorporated nucleotide in a given position of the DNA chain.

The flow cell, previously prepared in a cluster station, is then placed in the Genome Analyzer in order to sequence the amplified DNA. Complementary DNA strands are denatured and a sequencing primer is used to initiate the strand elongation reaction. The primer is complementary to the sequences of Illumina oligomers attached to each strand of the template. Next, the Genome Analyzer performs the sequencing by synthesis, adding one pair of bases at a time to the clustered DNA. Each base is colour-coded with a fluorophore, the so-called reversible terminator (6), and the Genome Analyzer CCD camera records the colour of each cluster to determine which base has been added. Before commencing the next cycle, the fluorophores are removed. This last feature prevents ambiguity in the determination of homopolymer sequences, e.g., poly(A) tails. Subsequently, Illumina software, including Sequencing Control Studio (SCS) v.2.6, Integrated Real Time Analysis and Gerald modules, converts the fluorophore information into sequencing data. The Gerald module is also capable of mapping this sequence onto the reference genome. For RNA-Seq, however, it is recommended to use software packages, which are also capable of mapping readings with gaps indicating implied introns (see next section).

RNA-Seq Data Analysis

One of the challenges created by the new sequencing technology is the so-called read mapping. The Illumina Genome Analyzer generates short sequences of length 36–100bp, called 'short reads', which constitute the reads of fragments of sequences from a longer DNA or RNA molecule present in the experimental sample. As opposed to assembly of an entire genome, in which reads are grouped in contigs to reconstruct a previously unknown genome, many next-generation sequencing projects start from a known genome, the reference genome. The full sequencing of genomes of a range of important species, including the human and the mouse (7,8), has practically eliminated the problem of *de novo* mapping of genomes or transcriptomes of higher mammals.

To understand information comprised in short sequences, it is necessary to determine their position in the reference sequence. This process is known as alignment or 'mapping' of the reads onto the reference sequence. In the case of mapping genomic sequence, reads can be aligned without big gaps in the alignment. A more complicated situation occurs in the case of RNA-Seq, in which alignment is characterised by the presence of large gaps corresponding to introns. Alignment of the reads from RNA-Seq experiments may require thousands of hours of central processing unit (CPU) operation using conventional mapping tools such as BLAST (9) or BLAT (10). Fortunately, new software packages have recently been introduced that have been designed to meet the computational requirements of short read sequence analysis.

Application of RNA-Seq to Study Splicing Patterns

One example of a software package that can identify splicing locations by means of the mapping of RNA-Seq sequence reads is TopHat (http://tophat.cbcb.umd.edu/). TopHat maps reads at splicing locations in the mammalian genome, without relying on gene annotation, at a speed of over 2 million reads per hour of CPU operation. First, TopHat maps the reads, which are not located at splice junctions but within exons, by means of Bowtie (11). Thus, precise expression levels for particular exons can be determined, as shown on **Fig. 2**. Bowtie indexes the reference genome using a technique borrowed from data compression, the Burrows-Wheeler transform. This structure of data, which uses the memory in an efficient way, allows the Bowtie program to scan the reads in relation to the mammalian genome using approximately 2GB of memory, which remains within the computational capacity of a standard desktop computer. Once intra-exonic reads are identified by Bowtie, TopHat identifies reads covering boundaries of two, not necessarily neighbouring, spliced exons.

Recently introduced RNA-Seq protocols, which generate pair-end reads, make the task of TopHat easier. The frequency of false-positive results is much lower because information about paired fragments dramatically reduces the number of potential splicing events that need to be considered. There are several other open-source short read mappers, each of them with their own pros and cons. For comprehensive coverage of these software packages, please refer to an excellent review by Trapnell and Salzberg (12).

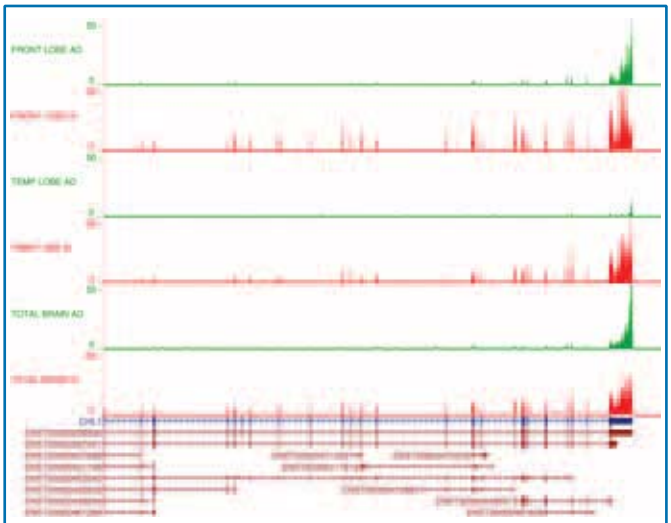

Fig. 2. Comparative analysis of the human cell adhesion molecule with homology to L1CAM gene (*CHL1*) between whole brain, as well as frontal and temporal lobe, from an Alzheimer's disease patient (shown in green) and reciprocal brain tissue samples from a healthy donor (shown in red) (Twine *et al.*, unpublished data). Gene expression and splicing pattern was evaluated using TopHat and Bowtie software and visualised using Ensembl browser. Short sequence reads were aligned to the reference *CHL1* gene sequence (shown in blue). Note that the reads map only to exon sequences shown as vertical bars along the reference gene. CHL1 alternatively spliced isoforms retrieved from Ensembl database (www.ensembl.org) are shown in dark red at the bottom.

RNA-Seq Helps to Understand the Complexity of the Transcriptome

Analytically dissecting each potential splice isoform across the genome has made it possible to discover a large number of alternatively spliced transcripts in human tissues. The consideration of various tissues has led to the discovery that as many as 95% of human genes containing several exons undergo alternative splicing. Exon skipping is the most frequent form of this regulation (13). These results considerably extend previous estimates, which suggested that approximately 60% of human genes undergo alternative splicing. Importantly, 92% of genes showed a second most frequently occurring isoform with relative frequency above 15% (14). Thus, in the majority of cases, several isoforms of the same transcript reach significant expression levels. The splicing patterns are usually tissue-specific, however, various alternatively spliced isoforms may also occur simultaneously in the same tissue. The latter phenomenon has been observed, for example, by the authors of this article in their own research of the transcriptome of the human brain using RNA-Seq (unpublished data).

RNA-Seq analysis conducted on the transcriptome of *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* provided an interesting insight into how simple, single-cell eukaryotes use alternative splicing as a mode of post-transcriptional regulation (15,16). In *S. pombe*, total splicing efficiency and transcript expansion levels seem to be positively correlated during vegetative growth and sexual differentiation. This implies coordination between transcription and splicing. Studies using RNA-Seq in *S. cerevisiae* showed a large number of alternative isoforms reflecting differential expression between the vegetative growth and a response to heat shock. What is interesting is that some of these isoforms probably encode proteins of various lengths. To sum up, it appears that splicing regulation can be used by single-cell eukaryotes to control and differentiate gene expression.

Bioinformatics tools, such as TopHat, help to identify specific levels of expression of various transcript isoforms in the enormous amounts of RNA-Seq data. In a situation when biological samples, subjected to RNA-Seq, come from patients suffering from complex diseases, there is a chance of identifying genes that, although having the same sequence at the genome level, differ by post-transcriptional processing of their mRNAs. This, in turn, may lead to the identification of molecular markers of the disease process. For instance, preliminary data from our laboratory suggest significant differences in the alternative splicing patterns between normal brain tissue and tissue affected by Alzheimer's disease (Twine *et al.*, unpublished data). Indeed, it is becoming clear that the complexity of the proteome of humans and other mammals does not mainly lie in the genomic sequence per se but in the enormous combinatory potential of the splicing process and other ways of processing of mRNA molecules (e.g., the regulation of 3'UTR length) before they eventually become translated.

Technological Perspective

NGS technologies, with their cost effectiveness and potential to sequence complex genomes in a single run, are revolutionising genomic research. The influence of the NGS technology on the analysis of gene regulation is particularly significant. RNA-Seq paved the way for new quality and scale in the study of the transcriptome. Indeed, it seems that within the next few years approaches based on sequencing will largely replace approaches based on microarray hybridisation. RNA-Seq makes it possible to sequence and to determine the number of transcriptomes with maximum resolution and dynamic range, regardless of the size of the transcript. Moreover, information on sequencing characteristics of the analysed genome is not required beforehand. The RNA-Seq technology has begun to change our way of thinking about examining the complexity and dynamics of transcriptomes and gene regulation. For example, surveying transcriptome profiles of a particular population might provide clues to our understanding of processes combining transcriptome plasticity with phenotypical diversity and evolution. The same holds true for comparative studies of the transcriptome on a single-cell level in complex organs such as brain. It might also be expected that with the rapid improvements of sequencing fidelity, throughput and cost-effectiveness, the next-generation technologies will soon be exploited in the fields of clinical diagnostics and pharmacogenomics.

References

1. Richard, H., Schulz, M.H., Sultan, M., *et al.* (2010) *Nucleic Acids Res.* **38**, e112
2. Sultan, M., Schulz, M.H., Richard, H., *et al.* (2008) *Science* **321**, 956-960
3. van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010) *PLoS Biol.* **8**, e1000371
4. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008) *Genome Res.* **18**, 1509-1517
5. Janitz M. (ed.) (2008) *Next-generation Genome Sequencing - Towards Personalized Medicine*, Wiley-VCH, Weinheim, ISBN: 978-3-527-32090-5
6. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., *et al.* (2008) *Nature* **456**, 53-59
7. Lander, E.S., Linton, L.M., Birren, B., *et al.* (2001) *Nature* **409**, 860-921
8. Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., *et al.* (2002) *Nature* **420**, 520-562
9. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *J. Mol. Biol.* **215**, 403-410
10. Kent, W.J. (2002) *Genome Res.* **12**, 656-664
11. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009) *Genome Biol.* **10**, R25
12. Trapnell, C., and Salzberg, S.L. (2009) *Nat. Biotechnol.* **27**, 455-457
13. Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008) *Nat. Genet.* **40**, 1413-1415
14. Wang, E.T., Sandberg, R., Luo, S., *et al.* (2008) *Nature* **456**, 470-476
15. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008) *Science* **320**, 1344-1349
16. Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bähler, J. (2008) *Nature* **453**, 1239-1243