PERSPECTIVE

Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes

Benjamin J. Blencowe,[1,2,4] Sidrah Ahmad,[1,2] and Leo J. Lee[1,3]

[1]Banting and Best Department of Medical Research, Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada; [2]Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada; [3]Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario M5S 3G4, Canada

Recent papers have described the first application of high-throughput sequencing (HTS) technologies to the characterization of transcriptomes. These studies emphasize the tremendous power of this new technology, in terms of both profiling coverage and quantitative accuracy. Initial discoveries include the detection of substantial new transcript complexity, the elucidation of binding maps and regulatory properties of RNA-binding proteins, and new insights into the links between different steps in pre-mRNA processing. We review these findings, focusing on results from profiling mammalian transcriptomes. The strengths and limitations of HTS relative to microarray profiling are discussed. We also consider how future advances in HTS technology are likely to transform our understanding of integrated cellular networks operating at the RNA level.

The transcriptome can be described as the complete list of all classes of RNA molecules, whether coding or noncoding, expressed in a particular cell, tissue, or whole organism. An RNA transcript can be subject to a wide array of different regulatory processes, and may itself serve in a critical regulatory or enzymatic capacity. The central role played by RNA, both as a template for protein expression as well as a regulatory molecule, has prompted growing interest in attempting to comprehensively catalog cellular (and viral) transcripts in biologically important contexts. These cataloging efforts will facilitate an integrated understanding of the diverse roles of RNAs comprising transcriptomes. This information will enable the elucidation of sequence- and structure-based RNA codes that operate to control different cellular regulatory programs. An ultimate goal of these efforts is to be able to accurately predict functional properties of RNAs from sequence features alone, and also how these functions are altered in human diseases. Technological advances permitting a detailed, quantitative, and rapid characterization of the transcriptomes of cells and tissues is a critical step toward achieving such an understanding. The development of high-throughput sequencing (HTS) methods for analyzing RNA populations, also known as "RNA-Seq" (or "mRNA-Seq" in the case of mRNA sequencing) has provided a major step forward in this direction.

During the past decade, microarray technologies have played a prominent role in shaping our understanding of transcriptome complexity and regulation. Currently, most microarray profiling systems employ glass slides containing thousands to millions of anchored oligonucleotides designed to hybridize to transcript sequences of interest. A major drawback of this approach is that profiling coverage is strictly limited by the probe sets available for specific hybridization on the microarray. Although genome-wide "tiling" microarrays have been widely used, most available systems employ spaced oligonucleotides and lack probes for the specific detection of junction sequences formed by RNA processing events. Further contributing to limited sensitivity and specificity is that detection is indirect, typically measured as a fluorescence signal, and is subject to a variety of confounding noise variables. In contrast, RNA-Seq provides a relatively unbiased and direct digital readout of cDNA sequence generated from an RNA sample. Several studies have demonstrated that RNA-Seq provides an extremely reproducible and quantitative readout of transcript abundance (e.g., Li et al. 2008; Marioni et al. 2008; Pan et al. 2008; Wang et al. 2008). Because RNA-Seq is performed using tagged libraries of short cDNAs, prepared from fragmented or unfragmented RNA, it does not require prior knowledge of the sequences to be profiled. This feature, coupled with the massively parallel nature of the technology, allows tens of millions of short sequence reads to be generated in a few days.

A current drawback of HTS is the overall cost required for generating such large data sets, which can run in the range of several thousand dollars for tens of millions of reads. However, when considering the price per sequenced base and the quality of the data obtained (which

[*Keywords*: Gene regulation; RNA-binding protein; RNA processing; RNA-Seq; transcriptome]
[4]Corresponding author.
E-MAIL b.blencowe@utoronto.ca; FAX (416) 946-5545.
Article is online at http://www.genesdev.org/cgi/doi/10.1101/gad.1788009.

can save significant time and therefore cost in data processing and downstream analyses), HTS is actually less expensive than microarray profiling methods. Currently, the most widely used systems to generate RNA-Seq data are those developed by Illumina (formerly Solexa), Applied Biosystems (AB), and 454 Life Sciences (Roche). The Illumina and AB systems can produce data sets comprising tens of millions of reads, currently at ~50 or more nucleotides per read, during a single 2- to 3-d run. The Roche system generates data sets typically consisting of a few hundred thousand reads at 400–500 nucleotides (nt) per read. These systems are rapidly evolving, and at the time of writing it is anticipated that at least one of these systems will afford a severalfold increase in read yield per run, and that the generation of substantially longer reads at higher yields will also soon be feasible.

The unparalleled ability of HTS to yield quantitative and unbiased information on transcript sequence abundance has afforded some remarkable new insights into transcriptome complexity and regulation. In this perspective, we primarily review recent results from applying HTS to the characterization of mammalian mRNA populations. Where relevant, we refer to recent microarray profiling studies and other approaches that have provided complementary insights. For a more detailed comparison of current HTS technologies and how these compare with microarray-based methods, we refer the reader to a recent review by Wang et al. (2009). A summary of additional applications of HTS technologies, including their previous use in profiling noncoding RNAs, can be found in Wilhelm and Landry (2009).

Deep surveying of mRNA processing complexity and regulation

Pre-mRNA transcripts undergo a series of modification and processing steps prior to the nuclear export of mature mRNA to the cytoplasm. As a pre-mRNA is synthesized by RNA polymerase II (pol II), it receives a 5′ m7G cap and, if it contains an intron, it is spliced by the spliceosome. All pol II transcripts are further processed at the 3′ end, in most cases by a tightly coupled cleavage and polyadenylation reaction. All of these steps are subject to regulation leading to transcript diversification. Differential promoter usage can lead to the formation of alternative 5′ ends, pairs of splice sites can be differentially used to generate transcript variants by alternative splicing, and different poly(A) sites can be selected. Additional processes such as RNA editing, in which individual bases are altered, can lead to further transcript diversity.

The realization that the transcriptomes of eukaryotes are highly complex and subject to regulation at multiple levels has spurred the ongoing development and application of experimental and computational approaches for the high-resolution characterization of transcriptome composition. Previously, analyses of mRNA populations employed alignments of expressed sequence tags (ESTs) and longer cDNAs to sequenced genomes. These afforded initial glimpses into the extent of alternative splicing and other forms of transcript processing complexity (Modrek and Lee 2003; Thanaraj et al. 2003; Zheng et al. 2005). However, these procedures were hampered by the lack of sufficient EST/cDNA coverage from individual cell and tissue types to yield significant information on the extent of regulated RNA processing events. Moreover, since ESTs are typically generated from the 5′ and 3′ ends of longer cDNA clones, detection of processing events was biased toward the ends of transcripts. The development of custom microarrays with probe sets designed to detect individual exons, or using combinations of probes specific for exon and splice junction sequences, overcame many of the obstacles encountered when analyzing EST/cDNA data and afforded a rapid means of profiling RNA expression and processing in different biological contexts (e.g., Clark et al. 2002; Johnson et al. 2003; Pan et al. 2004; for review, see Calarco et al. 2007; McKee and Silver 2007; Ben-Dov et al. 2008).

RNA-Seq represents the latest and most powerful tool with which to characterize transcriptomes (see Fig. 1 for an overview). An important first step in the analysis of RNA-Seq data is to identify reads that uniquely align to the genome and transcriptome, and remove reads that derive from nonunique sequences, including transcribed retrotransposons, duplicated or paralogous genes, and repetitive splice junction sequences. To date, there is no "gold standard" method for analyzing RNA-Seq data, and several laboratories have developed independent methods that score read alignments and counts in different ways. Initial methods and results from analyzing mammalian mRNA-Seq data were described by the Grimmond and Wold laboratories (Cloonan et al. 2008; Mortazavi et al. 2008). Mortazavi et al. (2008) analyzed mouse tissue transcriptomes using Illumina data sets consisting of 41–52 million 25-nt reads from three different mouse tissues. Mortazavi et al. (2008) identified unique mapping reads corresponding to ~17,000 previously unannotated regions of known genes. Many of these sequences appeared to correspond to extended 3′ or 5′ untranslated regions (UTRs), indicating that these genes contain longer and more variable UTR sequences than appreciated previously (Mortazavi et al. 2008). In addition, ~145,000 distinct splice junctions were detected (from a total of ~200,000 previously annotated junctions), and alternative splicing was detected in ~3500 genes.

In the parallel study by Cloonan et al. (2008), transcriptomic changes during differentiation of mouse embryonic stem cells (ESCs) and embryoid bodies (EB) were profiled using data sets comprising a total of ~100 million 25-nt reads generated using the AB SOLiD system (Cloonan et al. 2008). As in the Mortazavi et al. (2008) study, Cloonan et al. (2008) profiled whole transcript and splicing events by analyzing read alignments to exons and splice junction sequences. Approximately one-third of unique mapping reads were detected outside of known exons, and Cloonan et al. (2008) also reported a small number of candidate novel splicing events using their methods.

Shortly following publication of the above two studies, Sultan et al. (2008) reported analyses of Illumina RNA-Seq data obtained from the human embryonic kidney (HEK) 293 T and Ramos B-cell lines. Sultan et al. (2008)

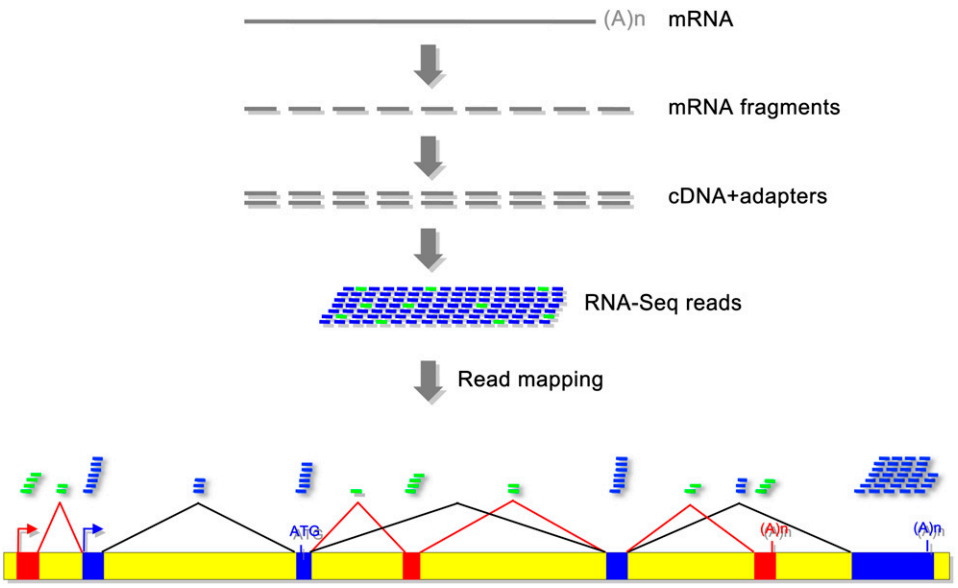

Figure 1. Schematic overview of the generation and analysis of RNA-Seq data. Poly(A)$^+$ mRNA is purified, fragmented, and then converted to a cDNA library with 5′ and 3′ adapter sequences. Short sequence reads are generated from the cDNA library. Reads are shown mapped to a hypothetical gene. Reads that map to previously annotated UTRs, exons, and splice junctions are shown in blue. Reads that map to novel expressed sequences, including alternative exons and corresponding splice junction sequences (indicated in red), are shown in green. (Forward arrows) Start sites of transcription; [(A)n] polyadenylation site.

also detected extended 5′ and 3′ UTR sequences. In addition, they determined the extent to which reads that do not match genomic sequence correspond to splice junction sequences. To address this, they scored read alignments to a synthetic set of splice junction sequences corresponding to all theoretically possible exon–exon junction combinations within annotated transcripts. Four-thousand-ninety-six putative novel junctions were detected in 3106 genes.

Subsequent RNA-Seq analyses focused on pre-mRNA processing events in human tissue and cell line transcriptomes. Using an Illumina-generated RNA-Seq data set consisting of a total of 400 million reads from 10 diverse human tissues and five mammary epithelial or breast cancer cell lines, Burge and colleagues (Wang et al. 2008) detected new exons and junctions by mapping sequence reads to a library of computationally predicted exons and splice junctions. This led to the detection of thousands of "high-confidence" (i.e., supported by two or more nonoverlapping reads) new candidate splice junctions (Wang et al. 2008). By plotting alternative splicing detection frequency as a function of increasing read coverage, it was estimated that alternative splicing occurs in 92%–96% of human genes, or ~98% or more of multiexon genes.

Using similar approaches, with the addition of a method that effectively discriminates true from false positive splice junction sequences, our group analyzed 17–32 million Illumina read data sets from six diverse human tissues (four of which were also used in the Burge [Wang et al. 2008] study) (Pan et al. 2008). New alternative splicing events were detected in ~20% of human genes and it was estimated that, overall, alternative splicing occurs in transcripts from 92%–97% of multiexon human genes with an average of approximately seven alternative splicing events per multiexon gene. The highest numbers of new alternative splicing events were detected in genes with the largest numbers of exons, including many "giant" genes such as Titin, Nebulin, and Obscurin, which are expressed in muscle tissue. However, the frequency of detection of alternative splicing per exon was found to be essentially independent of the number of exons per gene. Thus, despite the theoretical possibility of a quadratic (n^2) increase in the number of alternative splicing possibilities as the number of exons per gene increases, the number of alternative splicing events per gene was found to increase in a near linear fashion. This suggests that selection pressure may act to generally limit splicing complexity in large genes, an observation that could relate to earlier evidence (Lopez-Bigas et al. 2005) that genes with higher numbers of introns are statistically more often associated with human disease.

Cell-, tissue-, and individual-specific transcript variants

The RNA-Seq data sets generated from different mammalian cells and tissues have provided a rich source of data for the identification and characterization of regulated alternative splicing events. In the Wang et al. (2008) study, transcripts from 105,000 alternative splicing events mined from cDNA/EST data were analyzed for tissue-dependent variation. Expression of both isoforms was detected for more than one-third of these events in the 10 tissues analyzed and the majority of these appeared to display tissue-dependent variation in alternative

exon inclusion levels. When scoring an absolute inclusion ratio change of at least 10%, >22,000 tissue-dependent changes in alternative splicing were detected, which is substantially greater than the number of tissue-dependent differences in alternative splicing events detected using microarray profiling methods (Wang et al. 2008).

The analyses of these tissue-dependent events confirmed several previous observations obtained from microarray profiling experiments. For example, consistent with prior results (Xing and Lee 2005; Sugnet et al. 2006; Fagnani et al. 2007), Wang et al. (2008) found that human alternative splicing events that undergo the most pronounced tissue-dependent changes are significantly more often frame-preserving and flanked by evolutionarily conserved intronic sequences than are alternative exons that do not display pronounced tissue-dependent regulation. Together, these and other results from the analysis of RNA-Seq data have indicated that tissue-dependent alternative splicing events are more widespread than recognized previously (Pan et al. 2008; Wang et al. 2008), and are also more likely to have conserved functions compared with alternative splicing events that do not display such differential regulation.

An important consideration in the above RNA-Seq analyses was the extent to which the tissue-dependent variations in alternative splicing patterns may be associated with human individual-specific variation, particularly since some of the RNA-Seq data sets analyzed were derived from individuals. Previous studies based on exon tiling microarray experiments provided evidence that a subset of single-nucleotide polymorphisms (SNPs) located within exons or neighboring intronic sequences are associated with individual-specific variation in alternative splicing levels (Kwan et al. 2008; for review, see Graveley 2008). Using RNA-Seq data from cerebellar cortex tissue samples from six individuals, Wang et al. (2008) estimated that 10%–30% of alternative splicing events exhibited interindividual-specific variability, which is in agreement with a previous estimate of ~21% (Nembaware et al. 2004). Moreover, it was estimated that individual-specific variation in alternative splicing is twofold to threefold less frequent than tissue-dependent variation in alternative splicing (Wang et al. 2008).

As before, an important outcome of these results was the demonstration of the remarkable sensitivity and quantitative nature of RNA-Seq data as a means to detect alternative splicing variation. These results further indicate that RNA-Seq data will contribute a powerful source of data for linking individual- and population-specific genetic variation, as well as disease-associated mutations, to effects on transcript and processing levels. The increased power to detect such effects will in turn greatly facilitate linking transcriptome variation with disease and other phenotypic characteristics.

Linking RNA regulation with *trans*-acting factors

Between 0.5% and 1% of human genes contain one or more of the several types of canonical RNA-binding domains (RBDs) such as RNA recognition motifs (RRMs) and K homology (KH) domains (Clery et al. 2008). However, the majority of these RBD proteins have not been functionally characterized. Additionally, there are many genes containing other types of domains with possibly functionally important RNA-binding activities, such as specific classes of zinc-finger motifs. An important step toward understanding the functions of RBD proteins is to be able to accurately map their physiologically relevant binding sites. In addition to accelerating the elucidation of transcriptomic complexity, RNA-Seq is proving to be a powerful tool for connecting RNA-binding proteins to their target sites within regulated transcripts.

Several studies during the past 5 years set the stage for some of the observations stemming from recent RNA-Seq analyses that will be elaborated on below. In particular, microarray profiling, computational analyses of motifs, and large-scale RT–PCR assays have revealed sets of coregulated alternative splicing events, also referred to as "splicing regulatory networks" (SRNs). The regulated splicing events comprising these SRNs are often located in genes that are significantly enriched in common Gene Ontology (biological process and/or molecular function) annotations (for review, see Calarco et al. 2007; Ben-Dov et al. 2008; Moore and Silver 2008). The first example of such an SRN was described by Darnell and colleagues (Ule et al. 2005), who used microarrays combining probes specific for exons and splice junction sequences to profile RNA from the brains of wild-type mice and mice deleted for a KH-type RBD gene encoding Nova-2, a brain-specific alternative splicing regulator. Consistent with knowledge that Nova functions to regulate inhibitory synapse activity, genes containing Nova-2-regulated splicing events were found to be significantly enriched in functional annotations associated with synapse biology. Other SRNs have been identified by microarray-profiling alternative exons across diverse normal mouse tissues (Fagnani et al. 2007), during activation of a human T cell line (Ip et al. 2007), by depolarization of a human neuronal cell line (McKee et al. 2007), and more recently by profiling *Drosophila* cells following activation of the insulin response and wingless signaling pathways (Hartmann et al. 2009).

As in the case of Nova-2, coregulated sets of exons have also been detected by microarray or RT–PCR profiling of cells or tissues following RNAi, knockout, and/or overexpression of other alternative splicing regulators. Some examples include the widely expressed splicing repressor protein PTB (polypyrimidine tract-binding protein) and its neuronal paralog nPTB/brPTB (Boutz et al. 2007; Makeyev et al. 2007); human hnRNP proteins (Venables et al. 2008); the widely expressed TIA1/TIAR proteins (Aznarez et al. 2008); the muscle and neural-expressed Fox-1/2, MBNL1, and CUGBP1 proteins (Kalsotra et al. 2008; Zhang et al. 2008); and members of the *Drosophila* SR and hnRNP protein families (Blanchette et al. 2005; 2009).

Where studied, it was generally found that specific motifs corresponding to known binding sites of the targeted alternative splicing regulators are enriched in exons and/or flanking intron sequences of the coregulated alternative exons. For example, in the case of Nova, it was

found that clusters of the consensus (YCAY) Nova-1/2-binding sites concentrated in discrete zones near exon–intron boundaries are predictive of Nova-dependent alternative exon inclusion or exclusion (Ule et al. 2006). Similarly, it was found that the presence of Fox-1/2 consensus (U)GCAUG-binding sites downstream from a regulated exon correlated with increased Fox-1/2-dependent exon inclusion, whereas location of this binding in the upstream intron correlated with Fox-1/2-dependent exon skipping (Zhang et al. 2008).

HTS is beginning to contribute substantial new knowledge to the emerging landscapes of *cis*- and *trans*-acting factor-dependent global regulation of RNA processing. In a recent study from the Darnell laboratory (Licatalosi et al. 2008), HTS following in vivo cross-linking and immunoprecipitation (dubbed "HITS-CLIP" or "CLIP-Seq") was employed to provide a high-resolution map of Nova-2 binding in mouse neocortex tissue. Licatalosi et al. (2008) identified 168,632 unique CLIP-Seq tags from the Nova-2 immunoprecipitated RNAs, 73% of which mapped to known RNAs, and 27% of which mapped to intergenic regions. CLIP-Seq tags analyzed from two independent mice displayed remarkably similar mapping patterns and the immunoprecipitated tags exhibited a significant enrichment of YCAY-binding sites versus the nonimmunoprecipitated tags. By compiling data from 1085 CLIP-Seq tags identified from 71 Nova-2-regulated cassette exons and mapping these onto a "composite" pre-mRNA, Licatalosi et al. (2008) generated a map that demarcates distributions of Nova-binding sites that correlate with Nova-2-dependent inclusion or skipping of exons in the mouse brain, as detected by alternative splicing microarray profiling.

Similar approaches were reported recently by Yeo et al. (2009) to generate a Fox-2-binding map in human ESCs. This study revealed a network of Fox-2 targets that confirmed and extended predictions of Fox-1/2 position-dependent regulatory effects inferred from previous microarray and computational studies (Zhang et al. 2008).

Coordinated RNA processing events and regulatory factor multitasking

An emerging widespread property of RBD proteins is their ability to function in more than one step in the generation of mature mRNA transcripts. Several steps involved in transcription and RNA processing are tightly coupled and can influence one another (Maniatis and Reed 2002; Kornblihtt 2007; Moore and Proudfoot 2009), so it is perhaps not surprising to find that some regulatory factors can directly impact more than one step leading to mRNA translation. For example, specific SR family proteins originally identified as splicing factors have since been implicated in RNA pol II elongation, mRNA transport, and translation (for review, see Huang and Steitz 2005; Long and Caceres 2009). Many of these "multitasking" activities were discovered as a consequence of focused molecular and cell biological studies. However, RNA-Seq promises to greatly accelerate the discovery of "unexpected" functions for RBD proteins simply by virtue of its ability to provide an unbiased perspective of the RNA-binding target sites of a factor. For example, sequencing of CLIP tags immunoprecipitated by an antibody specific for the SR family protein SF2/ASF revealed candidate microRNA (miRNA)- and small nucleolar RNA (snoRNA)-binding targets, in addition to unspliced and spliced mRNA transcripts (Sanford et al. 2008, 2009). Similarly, sequencing of CLIP tags immunoprecipitated by an antibody to the widely acting splicing repressor hnRNPA1 revealed that this protein binds to the miRNA precursor pre-miR18a (Guil and Caceres 2007), and additional experiments confirmed that hnRNP-A1 is indeed important for the processing of this pre-miRNA.

Splicing and polyadenylation are closely coupled processes, and numerous studies have demonstrated that the binding of individual splicing factors to sequences proximal to poly(A) sites can influence 3′-end processing efficiency (for review, see Lutz 2008; Moore and Proudfoot 2009). Similarly, a subset of factors associated with mRNA 3′-end processing complexes influence the splicing of introns proximal to poly(A) sites. This cross-talk between the splicing and 3′-end formation machineries is important for the recognition of terminal exons and in some cases for the regulation of alternative poly(A) sites. Recent RNA-Seq- and microarray-based profiling studies have revealed that alternative poly(A) site selection, like regulated alternative splicing, is a far more common process than appreciated previously. Also emerging from these studies is evidence for a more extensive role for splicing regulators in the control of polyadenylation.

For example, in an interesting study from the Burge and Sharp groups (Sandberg et al. 2008) genome-wide exon tiling arrays were employed to profile alternative splicing and alternative poly(A) site selection during mouse T-lymphocyte differentiation. Remarkably, 86% of mapped alternative poly(A) sites in UTRs exhibited a directional shift resulting in shorter 3′ UTRs, coinciding with a late stage of differentiation following stimulation of resting T cells. Sandberg et al. (2008) further demonstrated that 3′ UTR shortening can provide a means of evading miRNA-mediated transcript degradation, thereby allowing for increased expression of specific genes at a late stage of T-cell differentiation.

In the aforementioned mRNA-Seq study by Burge and colleagues (Wang et al. 2008) mapping of sequence reads to alternative poly(A) sites revealed that differential alternative polyadenylation usage between tissues is even more frequent than different types of tissue variable alternative splicing. As for regulated "switch"-like alternative exons, increased sequence conservation was detected proximal to alternative polyadenylation sites, indicating an important role for these sequences in the regulation of alternative polyadenylation. To explore the linkages between alternative splicing and alternative polyadenylation, Wang et al. (2008) searched for common regulatory patterns and enrichment of hexanucleotide sequences adjacent to conserved alternative splicing and alternative polyadenylation events. A subset of the enriched motifs was common to the two types of RNA processing events. These motifs matched the consensus

sequences of known alternative splicing regulators, including Fox-1/2, CELF, MBNL, and members of the STAR family of RNA-binding factors.

These observations further emphasized that there may be widespread roles for alternative splicing regulators in the regulation of polyadenylation. Indeed, the previously mentioned Nova-2 CLIP-Seq study from the Darnell laboratory (Licatalosi et al. 2008) revealed that a subset of sequenced RNA tags clustered within a few hundred nucleotides of polyadenylation sites. Using exon tiling arrays to profile RNA from the brain tissues of wild-type and Nova-2 knockout mice, Licatalosi et al. (2008) found 297 changes involving alternative 3′ UTR sequences. RNase protection analysis on RNA from mouse brains confirmed a role for Nova-2-dependent alternative poly(A) site selection for a few of these genes.

The examples summarized above further illustrate the power of HTS in revealing important new biology. The combination of RNA-Seq profiling of transcriptomes with CLIP-Seq in particular will undoubtedly yield a wealth of interesting and important new information on how specific RNA-binding proteins function to coordinate different aspects of RNA biogenesis and regulation.

Current and future challenges in the emerging transcriptomics era

An ultimate goal of high-throughput and systems-based approaches for studying gene regulation is to be able to derive a "unified" network that encompasses all gene expression regulatory steps. Information provided by such a unified network should help to explain how different steps in gene regulation communicate with one another and respond to intracellular and extracellular signals and perturbations. Different subsets of genes are regulated at the transcriptional and RNA processing levels to achieve cell/tissue- and condition-specific gene expression programs (Keene and Lager 2005; Blencowe 2006; McKee and Silver 2007). However, as mentioned above, it is also known that multiple steps in the synthesis and processing of RNA transcripts are coupled and can influence one another (Maniatis and Reed 2002; Kornblihtt 2007; Moore and Proudfoot 2009). These statements are not conflicting, since accumulating evidence suggests that physical coupling mechanisms may generally serve to temporally coordinate and enhance the kinetics of individual steps in transcription and processing in a cell/tissue- or condition-independent fashion, although there are emerging exceptions in which condition-dependent changes in transcription and alternative splicing can involve significantly overlapping subsets of genes (Hartmann et al. 2009). Nevertheless, how multiple steps in transcription and RNA processing are coordinated to achieve a concerted cell/tissue type-dependent physiological outcome, whether involving coupling mechanisms that require direct physical interactions or not, is not well understood. Based on the initial results from HTS and complementary microarray-based studies described above, it appears that data sets with the quantity and quality of measurements necessary to accurately model multilayer transcript-level regulatory networks should soon be available.

However, there is a definite need for improvement in the current HTS technology to facilitate major advances involving integrated analyses. As mentioned earlier, current generation systems for producing RNA-Seq data are limited in that they do not yet provide an efficient means with which to comprehensively define the full complement of transcript isoforms in an RNA sample. This is in part because short-read profiling does not reveal the colinear structures of transcripts. An important step toward resolving this limitation is the use of "paired-end" (PE) sequencing, in which sequences at 5′ and 3′ ends of the same cDNA molecule can be determined. Sequencing of PEs with variable intervening distances should facilitate defining how multiple transcript features, including specific start and stop sites and exon/splice site combinations, are combined within individual transcripts. Moreover, systems currently in development are expected to permit single-molecule sequencing, and the use of such systems may provide a solution for this "transcript assembly" problem if sufficiently long read lengths can be obtained.

The other ongoing issue in using HTS methods is the depth of coverage. As mentioned above, the current maximal output of single-read RNA-Seq data from a single 2- to 3-d run is only sufficient to accurately quantify splicing levels for ~30% of expressed transcripts in a typical mRNA sample from a mammalian cell line or tissue. Accurate measurements of exon inclusion levels require ~20 or more reads that map specifically to the three junction sequences specifying inclusion or skipping of an alternative exon; achieving this density of coverage requires on average ~400 reads (at ~35 nt) per kilobase (Pan et al. 2008). Reaching this level of coverage by shotgun sequencing of mRNA samples is a diminishing returns situation, and an estimated ~700 million reads would be required to obtain accurate quantification of >95% of expressed transcripts (see Fig. 2). A possible solution to this problem is to generate pools of primers directed to specific transcript regions of interest, such that focused cDNA libraries can be sequenced. Further developments along these lines will be necessary before HTS methods, at the current levels of read output, afford more comprehensive profiling of transcriptome complexity and regulation.

Finally, the enormous complexity of transcript variants already revealed by RNA-Seq and other transcriptome profiling methods begs the question as to what extent the transcript variants generated by different RNA processing steps are functionally significant. Judging from the range of conservation levels of sequences surrounding these events, one can expect a spectrum of functional importance ranging from essential for viability to neutral activity and potential fodder for evolutionary adaptation. As an initial step to addressing questions regarding function, one can ask which transcripts are most likely to be translated. A very recent advance in this direction is the employment of RNA-Seq to characterize yeast mRNA sequences that are bound and protected by polyribosomes

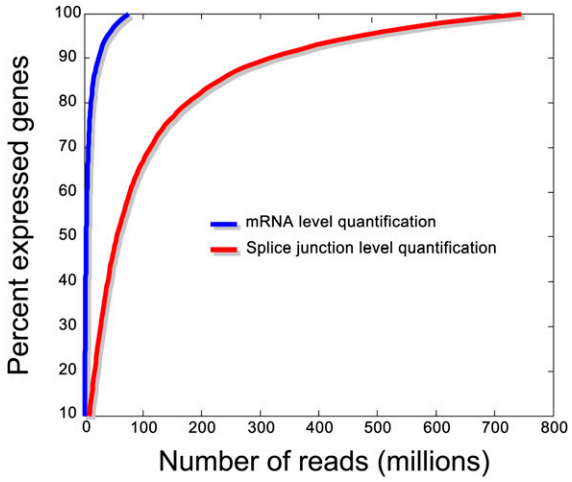

Figure 2. Plots showing the percent of genes for which accurate quantitative measurements of mRNA levels and individual splicing levels can be obtained, when analyzing different numbers of reads (at ~35 nt per read) in a representative human cell line RNA-Seq data set. Notably, <10 million reads is necessary to accurately quantify mRNA expression levels for >80% of genes, whereas the accurate quantification of splicing levels for 80% of genes would require ~200 million reads.

recovered by affinity purification (Ingolia et al. 2009). In principle, this method should be applicable to the characterization of the translated mRNA populations from any type of cell or tissue source. Comparisons of recent RNA-Seq data sets with peptide mass spectrometry data from the same sources should also yield information on the extent of translation of transcript variants. The time has clearly come to employ these and other profiling approaches, together with high-throughput screening methods linked to bioassays, to systematically address the specific roles of the myriad of transcript variants.

Acknowledgments

We thank John Calarco, Cori Hanson, Jim Ingles, Mathieu Gabut, Deb Ray, and Arneet Saltzman for helpful comments on the manuscript. Our research is supported by grants from the Canadian Institutes of Health Research, National Cancer Institute of Canada, and from Genome Canada through the Ontario Genomics Institute.

References

Aznarez I, Barash Y, Shai O, He D, Zielenski J, Tsui LC, Parkinson J, Frey BJ, Rommens JM, Blencowe BJ. 2008. A systematic analysis of intronic sequences downstream of 5′ splice sites reveals a widespread role for U-rich motifs and TIA1/TIAL1 proteins in alternative splicing regulation. *Genome Res* **18:** 1247–1258.

Ben-Dov C, Hartmann B, Lundgren J, Valcarcel J. 2008. Genome-wide analysis of alternative pre-mRNA splicing. *J Biol Chem* **283:** 1229–1233.

Blanchette M, Green RE, Brenner SE, Rio DC. 2005. Global analysis of positive and negative pre-mRNA splicing regulators in *Drosophila*. *Genes & Dev* **19:** 1306–1314.

Blanchette M, Green RE, MacArthur S, Brooks AN, Brenner SE, Eisen MB, Rio DC. 2009. Genome-wide analysis of alternative pre-mRNA splicing and RNA-binding specificities of the *Drosophila* hnRNP A/B family members. *Mol Cell* **33:** 438–449.

Blencowe BJ. 2006. Alternative splicing: New insights from global analyses. *Cell* **126:** 37–47.

Boutz PL, Stoilov P, Li Q, Lin CH, Chawla G, Ostrow K, Shiue L, Ares M Jr, Black DL. 2007. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes & Dev* **21:** 1636–1652.

Calarco JA, Saltzman AL, Ip JY, Blencowe BJ. 2007. Technologies for the global discovery and analysis of alternative splicing. *Adv Exp Med Biol* **623:** 64–84.

Clark TA, Sugnet CW, Ares M Jr. 2002. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* **296:** 907–910.

Clery A, Blatter M, Allain FH. 2008. RNA recognition motifs: Boring? Not quite. *Curr Opin Struct Biol* **18:** 290–298.

Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5:** 613–619.

Fagnani M, Barash Y, Ip JY, Misquitta C, Pan Q, Saltzman AL, Shai O, Lee L, Rozenhek A, Mohammad N, et al. 2007. Functional coordination of alternative splicing in the mammalian central nervous system. *Genome Biol* **8:** R108. doi: 10.1186/gb-2007-8-6-r108.

Graveley BR. 2008. The haplo-spliceo-transcriptome: Common variations in alternative splicing in the human population. *Trends Genet* **24:** 5–7.

Guil S, Caceres JF. 2007. The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. *Nat Struct Mol Biol* **14:** 591–596.

Hartmann B, Castelo R, Blanchette M, Boue S, Rio DC, Valcarcel J. 2009. Global analysis of alternative splicing regulation by insulin and wingless signalling in *Drosophila* cells. *Genome Biol* **10:** R11. doi: 10.1186/gb-2009-10-1-r11.

Huang Y, Steitz JA. 2005. SRprises along a messenger's journey. *Mol Cell* **17:** 613–615.

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324:** 218–223.

Ip JY, Tong A, Pan Q, Topp JD, Blencowe BJ, Lynch KW. 2007. Global analysis of alternative splicing during T-cell activation. *RNA* **13:** 563–572.

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302:** 2141–2144.

Kalsotra A, Xiao X, Ward AJ, Castle JC, Johnson JM, Burge CB, Cooper TA. 2008. A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. *Proc Natl Acad Sci* **105:** 20333–20338.

Keene JD, Lager PJ. 2005. Post-transcriptional operons and regulons co-ordinating gene expression. *Chromosome Res* **13:** 327–337.

Kornblihtt AR. 2007. Coupling transcription and alternative splicing. *Adv Exp Med Biol* **623:** 175–189.

Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. 2008. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* **40:** 225–231.

Li H, Lovci MT, Kwon YS, Rosenfeld MG, Fu XD, Yeo GW. 2008. Determination of tag density required for digital transcriptome analysis: Application to an androgen-sensitive prostate cancer model. *Proc Natl Acad Sci* **105:** 20179–20184.

Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, et al. 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456:** 464–469.

Long JC, Caceres JF. 2009. The SR protein family of splicing factors: Master regulators of gene expression. *Biochem J* **417:** 15–27.

Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo R. 2005. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett* **579:** 1900–1903.

Lutz CS. 2008. Alternative polyadenylation: A twist on mRNA 3′ end formation. *ACS Chem Biol* **3:** 609–617.

Makeyev EV, Zhang J, Carrasco MA, Maniatis T. 2007. The microRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol Cell* **27:** 435–448.

Maniatis T, Reed R. 2002. An extensive network of coupling among gene expression machines. *Nature* **416:** 499–506.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18:** 1509–1517.

McKee AE, Silver PA. 2007. Systems perspectives on mRNA processing. *Cell Res* **17:** 581–590.

McKee AE, Neretti N, Carvalho LE, Meyer CA, Fox EA, Brodsky AS, Silver PA. 2007. Exon expression profiling reveals stimulus-mediated exon use in neural cells. *Genome Biol* **8:** R159. doi: 10.1186/gb-2007-8-8-r159.

Modrek B, Lee CJ. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet* **34:** 177–180.

Moore MJ, Proudfoot NJ. 2009. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* **136:** 688–700.

Moore MJ, Silver PA. 2008. Global analysis of mRNA splicing. *RNA* **14:** 197–203.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628.

Nembaware V, Wolfe KH, Bettoni F, Kelso J, Seoighe C. 2004. Allele-specific transcript isoforms in human. *FEBS Lett* **577:** 233–238.

Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, Babak T, Siu H, Hughes TR, Morris QD, et al. 2004. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell* **16:** 929–941.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40:** 1413–1415.

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* **320:** 1643–1647.

Sanford JR, Coutinho P, Hackett JA, Wang X, Ranahan W, Caceres JF. 2008. Identification of nuclear and cytoplasmic mRNA targets for the shuttling protein SF2/ASF. *PLoS One* **3:** e3369. doi: 10.1371/journal.pone.0003369.

Sanford JR, Wang X, Mort M, Vanduyn N, Cooper DN, Mooney SD, Edenberg HJ, Liu Y. 2009. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res* **19:** 381–394.

Sugnet CW, Srinivasan K, Clark TA, O'Brien G, Cline MS, Wang H, Williams A, Kulp D, Blume JE, Haussler D, et al. 2006. Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Comput Biol* **2:** e4. doi: 10.1371/journal.pcbi.0020004.

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321:** 956–960.

Thanaraj TA, Clark F, Muilu J. 2003. Conservation of human alternative splice events in mouse. *Nucleic Acids Res* **31:** 2544–2552.

Ule J, Ule A, Spencer J, Williams A, Hu JS, Cline M, Wang H, Clark T, Fraser C, Ruggiu M, et al. 2005. Nova regulates brain-specific splicing to shape the synapse. *Nat Genet* **37:** 844–852.

Ule J, Stefani G, Mele A, Ruggiu M, Wang X, Taneri B, Gaasterland T, Blencowe BJ, Darnell RB. 2006. An RNA map predicting Nova-dependent splicing regulation. *Nature* **444:** 580–586.

Venables JP, Koh CS, Froehlich U, Lapointe E, Couture S, Inkel L, Bramard A, Paquet ER, Watier V, Durand M, et al. 2008. Multiple and specific mRNA processing targets for the major human hnRNP proteins. *Mol Cell Biol* **28:** 6033–6043.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* **10:** 57–63.

Wilhelm BT, Landry JR. 2009. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* doi: 10.1016/j.jmeth.2009.03.016.

Xing Y, Lee CJ. 2005. Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS Genet* **1:** e34. doi: 10.1371/journal.pgen.0010034.

Yeo GW, Coufal NG, Liang TY, Peng GE, Fu XD, Gage FH. 2009. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol* **16:** 130–137.

Zhang C, Zhang Z, Castle J, Sun S, Johnson J, Krainer AR, Zhang MQ. 2008. Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. *Genes & Dev* **22:** 2550–2563.

Zheng CL, Kwon YS, Li HR, Zhang K, Coutinho-Mansfield G, Yang C, Nair TM, Gribskov M, Fu XD. 2005. MAASE: An alternative splicing database designed for supporting splicing microarray applications. *RNA* **11:** 1767–1776.

Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes

Benjamin J. Blencowe, Sidrah Ahmad and Leo J. Lee

Genes Dev. 2009, **23:**
Access the most recent version at doi:10.1101/gad.1788009