BIOLOGY REPORTS

Published: 19 August 2010
© 2010 Faculty of 1000 Ltd

Recent advances in RNA sequence analysis

Steven L Salzberg

Address: Center for Bioinformatics and Computational Biology and Department of Computer Science, University of Maryland, College Park, MD 20742, USA

Email: salzberg@umd.edu

F1000 Biology Reports 2010, **2**:64 (doi:10.3410/B2-64)

Abstract

The latest high-throughput DNA sequencing technology can now be applied on a large scale to capture the complete set of mRNA transcripts in a cell, using a technique called RNA-seq. Although RNA-seq is only 2 years old, it has rapidly swept through the field of genomics, and it is now being used to analyze the transcriptomes of organisms ranging from bacteria to primates. The depth of sequencing allows researchers to quantify the level of expression of genes, to discover alternative isoforms in eukaryotic species, and even to characterize the operon structure of bacterial genomes.

Introduction and context

Sequencing the mRNA in a cell has been used as a high-throughput method for finding genes since the early days of the human genome project. Beginning in the early 1990s, the expressed sequence tag (EST) method was used to capture fragments of thousands of human genes [1] prior to the sequencing of the genome. EST sequencing relies on the fact that eukaryotic genes are polyadenylated after transcription, and the long poly-A tract can be used to capture the transcripts via reverse transcription PCR (RT-PCR). The EST method was subsequently applied to many other species, and EST databases (notably dbEST) became a vital resource for genome annotation. Recently, a 'next-gen' version of EST sequencing has emerged, allowing researchers to capture and sequence mRNA at dramatically lower cost, and higher volume, than was ever possible with the EST method. The new RNA-seq methods [2-5] are being applied to a rapidly growing variety of species, cell types, and scientific questions, revealing far more about the transcriptomes of these species than was known just a few years ago. The field is advancing so rapidly that a brief review cannot cover the work of the past 2 years; this review is just a sampling of a few highlights.

Major recent advances

Sultan *et al.* [6] analyzed approximately 8 million short reads and found that RNA-seq could detect 25% more genes as compared to microarrays. About one-third of transcripts in their experiments mapped to genomic regions not annotated as genes. Of the 94,241 splice junctions, 4096 were novel, and many of these indicated exon skipping events. This result has been amplified by subsequent studies that generated even more sequences and showed even larger numbers of novel splicing events. Trapnell *et al.* [7] generated approximately 430 million paired-end reads to recover 13,692 known isoforms from mouse myoblast cells, but also detected 12,712 novel isoforms, of which 7395 contained novel splice junctions while the rest represented novel combinations of known exons. This latter study also demonstrated the power of a new algorithm capable of detecting and quantifying alternative isoforms when aligning RNA-seq reads to a genome. In an RNA-seq study using liver RNA samples from humans, chimpanzees, and rhesus macaques, Blekhman *et al.* [8] found that alternative splicing events vary between closely related primates and also between the sexes within species. Wang *et al.* [9] generated approximately 600 million short reads from 15 cell types and found that 92-94% of human genes are alternatively spliced, and that many alternative splicing events are tissue-specific. RNA-seq is also being used to study genetic variation among individuals (expression quantitative trait loci, or eQTLs). Pickrell *et al.* [10] and Montgomery *et al.* [11] combined RNA-seq data and

HapMap data from 69 Nigerian individuals and 63 Caucasian individuals, respectively, and both groups identified variants responsible for alternative splicing as well as variation in expression levels among individuals.

In single-celled organisms, RNA-seq can reveal novel insights about polycistronic transcripts. In the first transcriptome analysis of *Trypanosoma brucei*, thousands of splicing and polyadenylation sites were identified and many genes were found to be differentially expressed between the parasite's two life-cycle stages [12]. In prokaryotes, RNA-seq can provide an extremely detailed transcription map, at the single-base level, as has been shown recently in an archaeal species, *Sulfolobus solfataricus*, and in a pathogen bacterium, *Helicobacter pylori*. In *S. solfataricus*, over 1000 transcriptional start sites were detected and 80 novel protein-coding genes were discovered [13]. In *H. pylori*, hundreds of transcriptional start sites within operons were found, as well as approximately 60 novel small RNA genes [14].

Future directions

The power of RNA-seq stems from its ability to generate deep coverage of the entire transcriptome of a cell with just a single run of a high-throughput sequencer, such as the Illumina HiSeq, which can produce up to 200 billion bases in a single run. The potential to characterize all genes, to capture alternative isoforms, and to measure differential expression has already been demonstrated in dozens of studies, but hundreds of species, and countless experimental conditions, are yet to be explored. Several groups have developed methods besides poly-A selection to capture all RNAs in a cell, for example, random hexamer priming [13,15], which allows them to analyze prokaryotic transcriptomes or to look at noncoding RNA in eukaryotes. It now appears that RNA-seq will replace microarray technology in the coming years, as it appears to be not only more comprehensive but also much more accurate than microarrays, particularly for transcripts with low expression levels [16]. As this new method becomes even more widely adopted, it should greatly expand our understanding of the complex interplay of genes in all phases of cell development.

Abbreviations

EST, expressed sequence tag; RT-PCR, reverse transcription PCR; poly-A, polyadenylase.

Competing interests

The author declares that he has no competing interests.

Acknowledgments

This was supported in part by National Institutes of Health grants R01-LM006845 and R01-GM083873.

References

1. Adams MD, Kerlavage AR, Fields C, Venter JC: **3,400 new expressed sequence tags identify diversity of transcripts in human brain.** *Nat Genet* 1993, **4:**256-67.

2. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5:**621-8.

3. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing.** *Science* 2008, **320:**1344-9.

   F1000 Factor 6.0 *Must Read*
   Evalauted by Bernd Weisshaar 15 May 2008

4. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: **Highly integrated single-base resolution maps of the epigenome in Arabidopsis.** *Cell* 2008, **133:**523-36.

   F1000 Factor 10.3 *Exceptional*
   Evalauted by Gregory Copenhaver 09 May 2008, Alisdair Fernie 13 May 2008, Craig Pikaard 20 May 2008, Elizabeth Dennis 22 May 2008, Matthew Sachs 30 May 2008

5. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM: **Stem cell transcriptome profiling via massive-scale mRNA sequencing.** *Nat Methods* 2008, **5:**613-9.

6. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML: **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.** *Science* 2008, **321:**956-60.

   F1000 Factor 3.0 *Recommended*
   Evalauted by Charles Auffray 02 Feb 2009

7. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28:**511-5.

8. Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y: **Sex-specific and lineage-specific alternative splicing in primates.** *Genome Res* 2010, **20:**180-9.

   F1000 Factor 3.0 *Recommended*
   Evalauted by Julin Maloof 09 Apr 2010

9. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456:**470-6.

   F1000 Factor 6.4 *Must Read*
   Evalauted by Ken Irvine 20 Nov 2008, Donald Rio 01 Dec 2008

10. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464:**768-72.

11. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET: **Transcriptome genetics using second generation sequencing in a Caucasian population.** *Nature* 2010, **464:**773-7.

12. Siegel TN, Hekstra DR, Wang X, Dewell S, Cross GA: **Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites.** *Nucleic Acids Res* 2010, [Epub ahead of print].

    F1000 Factor 3.2 *Recommended*
    Evalauted by Christine Clayton 22 Apr 2010, Marilyn Parsons 17 Jun 2010

13. Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R: **A single-base resolution map of an archaeal transcriptome.** *Genome Res* 2010, **20:**133-41.

F1000 Factor 3.0 *Recommended*
Evalauted by Shiladitya DasSarma 03 Mar 2010

14. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, Stadler PF, Vogel J: **The primary transcriptome of the major human pathogen *Helicobacter pylori*.** *Nature* 2010, **464:**250-5.

F1000 Factor 10.0 *Exceptional*
Evalauted by Vincenzo Scarlato 24 Feb 2010, Tracy Raivio 16 Apr 2010, Fiona Brinkman 21 Apr 2010, Michael Hensel 22 Apr 2010

15. Yoder-Himes DR, Chain PS, Zhu Y, Wurtzel O, Rubin EM, Tiedje JM, Sorek R: **Mapping the Burkholderia cenocepacia niche response via high-throughput sequencing.** *Proc Natl Acad Sci U S A* 2009, **106:**3976-81.

F1000 Factor 3.2 *Recommended*
Evalauted by Jo Handelsman 23 Mar 2009, Fiona Brinkman 18 May 2009

16. van Bakel H, Nislow C, Blencowe BJ, Hughes TR: **Most "dark matter" transcripts are associated with known genes.** *PLoS Biol* 2010, **8:**e1000371.

F1000 Factor 8.0 *Exceptional*
Evalauted by Daniel Reines 01 Jun 2010, Andre Nantel 03 Jun 2010