

Is sequencing enlightenment ending the dark age of the transcriptome?

Piero Carninci

Sequencing-based technologies for RNA discovery are playing a key role in deciphering the transcriptome and hold the potential to provide us with a census of RNAs and their functions.

A decade ago, before the appearance of the first draft of the human and other genomes, genes were identified via expressed sequencing tags. In the pregenome era, the main goal was to identify mRNAs produced by protein-coding genes. The computational tools used to predict these genes were far from accurate and required experimental data to support gene predictions. Contrary to expectations, just at the turn of the new millennium, initial genome studies suggested that the total number of genes was much lower than estimated earlier by analyses of expressed sequencing tag data¹, giving an upper bound of approximately 25,000 protein-coding mammalian genes.

The first doubts about this estimate came from the analysis of full-length cDNAs, which showed that in addition to protein-coding genes, there is an even larger number of RNA transcripts that do not encode proteins and are instead defined as noncoding RNAs (ncRNAs)². Additionally, the analysis of 1% of the human genome by the ENCODE (Encyclopedia of DNA Elements) Consortium not only suggested that 93% of the genome is transcribed but also revealed an unprecedented number of splice variants. These data put the number of predicted proteome variants at least fourfold above the number of genes³.

Indeed, an average cell contains at least 300,000 mRNA molecules and perhaps even more if we consider rare ncRNAs or

RNAs restricted to specific cell compartments. Different cells express different sets of mRNAs and ncRNAs, and their expression level spans 4–5 orders of magnitude from the highly to the rarely expressed RNA. Classical Sanger (dideoxy) sequencing of full-length cDNA, even if implemented with strategies to subtract the known genes to more efficiently search for new RNAs⁴, still requires the cloning of individual cDNAs before analysis of the clones; the largest full-length sequencing efforts so far have been limited to about 100,000 cDNAs². The main limitation of this approach is the need for very extensive handling and sequencing of individual cDNA clones. The advantage is that once the full-length cDNAs are sequenced, they not only serve to identify mRNAs, ncRNAs and their isoforms, but they can also be used in functional assays.

Tiling DNA microarrays, which probe the whole nonrepetitive part of the genome at high resolution⁵, are faster alternatives for transcript detection in that they instantaneously allow the comprehensive assessment of mRNA and ncRNA expression. Together with tagging technologies⁶ that identify novel RNA start and termination sites, tiling arrays have been instrumental in revealing a seemingly overwhelming degree of transcriptome complexity. However, they suffer from a lack of sensitivity in detecting rare transcripts and

from background owing to cross-hybridization of highly related sequences, making the analysis of repeat regions impossible. Additionally, tiling arrays cannot be used to identify new splice junctions: they will identify exons but do not provide their connections. Even arrays dedicated to the detection of predicted or known splice variants cannot detect novel splice sites that have not yet been experimentally identified or predicted *in silico*.

The arrival of second-generation sequencing has advanced unbiased transcriptome studies. The technology, known as RNA-sequencing (or RNA-seq) is procedurally quite simple: RNA is converted to a cDNA library without a lengthy cloning procedure, and a single cDNA preparation can yield over hundreds of millions of short reads that are then computationally aligned to the genome. Sequences mapping to a single exon are generally unambiguously assigned to the corresponding gene, often at a coverage of hundreds of sequences for each RNA. Chances are good that even rare RNAs can be identified in the library, although full assembly from the short reads is still unlikely for transcripts present in one copy per cell at the current coverage. With increased sequencing depth this may change, and it may become possible that RNA-seq will sensitively detect the whole transcriptome in a biological sample.

Indeed, a growing number of studies published in this journal and elsewhere^{7–11} have used RNA-seq to comprehensively investigate transcription. One of the obvious technical advantages of RNA-seq is the avoidance of spurious, false positive detection of RNA, an inherent problem of microarrays owing to cross-hybridization. RNAs that are produced by highly similar members of paralogous genes present an alignment challenge, but recently developed strategies also allow for the mapping of reads to splice sites or to paralogous sequences⁹. With read length increasing, it will become even easier to identify the exact location of most sequence reads.

In one of the first uses of RNA-seq, Mortazavi and colleagues developed an

Piero Carninci is at the Omics Science Center, RIKEN Yokohama Institute, Yokohama, Kanagawa, Japan.
e-mail: carninci@riken.jp

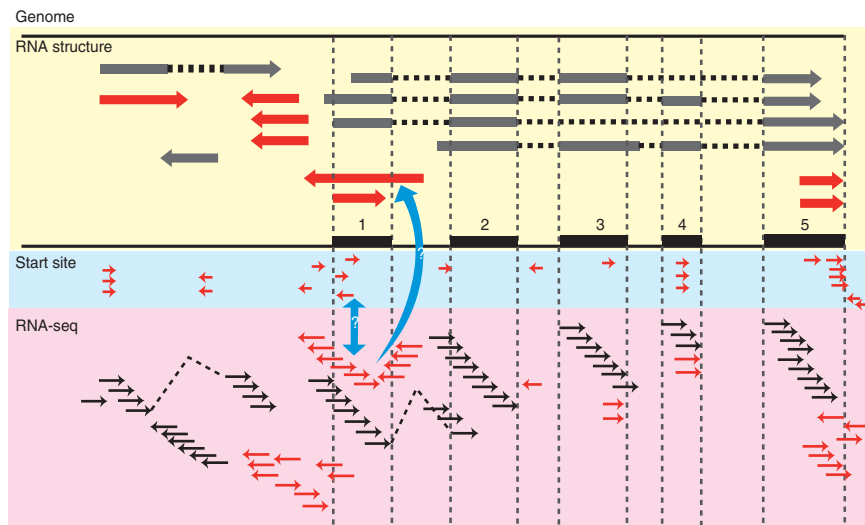


Figure 1 | Challenges in assembling RNA-seq data. A hypothetical locus often produces multiple RNAs (top), which may affect assembly of RNA-seq into an unambiguous gene model. A given annotated protein-coding gene (the known annotated region is represented by back boxes, and known exons are numbered 1–5) produces protein-coding mRNAs (gray boxes and arrows) and other noncoding RNAs (large red arrows). RNA-seq produces multiple overlapping signals (bottom), represented by black and red tiled arrows. Although RNA-seq allows reconstruction of exon structure and neighboring exon-exon junction definition, the connectivity of exons as well as all the splice combinations of individual transcripts may not be fully reconstructed. By looking at RNA-seq data, the reconstruction of connectivity in the various transcripts between distant exons is challenging. Additionally, RNAs have multiple transcription start sites (middle, small red arrows), identified, for instance, by cap analysis gene expression (CAGE tags), which cannot be detected efficiently by RNA-seq.

RNA-seq method in which oligo-dT-selected mouse mRNAs are fragmented before synthesis of random-primed, first-strand cDNA⁹. This RNA fragmentation proved essential for uniform coverage of all exons. Random priming performed on full-length mRNA would show a strong preference to a few particular positions on the mRNAs and thus bias coverage. Although the method was limited to the short 25-nucleotide reads available in the first version of the Illumina GA sequencer, the authors impressively provided a clear pattern of novel splice sites. These findings explore the world of protein-coding mRNAs, as noncoding RNAs often lack poly(A) tails, and their data suggest that a good fraction of protein-coding exons has already been identified because only a few percent of the tags mapped to intergenic regions.

In parallel, Sean Grimmond and colleagues went for even larger transcriptome coverage in human embryonic stem cells and embryoid bodies using Applied Biosystems' sequencing by ligation (SOLiD) technology¹⁰. They reasoned that RNAs are often transcribed bidirectionally and often overlap to form sense-antisense pairs. To capture these sense-antisense

RNA pairs, they devised a template switch strategy that maintains the 5'–3' directionality of the original RNAs in the shotgun library: the reverse transcriptase adds (albeit at low efficiency) some extra bases at the end of the first-strand cDNA, so the second-strand cDNA can be primed from the overhang. Another important procedural step was that they subtracted ribosomal RNAs from the total RNA preparation before cDNA library preparation; otherwise ribosomal RNAs would comprise a large part of the library, affecting sequencing yield. The authors showed that sense-antisense pairs are more concentrated at the 3' end of genes.

A remaining challenge is the detection of rare RNAs expressed only in a few subtypes of cells. This is a key issue in analyzing complex tissues, such as in the brain and those formed during embryonic development. A first step toward meeting this challenge has been made by Surani and colleagues, who prepared seven libraries from the limited amount of RNA in single mouse oocytes—cells in an early developmental stage¹¹. Owing to the protocol, which is based on oligo-dT priming followed by A-tailing of cDNA and subsequent cDNA

fragmentation before sequencing, the direction of the RNA is lost. There are no protocols yet for the removal of ribosomal RNA when only a few nanograms of RNA are available. Also, oocytes, though technically single cells, are much larger than an average cell, which leaves room for the development of technologies to miniaturize single-cell sequencing-based assays.

It is certainly reassuring that different sequencing and library preparation technologies have pointed unambiguously to similar levels of complexity and all have provided solutions for tackling this complexity.

Indeed, RNA-seq is contributing to the comprehensive analysis of splice events and novel splicing combinations. Even with relatively 'shallow' coverage, such as less than 10 million 27-nucleotide short sequences⁷, researchers found more than 2,000 new exons for each cell line derived from kidney and B cells. Deeper sequencing of longer fragments allowed investigators to detect millions of splice sites and ~90,000–145,000 novel splice site candidates per tissue^{8–11}. This clearly supports the notion that multi-exon genes are usually alternatively spliced, but these studies have also demonstrated that alternative splicing is restricted to only a fraction of exons.

RNA-seq also deals much better with RNA derived from repeat elements¹⁰, which have been previously excluded from arrays because of cross-hybridization. Notably, even 20-nucleotide sequence tags have been suitable for detection of widespread expression of retrotransposons¹².

Have we addressed all the challenges yet? A final unambiguous assembly of different splice isoforms is still outstanding because the connectivity between distant exons that do not share sequencing reads cannot be unequivocally predicted (Fig. 1). The ENCODE analysis working group is currently developing computational methods to assemble RNA-seq reads according to RNA models with accurate splice sites and to predict the most likely splice variants based on the expression of each exon. At the same time, sequencing technology is constantly improving, reads are longer—and thus the assembly of full mRNA sequences is becoming simpler—and greater support is being provided with the second- and third-generation sequencing technologies. The Roche 454 Life Sciences sequencer can produce over a million sequences around 500 nucleotides long; this will be helpful

for assembling splicing sites. The ABI SOLiD and the Illumina GAIIX have not only increased the sequencing length to 50 and 75 bases, respectively, but have also developed methods for sequencing from both ends of the cDNA fragments to help in connecting more distant exons.

Other challenges of RNA-seq are how to distinguish the various start and end sites of RNAs. It is becoming evident that there are often multiple overlapping RNAs encoded from the same genome region, and intron-derived RNAs are recycled to produce functional ncRNAs such as microRNAs. Another source of complexity comes from the secondary processing of mRNAs, which produces shorter, likely functional, RNAs. Thus, protein-coding genes are associated with a plethora of short ncRNAs, including short RNAs associated with promoters¹³, transcripts arising around termination sites and even exons. A fraction of these RNAs are produced by a novel cleavage and recapping mechanisms, resulting in capped RNAs that start in the middle of coding exons or in untranslated regions. These naturally truncated RNAs are likely to be ncRNAs that overlap larger mRNAs¹³. Another complication arises from the broad nature of many promoters¹⁴, which produce various capped RNAs from multiple transcription start sites. Technologies that identify the cap structure in such mixtures are needed to distinguish the RNA fragments obtained by RNA-seq. At present RNA-seq does not perform well at unambiguously identifying transcription start sites, and RNA-seq protocols need improvement to simultaneously decipher the long, short and capped RNAs so the RNAs' function can be assessed.

Some of the third-generation sequencers such as those from Pacific Biosciences and Oxford Nanopore—which will be able to read thousands of nucleotides¹⁵ of single cDNAs—may ultimately meet these challenges: their long sequences will quantitatively represent complete RNAs, and the use of tags and linkers that mark cap sites and other modifications will allow an all-in-one determination of transcriptome structure, including start and termination sites and the mapping of regulatory elements such as promoters. The accurate sequence of coding sequences will also help directed cloning of open reading frames in experimental verification of alternative splice isoforms^{16,17}.

Although many challenges are ahead, the direction is becoming clearer, and I am

beginning to wonder if the dark age of the transcriptome is giving way to rays of light.

- Liang, F. *et al. Nat. Genet.* **25**, 239–240 (2000).
- Carninci, P. *et al. Science* **309**, 1559–1563 (2005).
- ENCODE Project Consortium *et al. Nature* **447**, 799–816 (2007).
- Carninci, P. *et al. Genome Res.* **13**, 1273–1289 (2003).
- Kapranov, P. *et al. Science* **316**, 1484–1488 (2007).
- Harbers, M. & Carninci, P. *Nat. Methods* **2**, 495–502 (2005).
- Sultan, M. *et al. Science* **321**, 956–960 (2008).
- Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. *Nat. Genet.* **40**, 1413–1415 (2008).
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. *Nat. Methods* **5**, 621–628 (2008).
- Cloonan, N. *et al. Nat. Methods* **5**, 613–619 (2008).
- Tang, F. *et al. Nat. Methods* **6**, 377–382 (2009).
- Faulkner, G.J. *et al. Nat. Genet.* **41**, 563–571 (2009).
- Fejes-Toth, K. *et al. Nature* **457**, 1028–1032 (2009).
- Carninci, P. *et al. Nat. Genet.* **38**, 626–635 (2006).
- Turner, D.J., Keane, T.M., Sudbery, I. & Adams, D.J. *Mamm. Genome* **20**, 327–338 (2009).
- Djebali, S. *et al. Nat. Methods* **5**, 629–635 (2008).
- Salehi-Ashtiani, K. *et al. Nat. Methods* **5**, 597–600 (2008).

Engineered fluorescent proteins: innovations and applications

Michael W Davidson & Robert E Campbell

Despite expansion of the fluorescent protein and optical highlighter palette into the orange to far-red range of the visible spectrum, achieving performance equivalent to that of EGFP has continued to elude protein engineers.

Evolving proteins, evolving tools

During the past decade and a half, intrinsically fluorescent proteins have been under intense evolutionary pressure for 'fitness', not in the wild, but rather for utility in live-cell imaging experiments. This unnatural course of evolution has occurred on the benches of protein engineers around the world who have helped to drive progress in the ever-expanding repertoire of fluorescence imaging technologies.

An underlying theme that has guided advancements in fluorescent protein engineering is that, all other factors being equal, redder is better. It is generally accepted that excitation with longer-wavelength light entails less phototoxicity for the cells or tissue being examined and decreased autofluorescence and scattering. These desirable factors mean that red-shifted fluorophores generally provide improved contrast (owing to decreased background fluorescence) and superior performance in whole-organism imaging (owing to higher tissue 'transparency'). Early efforts to engineer red-shifted

Aequorea victoria GFP (avGFP) variants led to the development of enhanced GFP (EGFP) and yellow fluorescent proteins with emission maxima at approximately 507 nm and 529 nm, respectively (versus 508 nm for wild type)¹.

For a time, however, it appeared that fluorescent protein engineering had hit a 'yellow' wall in efforts to red-shift fluorescence emission. Fortunately, this barrier had already been surmounted by natural evolution, as was revealed in October 1999 with a report that the *Discosoma* sp. mushroom anemone harbored a fluorescent protein homolog, commonly known as DsRed, emitting in the orange-red region (583 nm)². Counterbalancing this favorable shift to the red were several undesirable properties, including oligomerization, 'contamination' by a green component and sluggish chromophore development, which dampened some of the initial enthusiasm.

The discovery of DsRed (and other Anthozoa fluorescent proteins of various hues) had a twofold impact on the

Michael W. Davidson is at the National High Magnetic Field Laboratory and Department of Biological Science, Florida State University, Tallahassee, Florida, USA. Robert E. Campbell is at the University of Alberta, Department of Chemistry, Edmonton, Alberta, Canada.
e-mail: robert.e.campbell@ualberta.ca