From single splicing events to thousands: the ambiguous step forward in splicing research

Emanuele Buratti, Marco Baralle and Francisco E. Baralle

Advance Access publication date 18 November 2012

Abstract

Since the discovery of RNA splicing in 1977 the knowledge of this important biological process has increased steadily following the identification of many of the mechanistic features of splicing: from the basic cis-acting splicing signals, through the detail composition and dynamics of the spliceosome, to the role played by accessory splicing factors and their interactions. Moreover, the realization that most genes undergo alternative splicing has had a strong impact in the overall cell proteome and metabolism research fields and also in better appraising the fundamental role played by splicing defects in human disease. This robust growth of knowledge is due in particular to the development of new powerful technical tools that range from methodologies useful to focus on single events in extreme detail to microarray and high-throughput RNA sequencing approaches that aim at providing a global vision of splicing changes. Here, we will discuss how these techniques relate to each other in terms of their respective strengths and weaknesses. In particular, we will focus on their value for evaluating the biological significance of splicing events. Finally, we provide some views on how these methodologies should move forward to improve our basic and applied knowledge of RNA splicing.

Keywords: RNA splicing; RT-PCR; Minigene; RNAseq; CLIP; splicing arrays

INTRODUCTION

RNA splicing was first described in the 1970s [1, 2] and since then has been shown to play a central role in regulating gene expression in higher organisms [3, 4], promoting proteome expansion [5], shaping evolutionary processes [6] and improving biotechnological processes [7]. In addition, many disease-causing mutations have been demonstrated to result in aberrant RNA splicing. It follows that an in-depth investigation of splicing molecular mechanisms integrated by clues provided from naturally occurring mutations is extremely important for both the basic research and clinical fields [8, 9]. For obvious reasons, substantial research has also been focused at elucidating how splicing itself is catalyzed [10]. As a result, we now have a fair understanding of the biology that underlies the action of the spliceosome: a highly dynamic macromolecular machinery that is physically responsible for the removal of introns and the joining together of exons with single-nucleotide fidelity [11, 12]. In parallel to these studies, a vast effort is still underway at clarifying the biological significance of RNA splicing in various tissues/developmental stages, both in normal and disease states [13].

To achieve these aims, researchers have developed several techniques capable of detecting, quantifying and eventually classifying splicing events. As will be discussed later, the goal of each new technique has been to try and find the best compromise between different aspects of splicing research namely quantification of naturally occurring constitutive and alternative splicing events, faithfully mimic splicing
processes that occur within a cell, and to obtain indications with regards to their biological significance.

**METHODOLOGIES THAT FOCUS ON SINGLE SPlicing EVENTS**

The biological ‘golden standard’ of any individual splicing event is represented by the direct observation of *in vivo* splicing patterns, either through northern blot or reverse transcriptase-polymerase chain reaction (RT–PCR)/quantitative RT–PCR (qRT–PCR) analyses (Figure 1a). This is relatively simple to achieve in animal and cellular models but more difficult to validate in human tissues. In fact, human samples suitable for RNA analysis may not always be available. In clinical genetics settings, they are often limited to leukocyte preparations. Even when available, direct analysis also requires to take into account the subject’s allele composition to unambiguously assign each splicing product to a specific context (i.e. considering the eventual presence of different splicing profiles from different alleles). Second, when looking at different human subjects it is also important to take into consideration individual physiological variations. Finally, it is now well known that many aberrant and regulatory transcripts are rapidly degraded *in vivo* by a surveillance mechanism such as nonsense-mediated decay and other...
types of quality-control mechanisms [14, 15]. This rapid degradation hampers direct observation in a sizable number of cases. These problems can be technically solved in a variety of ways. However, the investment of time and resources required to do so becomes significant and all these drawbacks limit the number of events that can be conveniently analyzed. Therefore, to gain a greater degree of independence from individual variations and patient’s sample availability, artificial systems have been set up.

*In vitro* splicing protocols [16] set up by researchers have proven to be very useful to characterize the splicing process itself. These include mechanisms of splice site recognition and regulation that can be recapitulated in the test tube and followed by a variety of techniques that monitor RNA-protein, RNA–RNA and protein–protein interactions, and dissect spliceosome assembly steps affected by regulatory factors. More recently, they have been adapted to investigate RNA–protein interactions in a functional manner through SELEX-based approaches (where SELEX stands for Systematic Evolution of Ligands by Exponential Enrichment) [17]. These systems, however, are not well suited to investigate many complex splicing events that frequently occur in nature. The reason is that unlike the cell, in the laboratories we lack the ability to synthesize long stretches of RNA *in vitro* and in any case nuclear extracts (that may be difficult to make to make with few exceptions) often cannot handle competently long RNAs. Therefore, *in vitro* approaches have been mostly replaced by minigene-based technologies that aim to exploit the natural transcriptional and splicing machinery of original cellular environment (Figure 1b). This approach was initially described almost 30 years ago [18] and has been continuously improved since then in a variety of ways [19, 20]. On the positive side, *in vitro* splicing systems have been essential to uncover the 200 or so factors that are involved in spliceosomal formation and function [21].

The limitations of *in vitro* splicing are overcome and complemented by minigenes that are probably the most common methodology used in both basic science research and the clinical setting as we know it today.

In addition to the simple mimicking natural or aberrant splicing processes, the use of minigenes has allowed to easily dissect complex regulatory events mediated by combinations of silencer and enhancer elements. For example, starting from the mapping of the first exonic splicing enhancer element in the EDA exon of the fibronectin gene [22] (Figure 2, left panels), minigene-based approaches have made it possible to explain very complex splicing phenomena such as mutually exclusive splicing events [23], of which the latest one has been studied in the Nav 1.6 sodium channel [24] (Figure 2, right panels). Finally, minigenes have also been proven to be very efficient at characterizing rather elusive splicing regulatory influences such as those provided by RNA secondary structure [25] or promoter architecture [26]. Initially, *in vitro* systems were very useful to test the influence of individual factors on a specific splicing event. This was achieved by immuno-depleting the various factors from the splicing-competent nuclear extracts, adding back bacterially expressed factors to observe the eventual rescue, etc. However, since the discovery of the RNAi process, functional analysis of the ligands interacting with the splicing regulatory elements (SREs) mapped by SELEX, using RNA-affinity *in vitro* binding techniques or *in vivo* binding assays such as the Cross-Linking and Immuno-Precipitation (CLIP) technique can now be carried out efficiently also in culture cells.

Nonetheless, because of their unavoidable out-of-origin-context nature, the use of minigenes is always accompanied by the risk of not mimicking the original biological splicing events in a proper manner. Even though this danger can often be reduced by optimizing the genomic milieu chosen for the analysis [27], the risk of falling into this trap will always be present and one must be especially careful of this possibility especially when studying exons that are already alternatively spliced. For this reason, minigene assays used for clinical molecular diagnostics studies need to be validated exon by exon by extensive comparisons with analyses of RNA derived from patients carrying mutations. Such comparisons have shown that in a majority of cases the assays were reliable, even if the exon was in a heterologous context [28, 29].

In conclusion, although the use of the minigene has greatly increased the number of splicing systems that can be analyzed in any given time, the biological information potentially provided by minigene studies may in some cases be lower than the one obtained from direct observation (Figure 1). For these reasons, it is not particularly suited to look at global splicing changes within specific organs/single cells.
METHODOLOGIES THAT FOCUS ON MULTIPLE SPICING EVENTS

Minigene and in vitro studies has given a plethora of information on single gene splicing functions and basic molecular mechanisms. Rather recently, the advent of high-throughput technologies gave the opportunity to assess the overall transcriptome status including RNA levels, splicing and alternative splicing events in most the cell’s RNAs. For these purposes, advanced microarray and high-throughput sequencing techniques have recently been applied for global analyses of RNA splicing processes [30].

An additional methodology that is not strictly confined to splicing analysis is the CLIP technique [31] (Figure 1c). CLIP has allowed the mapping in vivo of the binding sites of several splicing factors, such as NOVA [32], SRSF1 (ASF/SF2) [33], hnRNP A1 [34] and TDP-43 [35, 36] and Tra2β [37]. A very recent trend has been that of combining different techniques such as coupling CLIP analysis with or splicing-sensitive microarrays for various hnRNP proteins such as A1, F/H and M,U [38]. This is achieved by UV-crosslinking all the proteins bound to cellular RNA, followed by RNase digestion and immunoprecipitation against the specific factor under study. The bound RNA–protein complexes are purified, the protein digested, and the bound RNAs are reversely transcribed and sequenced. If used in conjunction with another global approach such as microarray or exon array analysis (see below), CLIP results can help to understand splicing regulatory events by providing a positional reference for the factor under study [39]. As with every technique, however, there are also some limitations [40]. The foremost, in our experience, is represented by the fact that in terms of functional consequences a high CLIP read around a specific RNA region for a specific factor is not necessarily more significant than regions in other transcripts that are CLIP-poor. In fact, factors such as abundance of the transcript itself, low crosslinking efficiencies, reverse-transcription stalling at the crosslinked site,
From single splicing events to thousands

and differential amplification efficiencies will strongly influence the end picture. In conclusion, the functional significance of the CLIPs cannot be easily predicted on the basis of their abundance alone, and needs to be verified on a case-by-case basis using the methods described in the previous section.

The analysis variations in global splicing levels is carried out by mainly two different approaches: high-throughput RNA sequencing (RNAseq) and splicing-junction or exon-array techniques [30] (Figure 1d and e). Both approaches allow the profiling of very large number of alternative splicing events in culture cells or in particular tissues. In the medical field, microarray and RNAseq approaches have also been used to try and identify functionally important changes in complex diseases from cancer [41] to neurodegeneration [42]. There are also several excellent examples of how global studies have brought insights of biological relevance that would be difficult to obtain using gene-by-gene approaches, for example by looking at the effect of RNA PolII elongation on global alternative splicing levels [43], or the transcriptome-wide effects of single splicing factors in different organs [44]. At present, the advantage of RNAseq (Figure 1d) over microarray techniques (Figure 1e) is that RNAseq does not require any previous knowledge of exons or splicing junctions to be spotted on the array chip. For this reason, RNAseq approaches can easily be used also to detect expression changes in the huge variety of non-coding RNAs as they become routinely mapped and classified. In addition, RNAseq also overcomes the problem of non-specific hybridization of the probes with the cDNA, is capable of detecting RNA editing events, and can quantitize splicing isoforms. Nonetheless, with respect to microarray approaches, one of the limitations of RNAseq is represented by the fact that the sequencing depth required to quantitatively measure variations in specific regions of a transcript is higher than for general expression analysis and this is particularly important for low- to mid-abundance transcripts. For a more comprehensive discussion of these issues, the reader is referred to the recent review by Malone and Oliver [45].

Until a few years ago, the relatively high cost of high-throughput RNA sequencing was the major limiting factor when deciding to use this kind approach to investigate splicing, as a limited number of ‘reads’ will always be associated with a reduced coverage of the genome. In recent times, however, the steep drop in sequencing prices has allowed to bypass this problem and nowadays there is the possibility to pay reasonable fees for several million sequencing reads from any RNA sample of interest. For this reason, a reasonable future-trend prediction is that RNAseq will probably become the preferred technical choice when looking at global splicing levels.

Together with microarray analysis, however, RNAseq suffers from the inevitable drawback that results cannot be connected in a straightforward manner to biological function. In fact, typical microarray and RNAseq experiments will yield a huge amount of possible targets, most of which will vary in expression between 1- and 2-fold with respect to control samples and sometimes even less. Many of these changes will therefore need to be validated on a case-by-case basis by qRT–PCR or other types of direct approach (and what is worse, the targets chosen for validation will of necessity reflect the subjective bias by the investigator). Therefore, in the absence of external lines of evidence all this huge amount of global information will be of limited use to guide researchers in identifying which splicing events are most important for a particular biological pathway or disease situation. One of the most promising future prospects to solve this problem would be to use microarray techniques that might focus on a limited number of target genes rather than the full analysis of all transcripts. This could be achieved by using RNA capture followed by sequencing with very high coverage.

Another major problem of these high-throughput approaches is represented by our still limited understanding of when different mRNA isoforms become biologically significant or can be considered to be biologically different. For example, a common and rather unexpected result of global studies performed so far is that multiple alternative splicing events occur in almost 98% of known genes [46]. This observation alone, however, does not necessarily mean that all splicing isoforms display significantly different biological properties (it is rather unlikely, actually). Current hypotheses are that many of these isoforms may be redundant. Many will be part of gene expression networks and will be rapidly degraded without ever being translated into proteins, and a fair fraction might even represent a biologically ‘neutral’ reservoir of protein species that are tolerated by the cell and might act as an evolutionary substrate to better adapt the cell in changing conditions. On the other hand, it has been recently found that many
tissue-specific, protein-coding exons code for ‘disordered’ regions (rather than distinct protein ‘domains’) that are enriched by post-translational modification sites or binding sites for other factors. This characteristic has been proposed to allow these alternatively spliced isoforms to be responsible for the required ‘rewiring’ of protein interaction network among different cell types [47, 48]. Whatever the answer, this is a controversial topic that has been recently discussed at length by Hsu and Hertel [49] and the reader is referred to this publication for further discussion. An additional possibility, discussed in the paragraphs later, is that many of these isoforms are simply splicing background noise.

The key issue that must be considered when evaluating high-throughput changes is where one must put the cut-off, especially in the lower range of splicing variations. In fact, it is quite probable that small changes in splicing efficiency (and hence protein production) could represent background oscillations of the technique or, even if real, may not be necessarily harmful. For example, it has been estimated that Cystic Fibrosis symptoms cannot be observed as long as normal production of the protein falls below a very low level (<10%) with respect to ‘average’ individual production levels (that is also by itself quite variable between asymptomatic carriers and non-carriers). This is probably not true for many other proteins. However, lacking this critical information causes researchers who look at high-throughput variations to face the problem of setting an arbitrary limit (i.e. usually 2- to 3-fold variations) and assume that the most represented variations will cause some kind of biological effect.

As mentioned before, there is an even more dangerous side to all these methodologies that can be summarized in the provocative question: how much of all this mountain of information is real and how much is just the outcome of a noise level due to the extreme amplification techniques used in these analyses? This seems a serious issue that is rarely presented in a rather structured form. We are going to limit ourselves to enunciate it and to point out the huge gap between the described splicing variants coming out of these analysis and the limited (several orders of magnitude lower) really established, biologically significant splicing variations. It is our impression that any report of high-throughput screening should be requested to really identify the meaning of at least 10% of the variations reported instead of the current 0.01–0.1%. What value has to report that 1000 genes go up and 1000 genes go down and then focus on the one or two of interest of the specific investigation? Or that there are hundreds of CLIPs and then out of those only a few are really binding sites for the specific protein and at most a fraction of them produces a meaningful effect. In our opinion, therefore, in future studies the focus should shift to better understanding the meaning of few changes rather than to describe thousands of changes and then speculate about their possible meanings.

**BIOINFORMATICS APPROACHES**

Because many of these high-throughput approaches heavily rely on bioinformatics data management, a final word should also be reserved to bioinformatics programs that predict strength of splicing signals or splicing efficiencies (Figure 1f). These approaches either target the basic cis-acting splicing elements (5’ss, 3’ss and branch-point) or the various SREs [50]. As with all in silico approaches their major advantage is that they can be applied to as many sequences as are available and running the analysis is practically cost-free. At the same time, however, the key question regarding their use is represented by the degree of reliance that one can place in each prediction. In general, because the donor and acceptor elements tend to be reasonably conserved, programs that evaluate their relative strengths seem to be rather more successful than those targeting the more loosely conserved SRE elements. Current SRE-targeting programs, in fact, have worked reasonably well in some contexts [51] but rather poorly in others [52]. In this respect, however, it should be noted that several steps forward have been made by linking SELEX- and minigene-based assessment of potential SRE sequences, strengths, and actions with deep sequencing of successfully spliced transcripts [53]. Nonetheless, the general consensus is that the combinatorial use of all these resources represents, to this date, the best chance of ‘predicting’ putative splicing mutations whether in conserved or less conserved regions [54]. In summary, notwithstanding their growing usefulness, all bioinformatics programs still show a considerable and unforeseeable degree of unreliability depending on the system(s) analyzed and this adds an additional difficulty in using them to evaluate biological consequences particularly in a clinical setting (Figure 1f).
**Table 1:** Systems used to investigate alternative splicing events/profiles

<table>
<thead>
<tr>
<th>Technology</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Cost estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>qRT–PCR</td>
<td>High precision and high sensitivity in direct detection of expression levels.</td>
<td>Has to be optimized for each single gene/isof orm of interest.</td>
<td>Low. Expenses are basically limited to cost of oligos and detection method</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Care must be taken with choice of method genes used to normalize results.</td>
<td>(SYBR Green, Taqman and molecular beacons).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RNA from patients/tissues is not always available.</td>
<td></td>
</tr>
<tr>
<td><strong>In vitro splicing assays</strong></td>
<td>Splicing intermediates and kinetics of spliceosome assembly can be followed easily.</td>
<td>Has to be optimized for each exon or sequence of interest.</td>
<td>Low. Expenses are basically limited to cost of synthesizing RNA and preparing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Splicing factors can be removed and replaced to assess exactly their activity.</td>
<td>splicing-competent nuclear extracts from culture cells.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Can be used to dissect the work of regulatory elements outside their original context.</td>
<td></td>
</tr>
<tr>
<td>Minigenes</td>
<td>RNA is processed <em>in vivo</em>.</td>
<td>Has to be optimized for each exon or gene region of interest.</td>
<td>Relatively low. Expenses are limited to maintaining standard cell culture</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laborious.</td>
<td>facilities and purchasing RT-PCR and qRT–PCR reagents.</td>
</tr>
<tr>
<td></td>
<td>Useful to analyze complex exonic/intronic regions.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Can be used to routinely screen for disease mutations in clinical screening analyses.</td>
<td>Cannot be used to investigate splicing intermediates.</td>
<td></td>
</tr>
<tr>
<td>Microarray</td>
<td>Genome-wide level analysis.</td>
<td>Results should always be checked against patient RNA.</td>
<td>Relatively high.* Cost of commercial Chips varies from $100 to $1000 but data analysis is usually included with purchase.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Huge range of Chips commercially available that can also be custom-made.</td>
<td>Results require extensive validation.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Handling and data analysis procedures are also widely available and optimized.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLIP</td>
<td>Genome-wide level analysis.</td>
<td>Functional relationship between CLIP hits and gene expression must be</td>
<td>Very High.* Same pricing applies as for RNA-seq analysis.</td>
</tr>
<tr>
<td></td>
<td>Provides a global picture of the transcripts and target sites bound by an RNA binding protein.</td>
<td>experimentally validated in every case.</td>
<td></td>
</tr>
<tr>
<td>RNA-seq</td>
<td>Genome-wide level analysis.</td>
<td>Good coverage is needed to obtain reliable information regarding low- to</td>
<td>Very High.* Depending on type of run and datalsize required each run can cost from $1000 to $3000 and more.</td>
</tr>
<tr>
<td></td>
<td>Sequences do not need to be known in advance.</td>
<td>mid-expressed transcripts.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RNA editing events can be detected. Splicing isoforms can be quantitated.</td>
<td>Unlike microarrays, there is still little knowledge of potential experimental bias in current technologies.</td>
<td></td>
</tr>
</tbody>
</table>

*In general, replicate runs should be considered per sample to allow for statistical handling of results. This substantially increases the cost of each experiment.

**CONCLUSIONS**

Research in splicing mechanisms has made several leaps forward in recent years, especially with regards to techniques that aim to study global changes in splicing levels *in vivo*. The move toward the global direction, however, is still plagued by a rather steep loss in the quality of the biological information obtained. Another problem that is often under-reported is the limited amount of overlap that can sometimes be observed between microarray experiments that in theory should yield similar if not identical results (e.g. when a particular factor is knocked out in similar or identical cell lines), an example has been previously discussed for the
TDP-43 nuclear protein [55]. There are many causes that could account for these differences and they include physiological condition of the cells, the many variables of each individual experiment (that are usually self-consistent), platform specificities, differing ‘data-crunching’ softwares, etc. Nonetheless, these occurrences should light a warning signal about the reliability of placing much confidence in microarray results that are not extensively validated. To this date, this problem cannot be easily solved in the absence of old-fashion and extensive validation studies. What we are still lacking, therefore, is some kind of system that allows the collection of multiple data without any loss in the quality of the biological information associated with it (Figure 1g). In the meantime, the various advantages and disadvantages of each technique discussed in this review together with their estimated costs have been summarized in Table 1.

However, this new highly reliable and widely applicable system (Figure 1g) could be achieved in several ways depending on the relative position of each technique. For example, direct detection and minigene analyses can be automatized to increase the number of samples analyzed in any given time. Alternatively, advances in data analysis and a better ability to quantify the results of high-throughput assays could be aimed at filtering out background ‘splicing noise’. A third possibility that could be implemented very easily would be to reach a trade-off between these existing techniques. For example, it might be viable to focus on genes that represent a priority in terms of costs/benefits for human health (i.e. BRCA1, NF1, etc.) and set up dedicated microarray approaches able to detect in a very precise manner all the known pathological splicing isoforms. Another possibility would be to prepare ‘minigene banks’ that cover the entire exon and most intronic sequences of selected genes that would allow the rapid assessment of any novel sequence variation obtained from clinical screening studies.

Aside from these considerations, however, the underlining conclusion that can be drawn from this overview is that researchers should not get carried away by thinking that just getting ‘more’ data will be a good result by itself. If anything, in fact, it will just add more confusion to a subject where many basic questions are still seeking a satisfactory answer.

**Key Points**

- Methodologies that look at RNA splicing have been optimized either to look at single events or global events. Both type of methodologies have advantages and drawbacks.
- Major advantages: methodologies that look at single splicing changes can be generally adapted to mimic real biological changes/significance. On the other hand, techniques that look for global changes allow to obtain a ‘bird’s eye’ view of the enormous complexity of the splicing process.
- Major drawbacks: looking at single splicing changes in detail always forces the big picture into the background. On the other hand, global techniques provide huge amounts of raw data from which it is always difficult to extract the ‘wheat from the chaff’.

**FUNDING**

This work was supported by Agenzia di Ricerca per la Sclerosi Laterale Amiotrotica (ARISLA).

**References**


