Minireview

Transcriptome content and dynamics at single-nucleotide resolution

Nicole Cloonan and Sean M Grimmond

Address: Institute for Molecular Bioscience, University of Queensland, 306 Carmody Road, St Lucia, 4072, Australia.

Correspondence: Sean M Grimmond. Email: s.grimmond@imb.uq.edu.au

Published: 18 September 2008

*Genome **Biology*** 2008, **9:**234 (doi:10.1186/gb-2008-9-9-234)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/9/234

Abstract

Massively parallel short-tag sequencing of cDNA libraries - RNAseq - is being used to study the dynamics and complexity of eukaryotic transcriptomes, giving new biological insights into the ‘active genome’.

With the advent of third-generation sequencing technologies - the so-called massively parallel sequencing technologies - it is now possible to generate tens of millions of short sequences (each typically 25-50 nucleotides long) in a single assay. This technology has enabled the recent ‘RNA sequencing’ (RNAseq), via random cDNA libraries, of the transcriptomes of yeasts, *Arabidopsis*, mouse embryonic stem (ES) cells and other mouse tissues, and human cell lines. These experiments are helping to redefine the understanding of transcriptome content, complexity, and dynamics in these species. A recent study by Bähler and colleagues [1] in the fission yeast *Schizosaccharomyces pombe* in particular shows how the new RNAseq technology is ideally suited to revealing the changes that occur in transcriptional activity at different stages in the yeast life cycle and in response to changes in external conditions.

Conceptually, the RNAseq approach is very simple. Sequence reads are generated from random locations along each RNA by either sequencing sheared double-stranded cDNA libraries [1-4] (strandless RNAseq), or by sequencing directional cDNA libraries prepared using either adaptor-tagged random hexamers [5], or serial ligation of adaptors [6] to fragmented RNA populations (stranded RNAseq). After sequencing *en masse*, the short reads are then mapped back against the appropriate reference genome or catalogues of all exon-junction sequences to provide a global survey of transcriptome activity (Figure 1).

Advantages of RNAseq for investigating the transcriptome

RNAseq has several advantages over microarrays, the traditional workhorse for transcriptomics. First, gene-expression profiling by RNAseq has been shown to be very robust and highly quantitative. The reproducibility of the approach has been shown to be extremely high (Pearson correlations of 0.99 have been reported for replicate RNAseq runs) and raw tag counts correlate well with quantitative real-time PCR results. For a microarray experiment, where image-derived intensities are used to determine relative abundance of transcripts, the dynamic range of expression is constrained to a maximum of four to five orders of magnitude. Although rare transcripts can be detected by prolonging image exposure, the image becomes saturated for the most highly expressed transcripts, and the relative expression of these transcripts is lost. In contrast, the dynamic range of RNAseq is potentially unlimited, as tag counts are used to directly determine transcript abundance.

RNAseq is also potentially far more sensitive than microarray platforms. When sequence depths of 10-100 million reads per biological sample are compared with expression arrays, many genes whose activities are below detection limit on the array are readily observed. Importantly, this sensitivity is tuneable by altering sequencing depth. In the case of *S. pombe*, for example, Wilhelm *et al.* [1] showed that less than 1 Gb of mappable sequence was required for

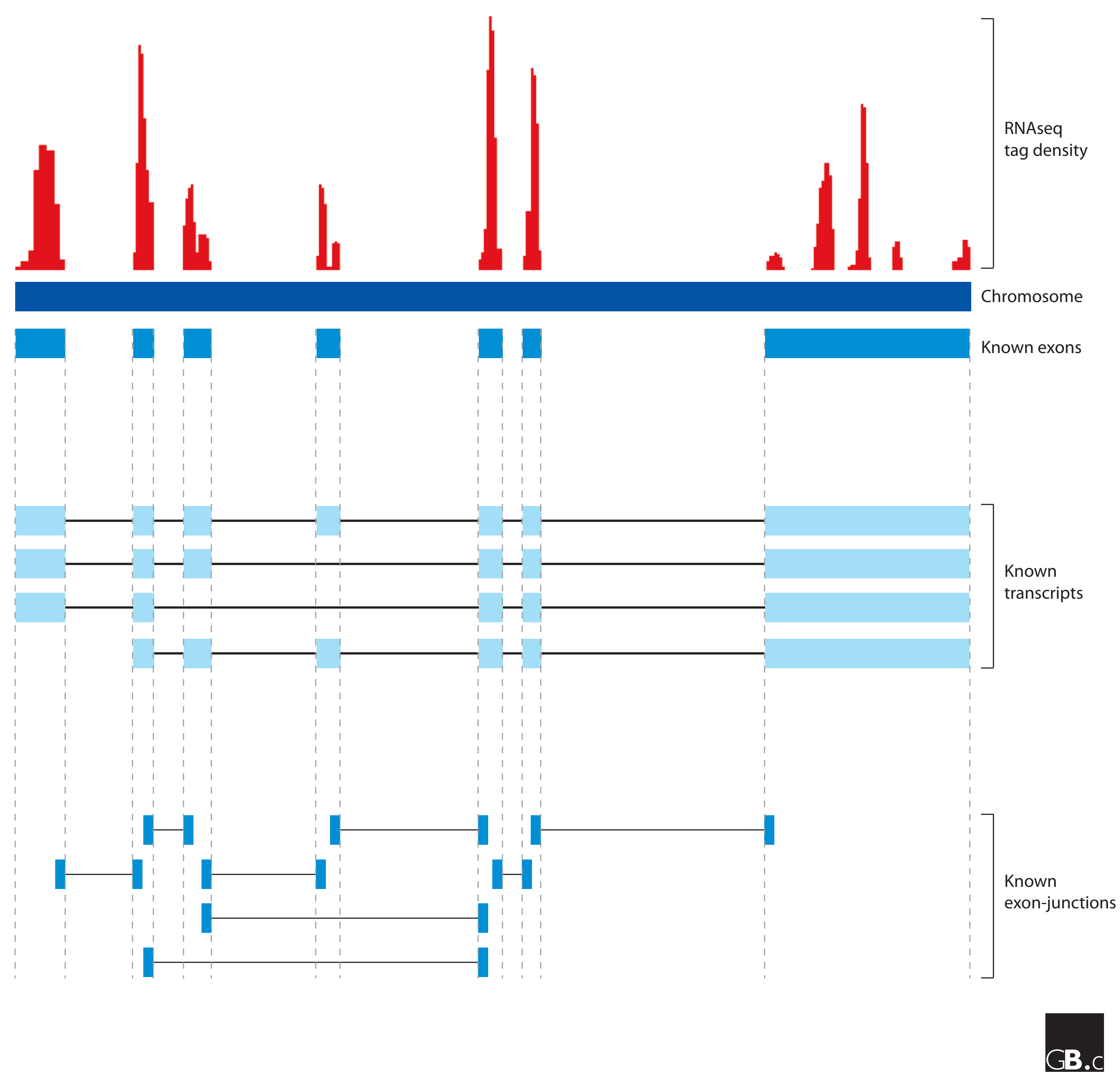

Figure 1
The identification of differential exon splicing by RNAseq. In this hypothetical genome-browser view, RNAseq tags (shown in red) are aligned to the genome sequence, giving a quantitative view of tag densities across the locus. Genome-aligned reads identify individual exons and exon-exon junction usage can be monitored by matching tags to a reference set of junction sequences. Differential exon-exon junction usage can be used to identify canonical and alternative splicing events.

complete mRNA coverage, and other work has shown that no more than 3-4 Gb is needed to obtain near complete coverage of mammalian transcriptomes [3,5].

Expression profiling by RNAseq is also far more precise than hybridization-based approaches, where RNAs sharing more than 75% sequence identity to the probe will cross-hybridize [7]. The high levels of sequence identity used for mapping (96-100%) allow one to profile highly homologous transcripts that would otherwise be confounded by cross-hybridization in microarray-based experiments. Repetitive sequences have always been excluded from array probe designs for this very reason, and these elements can now be profiled using RNAseq.

Unlike arrays that use a defined set of probes to interrogate RNA samples, RNAseq requires no previous assumptions about which parts of the genome are transcriptionally active. This provides the opportunity for transcriptome discovery, and large amounts of novel expression have been reported in budding and fission yeast [1,2], *Arabidopsis* [6], mouse ES cells [5], mouse tissues [3] and human cell lines [4]. As much as 25% of all observed expression in RNA sequence experiments falls outside known exons for mammalian genomes [3-5].

Biological insights from RNAseq

The large-scale survey of gene-expression space by traditional sequencing of expressed sequence tags (ESTs) [8,9], and the transcriptome annotation efforts of both the FANTOM [10-12] and ENCODE [13] consortia have shown that mammalian genomes are capable of generating 6-10 transcripts per locus. Until now, one of the major challenges in transcriptomics has been how to survey which of these known RNAs are present in a single biological state. In addition to monitoring gene activity, RNAseq can study alternative splicing events, and the usage of promoters and 3' untranslated regions (3' UTRs). These events can be detected by counting tags that match the portions of sequence unique to each transcript. These so-called 'diagnostic' sequences may correspond to cassette exons or the junction sequences arising from specific exon combinations.

The use of RNAseq has, for the first time, enabled researchers to rapidly place genome-wide surveys of both known and novel transcriptional complexity into a biological context. For example, a recent RNAseq survey of human embryonic kidney (HEK) 239T and Ramos B cells has shown exon skipping to be the most common form of alternative splicing in these cell lines [4]. However, by examining different tissues, treatments or time-points, patterns of transcriptional complexity can be placed into biological context. A common observation is the alternative use of extended UTRs [1-3]. In *S. pombe*, Wilhelm *et al.* [1] demonstrated that the lengths of 5' and 3' UTRs could be alternatively regulated under different environmental conditions. During a sexual differentiation time course, more than 20 genes were identified with dynamic 5' UTR lengths, predominantly transitioning from short to long UTRs as cells exit mitosis. Many of the genes identified have known functions in the cell cycle, or were associated with the cell wall/cell surface; however the mechanism governing UTR usage is not yet clear. The parallel finding that mRNAs known to be unstable had longer UTRs suggests that the extended UTRs may contain regulatory signals that affect the stability of the mRNA. If correct, this would enable tighter regulation of protein levels during specific biological processes, implying flexibility in regulatory networks.

In addition, Wilhelm *et al.* [1] showed that pre-mRNA splicing in fission yeast is dynamically and biologically regulated on a genome-wide scale. By surveying both rapidly proliferating mitotic cells and induced meiotic reproduction using a combination of strandless RNAseq with high-density tiling arrays, they showed that as the expression of transcripts increased, the efficiency with which those transcripts were spliced (and therefore the overall proportion of spliced to unspliced transcripts) also increased. These results point to a functional link between transcriptional and splicing machinery in *S. pombe*. A physical interaction between transcription and pre-mRNA splicing has already been established in mammalian cells (reviewed in [14,15]), and the findings of Wilhelm *et al.* in yeast could highlight a potential evolutionarily conserved mechanism to ensure the efficiency of pre-mRNA processing [16].

In addition to identifying the presence and relative abundance of known transcripts, RNAseq has regularly identified novel transcriptional content and complexity. This includes more than 200,000 retrotransposable elements identified as transcriptionally active in mouse ES cells, of which as many as 30,000 of these elements are dynamically expressed during mouse ES cell differentiation [5]. RNAseq is also not just a means for measuring the relative abundance of transcripts, it is a massive-scale survey of sequence content, enabling the simultaneous analysis of gene expression and screening for sequence variation. While this has not yet been pursued to any extent so far, the ability of RNAseq to identify novel single nucleotide polymorphisms (SNPs) in exons has been shown [5].

Challenges for RNAseq technology

As with any new technology, there are currently various limitations to RNAseq that will need to be addressed. First and foremost is that it is based on resequencing. This means that RNAseq is more useful for organisms that already have good-quality reference genome sequences. Furthermore, the ability to monitor transcriptional complexity as reported in the recent papers is built on the foundation of previous large-scale transcriptome annotation. In species where genome builds are not complete, or where there is limited EST and mRNA characterization, inferring the scale and scope of transcriptional complexity will be challenging.

Mapping of the RNA sequence tags to the genome sequence provides much needed precision in distinguishing the expression of homologous genomic sequences, but it is still not possible to discern the origin of a sequence tag that maps to more than one location. This means that parts of the transcriptome are undecipherable (known as multi-mapping or ambiguous regions). Given that tags are typically between 20 and 40 nucleotides, approximately 10-20% of tags can multi-map, especially when mapping strategies normally allow for a number of mismatches between the tag and the reference sequence to account for SNPs or systematic error. While tags of limited ambiguity (those that map to less than five different locations) can be assigned to their most likely origin using computational approaches, such events are undesirable. This is especially true when examining novel events associated with expression from those regions, such as novel splice variants or SNPs. Highly ambiguous exonic sequences will remain 'black holes' in the genome until advances in sequencing technology increase the read length to the point where the tag extends into sufficient unique sequence to allow unambiguous mapping. In addition, the effect of genomic variation (such as copy-number variations, structural variations or ploidy) on accurate and unbiased tag mapping has yet to be systematically studied. Nevertheless, it is anticipated that improvements and advances in these technologies will see tag lengths increase and systematic error decrease, both of which will dramatically shrink the current black holes in reference genomes.

There is also opportunity for improvement in almost every part of the RNA workflow. In the case of library preparation, most experiments have used double-stranded cDNA libraries that are sheared and then sequenced *en masse*. The problem with this approach is that the directionality of the fragment is lost, so that precise mapping of its origin to a specific strand in the genome is also lost. In higher eukaryotes, where overlapping sense and antisense transcripts are abundant, this is far from desirable. It is also unclear how PCR or ligation steps in other library preparation methods may bias the tag content of the libraries.

Current RNAseq methods are not yet suitable for samples where only small amounts of RNA are available. The various reported protocols use from as little as 20 mg to as much as 1 mg of total RNA. While it is conceivable that RNA amplification steps could be used to generate enough starting material, those libraries would be significantly less complex and, therefore, their sensitivity would be compromised.

Despite these caveats, RNAseq is heralding a new period in transcriptomics and is bringing much needed sensitivity and discrimination to global gene expression assays. The power of the new sequencing technologies means it is now feasible to sequence the complete transcriptome in short random fragments, thus providing the opportunity to measure the expression of all known transcripts as well as systematically screening for novel expression. As with microarray profiling, it is anticipated that as the number of biological states surveyed by RNAseq increases in each species, it will be possible to put much of this new-found complexity into biological context. Being able to accurately survey sequence variation and gene activity simultaneously should enable a single experiment to yield large amounts of diverse information: for example, screening for mutations, monitoring allele-specific expression and studying post-transcriptional events, such as RNA editing, simultaneously in a single pathological sample.

References

1. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J: **Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.** *Nature* 2008, **453:**1239-1243.
2. 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:**1344-1349.
3. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5:**621-628.
4. 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 M-L: **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.** *Science* 2008, **321:**956-960.
5. Cloonan N, Forrest AR, Kolle G, Gardiner BBA, 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:**613-619.
6. 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:**523-536.
7. Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ: **Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays.** *Nucleic Acids Res* 2000, **28:**4552-4557.
8. Boguski MS, Schuler GD: **Establishing a human transcript map.** *Nat Genet* 1995, **10:**369-371.
9. Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J: **The TIGR gene indices: clustering and assembling EST and known genes and integration with eukaryotic genomes.** *Nucleic Acids Res* 2005, **33:**D71-D74.
10. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest ARR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato X, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, *et al.*: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309:**1559-1563.
11. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, Adachi J, Fukuda S, Aizawa K, Izawa M, Nishi K, Kiyosawa H, Kondo S, Yamanaka I, Saito T, Okazaki Y, Gojobori T, Bono H, Kasukawa T, Saito R, Kadota K, Matsuda H, Ashburner M, Batalov S, Casavant T, Fleischmann W, *et al.*: **Functional annotation of a full-length mouse cDNA collection.** *Nature* 2001, **409:**685-690.
12. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, Yamanaka I, Kiyosawa H, Yagi K, Tomaru Y, Hasegawa Y, Nogami A, Schonbach C, Gojobori T, Baldarelli R, Hill DP, Bult C, Hume DA, Quackenbush J, Schriml LM, Kanapin A, Matsuda H, Batalov S, Beisel KW, Blake JA, Bradt D, *et al.*: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420:**563-573.
13. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng ZP, Snyder M, Dermitzakis ET, Stamatoyannopoulos JA, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, *et al.*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447:**799-816.
14. Allemand E, Batsché E, Muchardt C: **Splicing, transcription, and chromatin: a ménage à trois.** *Curr Opin Genet Dev* 2008, **18:**145-151.
15. Kornblihtt AR, De La Mata M, Fededa JP, Munoz MJ, Nogues G: **Multiple links between transcription and splicing.** *RNA* 2004, **10:**1489-1498.
16. Hicks MJ, Yang C-R, Kotlajich MV, Hertel KJ: **Linking splicing to Pol II transcription stabilizes pre-mRNAs and influences splicing patterns.** *PLoS Biol* 2006, **4:**e147.