Cell Research (2010) 20:866-868.

www.nature.com/cr

RESEARCH HIGHLIGHT

Embracing the complexity of pre-mRNA splicing

Peter J Shepard[1], Klemens J Hertel[1]

[1]*Department of Microbiology & Molecular Genetics, University of California, Irvine, Irvine, CA 92697-4025, USA*
Cell Research (2010) **20**:866-868. doi:10.1038/cr.2010.98; published online 6 July 2010

Pre-mRNA splicing is a fundamental process required for the expression of most metazoan genes. It is carried out by the spliceosome, which catalyzes the removal of non-coding intronic sequences to assemble exons into mature mRNAs prior to export and translation. Defects in splicing lead to many human genetic diseases [1], and splicing mutations in a number of genes involved in growth control have been implicated in multiple types of cancer [2]. Given the complexity of higher eukaryotic genes and the relatively low level of splice-site conservation, the precision of the splicing machinery in recognizing and pairing splice sites is remarkable. Introns ranging in size from less than 100 up to 100 000 bases are removed efficiently. At the same time, a large number of alternative splicing events are observed between different cell types, developmental stages, and during other biological processes. Of the approximately 25 000 genes encoded by the human genome [3], more than 90% are believed to produce transcripts that are alternatively spliced [4]. Thus, alternative splicing of pre-mRNAs can lead to the production of multiple protein isoforms from a single pre-mRNA, significantly enriching the proteomic diversity of higher eukaryotic organisms [5]. Because regulation of this process can determine the timing and location in which a particular protein isoform is produced, changes in alternative splicing patterns modulate many cellular activities. This extensive alternative splicing implies a significant flexibility of the spliceosome to identify and process exons within a given pre-mRNA.

Over the last few years, research from different laboratories has demonstrated that the regulation of alternative splicing depends on many different *cis*-acting RNA elements [6]. While most of these RNA elements appeared to have a common mechanistic goal, the initial recruitment of spliceosomal components, the sheer number of RNA sequences involved in regulating alternative splicing became more and more daunting. As a consequence, the enthusiasm to derive alternative splicing predictions based only on pre-mRNA sequence, appropriately referred to as the "splicing code", gave way to the growing realization that alternative splicing is much more complex than initially anticipated. Computational analyses of the identified RNA elements proved to be the first efficient tools to classify and group these RNA elements, while at the same time expanding the repertoire of additional RNA sequence elements likely to be important for splicing regulation [7]. Sequence motifs that define the exon/intron boundary, the presence or absence of splicing enhancer or silencer sequences, RNA secondary structures, and the length of exons and their flanking introns are the most prominent classes of RNA elements involved in mediating efficient exon splicing [6]. While it was appreciated that the relative contribution of each of these RNA sequence elements controls how efficient splice sites are recognized and flanking introns are removed, efforts to describe how their relative contributions are weighted in different cell types were limited in success due to insufficient splicing profiles from unique tissues. The "splicing code" remained elusive.

With the advent of high-throughput approaches such as microarrays and deep sequencing, the mRNA landscape changed dramatically. Extensive mRNA profiles derived from unique cell lines or tissue types were reported within a short timeframe, thus permitting more specific alternative splicing analyses. These new tools have demonstrated that essentially all multi-intron containing genes undergo alternative splicing, generally in a tissue-specific manner [4, 8]. The May 6 issue of *Nature* reported on the combined efforts from the Blencowe and Frey laboratories designed to take a stab at deciphering the "splicing code" [9]. Equipped with extensive microarray results evaluating the inclusion levels of more than 3 500 exons from 27 diverse mouse tissues, Barash and colleagues set out to derive an algorithm capable of predicting alternative splicing outcomes based solely on RNA sequence elements. Unique to their ambitious approach was the requirement that many of the known RNA sequence features from all different classes of splicing parameters were included in their

Correspondence: Klemens J Hertel
E-mail: khertel@uci.edu

"splicing code", thus fully embracing the highly complex character of splicing regulation. A machine learning approach exploiting information entropy was used to infer which features influenced tissue-specific alternative splicing and optimal splicing was achieved when evaluating ~200 feature elements. Interestingly, most of these elements represent putative binding sites for *trans*-acting factors.

The performance of this first encompassing "splicing code" is impressive. The code correctly predicts the tissue-specific directionality of change in exon inclusion in 82.4% of cases using microarray data and 93.3% of cases using RT-PCR data. The code is also able to distinguish alternatively spliced exons from constitutively spliced exons, with a true detection frequency around 60% at a false positive rate of 1%. This means that the "splicing code" is capable of predicting likely exon inclusion or exclusion events in more than half of interrogated cases. While there is obvious room for improvement, it should be emphasized that this success rate exceeded most expectations, given the prevailing sentiment in the field that attempting to mix the many splicing regulatory elements into one pot would dilute any predictive power. One important take-home message of the work by Barash and colleagues is that a "splicing code" is attainable and that it can be made apparent using appropriate transcriptome data sets and computational tools. One of the obvious benefits of a "splicing code" lies in its predictive power that permits evaluating the functional consequences of disease mutations or disease-associated SNPs *in silico*. While the current detection rate of 60% may cast some doubt over whether an experimental team should dive into exhaustive molecular analyses, the first generation "splicing code" can be viewed as a potential splicing blueprint that guides experimental verification.

The alternative splicing analysis

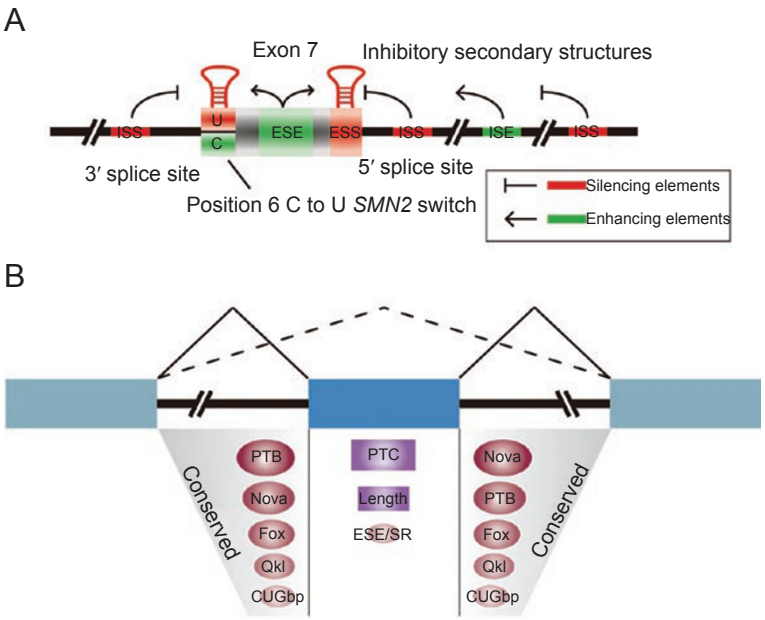

Figure 1 Multiple RNA elements influence alternative exon recognition. **(A)** Summary of the RNA elements that have been demonstrated to influence selection of *SMN* exon 7, a gene associated with Spinal Muscular Atrophy. Exhaustive biochemical analyses have identified splicing enhancer, silencer, and RNA secondary structure elements that modulate constitutive exon inclusion. A C to U transition occurs in the duplicated *SMN* gene, ultimately resulting in preferential exon 7 skipping. **(B)** The most influential tissue-specific alternative splicing features identified by Barash and colleagues (PTB, Nova, Fox, Qkl, and CUGbp) function mainly from within the flanking introns. The functional effects of this surprisingly small subset of splicing regulators are complemented with exonic features such as exon length, the generation of premature termination codons (PTC), or the activity of ESEs/SR proteins. The combinatorial contribution of these features dictate tissue-specific alternative exon inclusion. The size of the ovals (representing binding sites for splicing regulators) and boxes (representing architectural and RNA quality control features) indicates differences in the magnitude of their average regulatory influence.

makes the additional important point that inclusion or exclusion of exons relies on multiple RNA feature elements [9]. On average, 12 tissue-specific RNA features were necessary to define exon inclusion or exclusion in central nervous system tissues and 19 RNA features define embryo-specific alternative splicing. While certain tissue-specific exons share a subset of the same RNA feature elements, the final set of RNA features that make up an exon's identity is likely to be unique for each exon. From now on, it should be the expectation that alternative splicing is modulated by multiple RNA features. This expectation is fully supported when evaluating one of the biochemically most analyzed alternative splicing events, the inclusion of the *Survival of Motor Neuron* (*SMN*) exon 7, a gene associated with Spinal Muscular Atrophy [10]. Several splicing enhancers and silencers, either intronic or exonic have been identified through mutational analyses in addition to RNA secondary structure elements (Figure 1A), suggesting that even the splicing of constitutive exons, such as *SMN* exon 7 in its normal context, is mediated by multiple splicing RNA elements. For alternatively spliced exons it can be expected that additional RNA features

participate to ensure regulated exon selection. It should also be anticipated that tissue-specific alternative splicing is under the control of splicing networks that consist of unique RNA feature elements. The identity of these splicing networks varies between tissues, thus promoting tissue-specific alternative splicing.

What RNA features other than the general splice junctions have the most significant impact on tissue-specific alternative exon inclusion? As expected, splice junction sequences are of importance, but they are not believed to mediate regulation of alternative exon selection. Rather, they set in stone a baseline level of spliceosomal recognition. Among the various splicing regulatory factors known to date, a surprisingly small ensemble emerged as most influential. CUG-binding proteins (CUGbp) such as Mbnl or CELF-like factors, the Fox proteins, Nova1 and 2, Quaking-like (Qkl), and the PTBs lead the list acting mainly from within the flanking introns, while SR proteins appear to be less important in dictating alternative splicing. In fact, the most prominent exonic features that associate with tissue-specific splicing are the exon length and whether the inclusion or exclusion of the alternative exon results in the generation of a premature termination codon (PTC) (Figure 1B). This places the most significant regulatory signatures into the flanking introns, a conclusion supported by the observation that the "splicing code" functions optimally only when requiring high levels of evolutionary conservation.

It is clear that the foundation for deciphering the "splicing code" is the volume of mRNA isoform information. Without the recently developed high-throughput approaches, tissue-specific splicing comparisons would not have been possible at the levels required to achieve sufficient statistical power. While acknowledging the advance Barash and colleagues presented in their encompassing computational analysis of alternative splicing, the "splicing code" is far from being completed. With the advent of increased transcriptome information through deep sequencing, the immediate future promises a flood of detailed alternative mRNA isoform information unique to cellular environments or biological contexts. This expanding reservoir of mRNA transcript information will nurture the evolution of the "splicing code". It will refine and improve the performance of the current algorithms predicting alternative exon inclusion and to include the many other forms of alternative splicing, such as alternative splice site selection, intron retention, or mutually exclusive exon inclusion. For now, it should be appreciated that alternative splicing can be predictable from RNA sequence alone even in the face of highly complex regulatory networks.

References

1 Cooper TA, Wan L, Dreyfuss G. RNA and disease. *Cell* 2009; **136**:777-793.
2 Venables JP, Klinck R, Bramard A, *et al.* Identification of alternative splicing markers for breast cancer. *Cancer Res* 2008; **68**:9525-9531.
3 International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004; **431**:931-945.
4 Wang ET, Sandberg R, Luo S, *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008; **456**:470-476.
5 Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature* 2010; **463**:457-463.
6 Hertel KJ. Combinatorial control of exon recognition. *J Biol Chem* 2008; **283**:1211-1215.
7 Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* 2008; **14**:802-813.
8 Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 2008; **40**:1413-1415.
9 Barash Y, Calarco JA, Gao W, *et al.* Deciphering the splicing code. *Nature* 2010; **465**:53-59.
10 Singh RN. Evolving concepts on human SMN pre-mRNA splicing. *RNA Biol* 2007; **4**:7-10.