Reflections for the 20th anniversary issue of *RNA* journal

BENJAMIN J. BLENCOWE
Donnelly Centre and Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada M5S 3E1

The 20 years since the launch of the *RNA* journal have seen numerous remarkable advances in RNA research. Those that have had a transformative impact on biomedical research as a whole include the discovery of large repertoires of short and long non-coding RNAs and their widespread roles in normal and disease biology, and the development of powerful methods for manipulating the expression and function of individual genes, based in part on the former discoveries, as well as the more recent development of RNA-guided genome editing tools.

In parallel with these discoveries, and similarly impacting our understanding of essentially every biological process and pathway, has been the illumination of entire landscapes of RNA processing events through the development and application of high-throughput profiling technologies. This work has led to a fundamental rethinking of the nature of biological complexity and diversity, as well as the organization, regulation, function and evolution of biological networks. Here I reflect on some of the key developments in the burgeoning field of transcriptomics during the past ~15 years, while also considering future directions in this area of research.

Twenty years ago, it was widely assumed that differences in biological complexity and phenotypic attributes of species would relate to the numbers and types of genes that they possess. It was also suggested that only a small fraction of human genes generate multiple transcript isoforms by processes such as alternative splicing (AS). These views, however, radically changed around the year 2000, following the completion of (draft) genome sequences of *C. elegans*, *D. melanogaster* and *H. sapiens*. These species were shown to have comparable numbers and repertoires of protein coding genes. This landmark observation shifted attention towards understanding the nature of gene regulatory differences between species. It was proposed by Eric Lander and colleagues in their 2000 human genome sequence paper that AS might provide a major mechanism for the generation of biological and regulatory diversity in species.

Shortly after joining the faculty at the University of Toronto in 1998, I set my group's sights on tackling the following questions: (1) How do splicing patterns differ between cell types, cell states (including disease versus normal) and species? (2) Which combinations of cis-regulatory sequences form a "splicing code" that governs regulated splicing decisions? (3) Which AS events are functionally important and what are their specific biological roles? The publication of the 2000 human genome sequencing paper not only emphasized the importance of these questions, but also the need for a new generation of tools capable of tackling them. This challenge became a major focus for several laboratories, including my own.

In the early 2000's, expressed sequenced tag (EST) sequencing data afforded a first glimpse into the extent of human transcriptomic complexity. Using alignments of ESTs to the draft human genome sequence, Christopher Lee and colleagues (UCLA) reported that ~40% of human genes produce transcripts that are subject to AS. However, the low coverage afforded by EST and longer cDNA sequences did not allow researchers to systematically detect—or reliably quantify—AS or other RNA processing events. Efforts were therefore initiated to establish custom microarrays and associated computational tools to address this challenge. Between 2002 and 2005, the first systems and resulting insights were reported, including work from Manny Ares (UCSC), Xiang-Dong Fu (UCSD), Jason Johnson (Rosetta Inpharmatics, later Merck), Robert Darnell (Rockefeller University, in collaboration with Affymetrix), and my own laboratory.

By developing a quantitative AS profiling system, we discovered thousands of previously unknown mammalian tissue-regulated AS events. The genes harboring these events by and large did not overlap genes differentially regulated between the same tissues at the transcriptional level. By analyzing RNA from mice deficient of the neuronal RNA binding protein Nova, Darnell's group detected Nova-dependent alternative exons that are enriched in genes functionally associated with the inhibitory synapse and axon guidance. These and numerous subsequent studies revealed that tissue-dependent AS represents a distinct "layer" of gene regulation, and that regulated exons are typically organized into functionally coordinated, biologically coherent "networks" that act orthogonally to other gene regulatory networks. These findings extended the concept of "RNP operons"

Corresponding author: b.blencowe@utoronto.ca
Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.051003.115. Freely available online through the *RNA* Open Access option.

proposed a few years before by Jack Keene and colleagues (Duke University).

AS microarrays were soon generating sufficient quantities of data with which to train sophisticated machine learning algorithms capable of inferring the combinations of cis-regulatory elements that govern regulated splicing. In 2010, in collaboration with Brendan Frey and colleagues at the University of Toronto, we described a predictive code for mammalian tissue-dependent splicing, and Darnell and colleagues integrated various data types to infer an expanded Nova-regulated exon network. These studies defined important features of the splicing code responsible for differentially regulating AS, and they also enabled the discovery and characterization of new mechanisms and biological functions of splicing.

Despite these advances, it was clear that splicing-sensitive microarrays were limited in terms of detection coverage, specificity and sensitivity. Fortunately, these limitations would soon be overcome by the arrival of high-throughput RNA sequencing (RNA-Seq) technologies. Chris Burge's laboratory at MIT and my group, were fortunate to gain access to the first human tissue RNA-Seq datasets from Gary Schroth of Illumina Inc. Our groups set about developing custom analysis pipelines and in late 2008 we simultaneously published results showing that more than 95% of multi-exon human genes produce alternatively spliced transcripts. However, to what extent splicing patterns vary between species, and the extent to which the massive repertoires of detected splice variants might be functionally relevant remained important open questions.

By comparing RNA-Seq AS profiles of several common organs across a diverse range of vertebrate species, we observed that exon skipping rates are significantly higher in primate organs than in the equivalent organs of species such as mouse, chicken and frog. Moreover, it was observed that vertebrate AS profiles have diverged so rapidly, that they are more similar between different organs within a species than they are between the equivalent organs from different species. In stark contrast, differential gene expression patterns have evolved much more slowly, such that they are more similar between equivalent organs from different species than between different organs from the same species. Collectively, these observations, which were published back to back with a related study from Burge and colleagues in 2012, provided strong evidence in support of predictions from the 2000 human genome sequencing paper, namely that AS has the potential to contribute in a major way to the enormous range of biological complexity and phenotypic diversity among vertebrates and other species.

Despite the extensive lineage and species-specificity of AS, cell or tissue-dependent AS events, which represent 10%–30% (depending on the tissue type) of all AS events, tend to be evolutionarily conserved. However, these events are significantly underrepresented in terms of their overlap with functionally defined, modular protein domains. While there are striking examples of such events, for example those that alter DNA binding domains of transcription factors to control cell fate decisions, tissue-regulated exons overlapping coding regions are highly enriched in intrinsically disordered regions of proteins. These regions typically reside on protein surfaces and are known to mediate ligand interactions. To investigate the functional significance of this feature of alternative exons, in collaboration with Jeffrey Wrana and colleagues (Mt. Sinai Hospital and University of Toronto), we employed a high-throughput protein–protein interaction assay to screen neural exons for functions in controlling protein–protein interactions. Remarkably, approximately one third of the analyzed exons were found to positively or negatively affect one or more partner interactions, including an interaction that is important for efficient neural cell endocytosis. Collectively, these and other studies have provided strong evidence that networks of tissue co-regulated alternative exons provide a powerful mechanism for remodeling protein-interaction networks required to establish and maintain cell type.

Unbiased transcriptome profiling by RNA-Seq coupled with the appropriate computational tools can yield entirely unanticipated and surprising findings and contribute important functional insight. One such example is the recent illumination of 3–27 nt neuronal-spliced "microexons," by Manuel Irimia and others in my group, as well as by Chris Ponting and colleagues (University of Oxford). This class of under-appreciated AS is particularly striking: it is more highly conserved than any other class defined to date, it displays the most dynamic degree of differential splicing during the differentiation and maturation of neurons, and, unlike longer alternative exons, it has the propensity to overlap and functionally alter modular interaction domains in proteins. Remarkably, microexons are frequently misregulated in the brains of individuals with autism spectrum disorder, and they are also highly enriched in genes with genetic and functional links to autism. An exciting picture thus emerges in which misregulated microexons (as well as longer exons) alter protein interaction networks associated with the proper functioning of neurons, in ways that may cause or contribute to the disorder. These studies thus exemplify the immense power of unbiased profiling methods in providing new insight into transcriptomic diversity and function, and also in illuminating pathways that may be commonly altered in complex diseases and disorders such as autism. It is remarkable how far such technologies have taken us in the past two decades. Nevertheless, it is likely that we are only scratching the surface in terms of understanding the full scope and significance of transcript isoform diversity and regulation.

Research in this area will advance through the continued development and application of new technologies that are geared for the high-throughput interrogation of transcript composition, regulation and function. For example, there is a need for improvement in methods for long-range, deep-coverage sequencing to determine the linear exon composition of full-length transcripts. There is also a need for

experimental methods that can rapidly link cis- and trans-acting regulatory factors to specific AS and other RNA processing events. Technologies permitting genome-wide knockdowns and CRISPR-mediated editing, coupled to barcoded sequencing of transcripts, such as the powerful method recently described by Juan Valcarcel and colleagues (CRG, Barcelona), will be particularly useful in this regard. More systematic profiling of relatively poorly understood aspects of transcript regulation, including the role of RNA modifications, RNA structure, and RNA–RNA interactions, will also be particularly important for filling gaps in our understanding of AS regulation.

In this reflection piece, I have primarily touched on developments in the "transcriptomics revolution" during the past ~15 years that have shed light on the scope and biological relevance of splice isoform diversity. However, when taking stock of developments in the "-omics" revolution as a whole, it is apparent that we will soon be able to model ways in which complex molecular systems are integrated and communicate with each other to control aspects of cell and organism behavior. A wealth of data is being generated that provide quantitative snapshots of each of the major steps in the gene expression pathway across diverse cell and tissue types, perturbations and developmental transitions, as well as from global-scale measurements for the functions of the multitude of cis- and trans-acting regulatory factors. The prospect of harnessing such datasets to model multi-dimensional networks in ways that illuminate new mechanisms and predict phenotypic outcomes seems very promising. Such advances will also undoubtedly greatly benefit our understanding of the mechanisms that contribute to and cause human diseases, thereby ultimately leading to more informed approaches to the development of new diagnostic and therapeutic strategies.