Finding the rules of splicing, and using them … alternatively

BENOIT CHABOT

Department of Microbiology and Infectious Diseases, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, Quebec, Canada J1K 2R1

The last 20 years have witnessed a convergence of efforts aimed at understanding how genes produce biological complexity. In 2001, the discovery that only 25,000 genes were required to produce a human being came as a surprise because less complex organisms such as worms and flies have a similar number of genes. Recalling that the first splicing events to be documented involved the use of alternative splice sites led some to anticipate the contribution of alternative splicing to biological complexity. Although we have now reached the upper limit of close to 95%, in terms of the number of genes using alternative splicing, one exciting challenge ahead remains to describe this diversity when associated with processes such as organismal and organ development using, for example, 3D model systems. Another critical task is to continue describing the role of alternative splicing in disease and aging, and how the production of splice variants is modulated by our environment, including our microbiomes. A parallel challenge that pervades all the above tasks is to develop tools that can assess in a systematic manner the functional attributes of this constellation of splice variants.

To document the existence and importance of alternative splicing is not enough; RNA researchers want to understand the underlying molecular mechanisms that regulate splice site selection. Building on the regulatory principles enunciated by pioneering studies on sex-specific splicing in *Drosophila*, the search for regulatory rules has been one of the busiest roads of the last two decades. Sequence elements and factors regulating alternative splicing were rapidly categorized as activators and repressors, but the elegant maps linking the binding positions with the activity of these regulators (pioneered by the group of Bob Darnell in New York, USA) revealed that regulators can impact splicing positively or negatively depending on where they bind on a pre-mRNA. The sequences targeted by regulators caught in "flagrant délit" of interaction also appear to be weaker than the strong binding sites typically identified by in vitro selection assays. Thus, like the splice sites of regulated exons, the binding sites for regulatory factors are suboptimal, possibly offering a wide dynamic range and a robust response to small changes in the activity of the regulators.

Great strides have been made in our understanding of the principles regulating splice site selection. Although most cases involve elements and factors bound in the immediate vicinity of the regulated splice sites, additional molecular strategies may soon include many more examples of longer range interactions mediated by proteins, and possibly stabilized by higher order structure of the pre-mRNA, or vice-versa. Even RNA binding proteins (RBPs) bound at distant sites on a pre-mRNA may engage in homo- or heterotypic interactions that can potentially modulate splice site selection by bringing distant splice sites in closer proximity, thus facilitating their commitment, or by looping out splice sites to repress their use.

A detailed understanding of the molecular mechanisms regulating splice site selection is currently limited to a handful of cases out of the 30 or so well known regulators that belong to the broad families of SR and hnRNP proteins. One recent surprise was the finding that mutations in splicing factors categorized as "constitutive" or "generic" could also impact splice site selection. However, such surprises may be expected when our rules for constitutive splicing are derived from the use of nonconstitutively expressed pre-mRNAs (most commonly β-globin or an adenoviral pre-mRNA). Nevertheless, it remains to be seen if controlling the activity of generic splicing factors is a cellular strategy used to regulate splice site selection. The group of Thomas Tuschl (New York, USA) has estimated at 1500 the number of human RBPs, not counting their splice variants, which may display different activities. Because the biological functions of at least one-third of these RBPs are unknown or were inferred from homologs, we can anticipate that many more RBPs will be implicated in splicing control. Although our knowledged of the diversity of these molecular mechanisms is guaranteed to expand, let us hope that this quest will be more rapid than in other fields; in 2011, 10 years after the human genome was sequenced, more than 75% of protein research focused on the 10% of proteins that were known before the genome was sequenced (based on an analysis by Aled Edwards in Toronto, Canada).

Because RBPs shuttle between different cellular compartments and interact with different RNA molecules, they are prime candidates to integrate pre-mRNA splicing with other layers of gene expression, including signaling and transcription. In some instances, they may even be drafted to participate in apparently unrelated processes (for example, hnRNP A1 in telomere biogenesis, as shown by our group). Based on recent studies in yeast and viral systems, another emerging feature of RBP function may be that these proteins are themselves controlled by increasing or decreasing levels of transcripts, not unlike the recently documented sponging effects of noncoding RNAs on miRNA function.

Bioinformatics has been an ever-increasing ally in the quest to compile and understand the complexity of transcriptomes, produce splicing maps for RBPs, and anticipate the impact of genetic polymorphisms on splicing profiles. Following the initial success of using complex sets of known regulatory elements to identify tissue-specific regulatory programs (Ben Blencowe and Brendan Frey groups in Toronto, Canada), future efforts aimed at uncovering the splicing code will undoubtedly evolve to integrate data from transcription elongation speed, signaling input, levels of RBPs and transcripts that compete for their binding.

Projecting what the field will look like 20 years from now is an uncomfortable exercise, in part because our funding agencies have trained us not to plan more than five years in advance. Nevertheless, one easy first prediction of the direction that splicing research may take emerges from the spectacular development of CRISPR-Cas9 technologies that allow manipulation of endogenous genes. The ability to modify the mammalian genome may also evolve rapidly from RNA-mediated recombination recently documented in yeast (group of Francesca Storici in Atlanta, USA). Applying these approaches to the analysis of splicing regulation will help validate rules that were derived mostly from the use of reporter genes, and will offer unprecedented means to address how regulation of splice site selection is intertwined with transcription and chromatin modifying processes in the natural environment of genes.

Another safe bet deals with the importance of splicing decisions in disease and our need to devise tools to correct the production of aberrant and defective splice variants. Promising results are already emerging from the use of antisense oligonucleotides and small molecules. Another potential source of progress in this direction is provided by the evidence showing that cooperativity and synergy between distinct and relatively low affinity RBPs can build specificity. This principle could be exploited once small molecules that inhibit individual RBPs are identified. For example, using a mixture of compounds, each at a concentration that has a weak effect on the target RBP and with limited off-target impact, may still, in combination, prevent synergy, bestow specificity of action for one or a small set of splicing events, and remain innocuous.

Another challenging area deals with repetitive or repeat-derived sequences which constitute more than 50% of the human genome. The current limitations of RNA sequencing technologies have prevented an assessment of the contribution of most repetitive elements, including the highly polymorphic minisatellites (VNTRs), to the function of splice variants and to biological complexity. The prevalence of such repetitive elements in genes and their important contribution to the diversity of transcriptomes (our work) warrant a closer look. Moreover, the intrinsically unstable nature of these repeats in the human population may have, and may still contribute, to produce splice forms endowed with novel functions that drive organismal evolution.

One of the largest global efforts in the upcoming decades will be to understand brain function, a goal that will demand a comprehensive investigation of the role of alternative splicing. The brain is the organ in which alternative splicing is the most prevalent; however, inheritable changes that have sculpted brain-specific functions must have occurred first in germ cells. Interestingly, alternative splicing is also prevalent in testis, and gene expression profiles exhibited by brain, cerebellum, and testis are most similar to each other as compared with other tissues. Moreover, changes in the transcriptome through evolution were most rapid in testis (group of Henrik Kaessman in Lausanne, Switzerland). One pathway that may produce genomic change in germ cells is transcription-associated genomic instability, a process driven by R-loops that is stimulated when the level of RBPs drops (work pioneered by the group of Andrés Aguilera in Sevilla, Spain). If relevant segments of a pre-mRNA emerging from a transcribing polymerase are not appropriately covered by RBPs during testis development (either because the expression or binding activity of a RBP is low or because of faster speed of transcription), then these RBP-free regions may trigger the formation of R-loops. These single-stranded DNA loops would in turn become substrates for mutations at or near the binding sites of regulatory RBPs, with the likelihood of permanently impacting future alternative splicing events at these locations. The strong overlap between testis and the brain transcriptomes argues in favor of the notion that many mutations created by the transcription-associated pathway in germ cells may affect alternative splicing in the brain, with a potential impact on function. If this process occurs in genes already carrying intrinsically unstable VNTR in coding regions, it may produce an exceptional rate of mutation that may have stimulated brain-specific alternative splicing and may still be molding brain function. Experimental systems tailored to study germ cell and brain development are required to test these speculations.

The last two decades have set the stage for recognizing the immense contributions that the vast collections of splice variants make to biological complexity. I look forward to seeing, in 2035, how far from the mark were my predictions, and how many surprising and exciting directions the field will have taken. While I am eagerly marking my calendar for that date, I will certainly check back regularly.