Role of small nuclear RNAs in eukaryotic gene expression

Saba Valadkhan¹ and Lalith S. Gunawardane

Center for RNA Molecular Biology, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, U.S.A.

Abstract
Eukaryotic cells contain small, highly abundant, nuclear-localized non-coding RNAs [snRNAs (small nuclear RNAs)] which play important roles in splicing of introns from primary genomic transcripts. Through a combination of RNA–RNA and RNA–protein interactions, two of the snRNPs, U1 and U2, recognize the splice sites and the branch site of introns. A complex remodelling of RNA–RNA and protein-based interactions follows, resulting in the assembly of catalytically competent spliceosomes, in which the snRNAs and their bound proteins play central roles. This process involves formation of extensive base-pairing interactions between U2 and U6, U6 and the 5′ splice site, and U5 and the exonic sequences immediately adjacent to the 5′ and 3′ splice sites. Thus RNA–RNA interactions involving U2, U5 and U6 help position the reacting groups of the first and second steps of splicing. In addition, U6 is also thought to participate in formation of the spliceosomal active site. Furthermore, emerging evidence suggests additional roles for snRNAs in regulation of various aspects of RNA biogenesis, from transcription to polyadenylation and RNA stability. These snRNP-mediated regulatory roles probably serve to ensure the co-ordination of the different processes involved in biogenesis of RNAs and point to the central importance of snRNAs in eukaryotic gene expression.

Keywords:
group II intron, ribozyme, small nuclear RNA, spliceosome.

¹To whom correspondence should be addressed (email saba.valadkhan@case.edu).
Introduction: the challenge of splicing and evolution of eukaryotic snRNAs (small nuclear RNAs)

A fascinating feature of modern eukaryotic genes is the nearly ubiquitous presence of intervening sequences or introns, which interrupt the continuity of the information content of genes. Thus, before primary gene transcripts can be used by the cell, introns must be accurately removed or 'spliced'[1]. In addition, recent research indicates that introns themselves often harbour regulatory or otherwise functional sequences, and their accurate and timely removal is often critical for their cellular function [2]. The intronic sequences in higher eukaryotic genes are much longer than the non-intronic sequences, the exons, and the sequence-based information that specifies the intron–exon boundaries is highly complex and poorly understood. Thus accurately distinguishing these two sets of functional sequences that co-exist in eukaryotic primary transcripts is a highly challenging task for the eukaryotic gene expression machinery.

Although modern mammals have one of the most complex splicing patterns among extant eukaryotes, it is likely that even in primordial eukaryotes splicing was already a highly complex process. On the basis of currently accepted models of evolution of eukaryotes, introns probably originated from self-splicing ribozymes that dated from pre-cellular life and constituted the majority of the genomes of ancient eukaryotes [3]. Later on, probably in order to prevent genomic instability, the introns lost their self-splicing capacity and, instead, the splicing function was delegated to a cellular machine, the spliceosome, which acted in trans to remove introns from primary transcripts. Although the origin and evolution of the early spliceosomes is still largely mysterious, several lines of evidence suggest that they probably evolved from self-splicing introns [4]. This hypothesis is partly based on the fact that the mechanism of intron removal by the spliceosome, performed through two consecutive transesterification reactions resulting in removal of a branched lariat intron, is identical with the splicing reaction performed by a class of extant self-splicing introns called the group II introns [5] (Figures 1 and 2). These introns, which are found in all three kingdoms of life, are RNA-centric catalytic sequences composed of a number of base-paired RNA structures called ‘domains’ (Figure 2) [6]. Extensive research has elucidated the identity of the catalytically essential sequences in these introns (Figure 2) [6]. Intriguingly, the RNA components of the spliceosome, the snRNAs, show unmistakable similarities to fragments of the catalytically essential domains of group II introns in sequence, secondary structure and function. Of the five major spliceosomal snRNAs (U1, U2, U4, U5 and U6), three of them (U2, U5 and U6) have clear structural and functional similarities to critical domains of group II introns (Figure 2). Domain-swapping experiments have indicated that isolated domains of group II introns and U5 and U6 snRNA substructures could functionally replace each other, proving their functional equivalence [5,6]. Another set of sequences in group II introns are functionally equivalent to U1 snRNA, despite the lack of structural similarity (Figure 2). Currently there are no known functional or structural equivalents for the U4 snRNA in group II introns and the evolutionary origin of this snRNA is completely unknown. On the other hand, a number of domains of the group II introns do not have an equivalent among the spliceosomal snRNAs and it is likely that, in the spliceosome, these RNA domains are replaced by spliceosomal proteins. While the possibility
Figure 1. U6 and U2 snRNAs and the mRNA at the time of first and second steps of splicing

The location of U6, U2 and the U6 ISL is shown. The intron is shown by a thick light blue line connecting the two exons. Position of the 5′ splice site (5′SS), 3′ splice site (3′SS) and branch site are shown. Solid arrows point to the site of the nucleophilic attack during the two steps of splicing. The first step involves a nucleophilic attack by the 2′ hydroxy group of a specific adenosine residue in the intron, the branch site adenosine (the bulged A), on the 5′ splice site. This leads to a trans-esterification reaction in which the 2′ oxygen of the branch-site adenosine replaces the 3′ oxygen of the last nucleotide of the upstream exon. The result of this reaction is the release of the first exon and the formation of an unusual 2′–5′ linkage between the branch site adenosine and the first nucleotide of the intron (right-hand panel). During the second step, the free 3′ hydroxyl moiety of the newly released exon is activated for a similar nucleophilic attack on the 3′ splice site, resulting in ligation of the two exons and release of the intron as a branched lariat. Base-pairing interactions are shown by short black lines. The location of the 2′–5′ linkage formed after the first step of splicing at the branch site is shown.

Figure 2. The structural organization of the group II self-splicing intron al5γ

The location of domains I–VI, the two exons (shown in green), the splice sites and the branch site (5′SS, 3′SS and BS respectively) are shown. The position of J2/3 and the AGC sequence are indicated. The metal-binding site of domain V is shown by a red ‘Mg’ sign. Broken lines connect regions which are juxtaposed to form the catalytic core. The circles denote the functional equivalent of each domain or subdomain in the spliceosome. ϵ and ϵ′ sites, which are involved in an interaction important in recognition of the 5′ splice site, are shown.
of convergent evolution cannot be formally ruled out, the above-mentioned similarities strongly suggest that at least a number of snRNAs are evolutionary remnants of primordial self-splicing ribozymes.

**Roles of snRNAs in the spliceosome**

The snRNAs were first discovered in 1970s as small highly abundant nuclear RNAs which formed the core of RNP (ribonucleoprotein) particles which showed strong reactivity with the immune sera from patients with autoimmune disorders [7,8]. Further analysis indicated the presence of sequence complementarity between one of the snRNAs, U1, and the sequences found at the 5′ splice site of primary transcripts, ultimately leading to the discovery of their involvement in splicing [9]. Further research identified a second set of snRNAs, named U11, U12, U4atac and U6atac, which are functional counterparts of U1, U2, U4 and U6 snRNAs respectively, and participate in the formation of a ‘minor’ spliceosome which is responsible for removal of an atypical subset of introns, most of which have alternative consensus sequences at the splice sites and branch site [10]. U5 snRNA is found in both ‘major’ and ‘minor’ spliceosomes. Each of the nine spliceosomal snRNAs are stably associated with a set of proteins, creating the snRNP (small nuclear ribonucleoprotein) particles which form the main functional subunits of the spliceosome. Analysis of the spliceosomal function suggests that snRNPs and several non-snRNP spliceosomal proteins assemble on each intron in a stepwise elaborate fashion through a large number of conformational rearrangements which start from the recognition of the splice junctions and culminate in splicing catalysis, followed by disassembly and recycling of the spliceosomal components [1,11] (Figure 3).

**Recognition of the 5′ splice site by U1 snRNP**

Recognition of introns in primary transcripts is partly mediated by detection of ‘consensus’ sequences found at the junction of introns and exons, the 5′ and 3′ splice sites. These consensus sequences are rather short: 5′-AG/GURAGU and 5′-YAG/G for 5′ and 3′ splice sites respectively, where R denotes either a G or an A, and Y denotes a C or a U and / marks the location of the splice site [1,11]. Interestingly, in higher eukaryotes and especially in mammals, the splice site sequences are highly degenerate and, in many cases, significantly deviate from the consensus sequence, thus necessitating an elaborate multi-step recognition mechanism mediated by a combination of RNA–RNA base pairing and RNA–protein interactions [1,11].

The association of U1 snRNP and its functional equivalent in the ‘minor’ spliceosomes, U11, with the 5′ splice site is one of the earliest and arguably most important events in the spliceosomal assembly pathway [11,12] (Figure 3, the E complex). The recognition and binding of the 5′ splice site is mediated both by base pairing of a single-stranded sequence at the 5′-end of U1 snRNA to the 5′ splice site and through an intricate web of interactions between the pre-mRNA and U1C, a U1-specific protein [5,11,13]. Interestingly, a high-resolution structure of U1 snRNP indicated the presence of a number of interactions between U1C and the nucleotides at the 5′ end of U1 which base pair to the 5′ splice site, thus providing a structural basis for the dual RNA–protein recognition of the 5′ splice site [12,13]. While this RNA–protein recognition of the 5′ splice site by U1 is functionally critical for the majority of cellular transcripts in vivo, several other proteins also contribute to the selection of the 5′ splice site [12]. Furthermore, at least some primary transcripts can be spliced in the absence of U1 snRNP in
vitro, pointing to extensive redundancy in the splicing machinery. Finally, the binding of U1 is not necessarily synonymous with productive splicing, since the binding of U1 to sequences involved in negative regulation of splicing has been documented [12]. Current data suggest that the binding of U1 to such elements is important in splicing regulation and exclusion of pseudo splice sites, underscoring the importance of sequence context in splicing.

**The U2 snRNP and recognition of the branch site and 3’ splice site**

Once bound to the 5’ splice site in a sequence context which is conducive to splicing, U1 helps initiate the spliceosomal assembly by forming a network of interactions with U2 snRNP that plays a dominant role in recognition of the 3’ splice site and the branch site, another region in the introns which is recognized by the spliceosomes [12,14]. Branch sites, which are typically located ~30 nt from the 3’-end of introns, contain the adenosine which acts as the nucleophile of the first transesterification step of splicing and forms the branched lariat structure found in the splicing intermediates and post-splicing introns (Figures 1 and 3). In early spliceosomes, U2 snRNP is loosely associated with the end of the intron through protein-mediated interactions. However, in an ATP-dependent step which involves the displacement of intron-bound proteins and remodelling of base-pairing interactions within the U2 snRNA, U2 forms a stable interaction with the branch site and 3’ splice site [11,13,15] (Figure 3, the A complex).
interaction is partly mediated through a base-pairing interaction between U2 and sequences flanking the branch site adenosine (Figure 4), and is stabilized by several RNA–protein interactions. The base pairing between U2 and the branch site leaves the branch site adenosine in an unpaired extrahelical conformation necessary for efficient splicing (Figure 4) [5,6]. In addition, the interaction between U1 and U2 snRNPs generates a loop that brings the 5′ and 3′ splice sites together and helps to ‘define’ the introns and exons [14].

**Formation of a catalytically active spliceosome**

Three of the snRNPs, U4, U5 and U6, form a ternary complex termed the ‘tri-snRNP’ and collectively integrate into the assembling spliceosomes [1,5] (Figure 3, the B complex). Of all of the spliceosomal snRNAs, U6 is the most conserved and contains two invariant domains, the ACAGAGA and AGC boxes, which play a critical functional role in splicing (Figure 4) [5,16]. Furthermore, it contains an ISL (intramolecular stem-loop) which is almost identical with the catalytic domain V of group II introns and, similar to its group II intron counterpart, binds a functionally required divalent cation, pointing to a critical role in splicing catalysis for this snRNA (Figures 2, 4 and 5) [5,16]. Perhaps in order to prevent it from prematurely forming a catalytically active structure, within the tri–snRNP U6 is kept in an inactive conformation.
through a base-pairing interaction with U4 snRNA that prevents the formation of its functionally critical ISL. Current data do not indicate any additional functions for U4 snRNA except acting as a negative chaperon for U6. Once the base paired U4/U6 complex joins the spliceosome in association with U5 within the tri-snRNP, the U4/U6 duplex is unwound in a tightly controlled manner and U2 snRNA replaces U4 as the base-pairing partner of U6 (Figure 4) [1,5,16]. In addition, U6 replaces U1 at the 5′ splice site, forming canonical and non-canonical base-pairing interactions with this sequence (Figure 3, B act complex and Figure 4). The base-pairing interactions between U6 and U2 allow the formation of the U6 ISL and, further, serve to juxtapose the branch site, which is bound to the branch-binding sequence in U2, and the 5′ splice site, which is bound by U6 (Figure 4). At the same time, U5 snRNA forms non-canonical base-pairing interactions with the exon sequences immediately adjacent to the splice sites and participates in aligning the exons to ensure their optimal positioning for the second step of splicing (Figure 4) [1,5,16]. These base-pairing rearrangements are accompanied by an extensive rearrangement of protein–protein and protein–RNA interactions, culminating in the formation of catalytically competent spliceosomes (B* and C complexes, Figure 3). U1 and U4 are not stably associated with fully assembled spliceosomes, thus leaving U2, U5 and U6 snRNAs as the only spliceosomal RNA components required for catalysis (Figure 4) [5]. Interestingly, as mentioned above, these three snRNAs have clear structural and functional counterparts in self-splicing group II introns, raising the possibility that the catalytic core of the two splicing systems may be closely similar (Figure 2) [5,6,17–19].

Catalysis of splicing: the role of snRNAs

Mutagenesis studies have shown that, at least in vitro, the conserved loop of U5, which was previously shown to be the functionally important domain of the molecule (Figure 4), was in fact dispensable for splicing [1,5]. On the other hand, it has been shown that, under certain
conditions, several positions within the branch binding sequence of U2 snRNA can be mistakenly recognized as the 5′ splice site [20]. These results imply that the branch binding sequence of U2, which is functionally the most critical region of this snRNA, is not essential for spliceosomal catalysis, at least under certain conditions. As the rest of U2 seems to mainly fulfill structural roles by forming base-pairing interactions with other U2 sequences or with U6 snRNA [5,15], these results suggest that U6 snRNA may be the only RNA that is absolutely crucial for splicing catalysis, at least under the conditions studied so far.

Several additional lines of evidence suggest that U6 may form part of the catalytic domain of the spliceosome [5,16]. U6 is the most conserved of all spliceosomal snRNAs and several point mutations in its two evolutionarily invariant sequences, the ACAGAGA and AGC boxes, lead to a block in splicing, pointing to a critical function for these two sequences (Figures 4 and 5). Cross-linking and mutational complementation analyses have indicated that the first step of splicing occurs in close proximity to the ACAGAGA box, suggesting that this sequence is in the immediate vicinity of or even forms part of the spliceosomal active site. Current data suggest that in group II introns, the active site is formed by juxtaposition of the AGC triad and the asymmetric internal bulge of domain V along with a short purine-rich sequence (J2/3, Figure 2) which is considered functionally equivalent to the ACAGAGA box in U6 [19]. Interestingly, U6 contains the equivalent of all of these sequences which form the active site in group II introns, and close similarities in phosphorothioate interference patterns between catalytic domain V of group II introns and U6 (Figure 5) suggest that they may be functionally related [5,17–19]. Furthermore, hydroxyl radical footprinting and in vivo mutagenesis studies have pointed to the proximity of the AGC triad, the ACAGAGA box and the area near the bulged residue in ISL in functional spliceosomes (Figure 4) [5], perhaps in an arrangement similar to or even identical with the one found in group II introns [5,17–19].

Interestingly, analyses on in-vitro-transcribed, protein-free U6 and U2 snRNAs indicate that they can efficiently form a base-paired complex in vitro which in many respects resembles the one formed in the activated spliceosomes (Figure 4) [5,21–23]. Furthermore, it has been shown that the in-vitro-assembled human U2–U6 complex can indeed catalyse a two-step splicing reaction which closely resembles the one catalysed by the self-splicing group II introns and the spliceosome [24,25]. On the basis of the data above, the snRNAs seem to be fully competent to form the majority, if not all, of the spliceosomal active site and to perform catalysis, similar to the self-splicing group II introns, albeit with much lower efficiency.

If we assume that the spliceosome is an RNA catalyst, the snRNAs are unusual ribozymes in many respects, perhaps most importantly they are unusually small compared with other natural ribozymes catalysing splicing reactions. The larger size of other natural splicing ribozymes is thought to allow them to fold into complex tertiary structures, which in turn enable them to create sophisticated active sites necessary for such complex reactions. It is conceivable that due to their short length, the U6 and U2 snRNAs at best form an inefficient splicing ribozyme, which requires other spliceosomal factors for stable positioning of the active-site elements and the reacting groups. Although the exact role played by the proteins in the spliceosomal catalytic core is mostly unknown, their possible roles could range from assisting the snRNAs in assuming their functional structure, assisting in or independently co-ordinating critical metal ions and participating in the positioning of the substrates, to independently forming part of the active site and even direct involvement in catalysis [26].
Beyond splicing: other biological roles of snRNAs

Although the spliceosomal snRNAs (and their minor spliceosomal counterparts U11, U12, U4atac and U6 atac) play major roles in spliceosomal function, data suggest additional roles in regulation of gene expression for the snRNP particles. The interaction of U1 and the 5′ splice site, in addition to its function in splice-site selection, also seems to play a role in stabilization of some messages [12]. Both U1 and U2 snRNPs have been implicated in transcriptional regulation through stimulation of the rate of formation of the first phosphodiester bond at transcription initiation and also through interaction with a component of the pre-initiation complex, TFIIH (transcription factor II H) respectively [27,28]. U1 snRNP seems to also regulate the efficiency of polyadenylation via the interaction of a U1-specific protein, U1A, with a component of the CPSF (cleavage and polyadenylation stimulating factor) [12]. In addition, it has been reported that binding of U1 to a 5′ splice site-like sequence in the 3′-UTR (untranslated region) of some mRNAs inhibits their polyadenylation, leading to degradation of the RNA [12]. Thus the snRNAs and their bound proteins seem to act in co-ordination of the various steps in gene expression, in addition to playing the central role in splicing.

Another spliceosomal snRNP, the SL (spliced leader) particle, plays a critical role as the splice donor in a non-canonical trans-splicing reaction mainly observed in some protozoa and lower invertebrates [29]. In the trans-splicing reaction, SL is treated as a mini exon plus a short intron, with the exonic sequences ‘spliced’ in trans to a 3′ splice site on the primary transcripts of trans-splicing organisms. Thus, unlike the other spliceosomal snRNAs, SL is consumed during the trans-splicing reaction.

Although the majority of snRNAs play a role in splicing, there are other abundant small nuclear-localized RNAs which play critical roles in other cellular processes. A non-spliceosomal snRNA, U7, functions in 3′-end processing of replication-dependent histone mRNAs, which are not polyadenylated and instead terminate in a conserved stem-loop structure (SL element) generated by endonucleolytic cleavage of the pre-mRNA [30]. Similar to the spliceosomal snRNAs, U7 also forms an snRNP by associating with a set of proteins which together form the so-called Sm ring, and is recruited to histone pre-mRNA primarily through base-pairing interactions via its 5′-end with a purine-rich HDE (histone downstream element) which is located in the vicinity of the cleavage site. Together with a protein which binds the SL element, U7 recruits a complex that triggers endonucleolytic cleavage between SL and the HDE by the CPSF73 endonuclease, thus forming the mature histone mRNA.

Conclusion

As detailed above, the ability of snRNAs to form strong, specific interactions via base pairing with another RNA is extensively utilized in the spliceosome and during processing of the histone 3′-ends. Base-pairing interactions contribute to substrate recognition (U1, U2 and U7), positioning of the branch site in a strained catalytic bulged conformation (U2), regulation of the activity of another snRNA (U4) and juxtaposition of reactive substrates (U2, U5 and U6). Since RNA–RNA interactions similar to those formed by some of the spliceosomal snRNAs
play identical or closely related roles in group II introns, it is conceivable that the robustness of RNA–RNA interactions has led to their preservation throughout the evolution of the spliceosome from group II-like ancient ribozymes. Although RNAs can perform the above-mentioned tasks with ease and even more effectively than proteins, when it comes to catalysis, proteins seem to have an advantage over RNA, at least in the case of natural ribozymes. The evolutionary reason behind the preservation of U6 snRNAs as a constituent of the spliceosomal catalytic core remains an open question.

Another feature of the snRNAs is their participation in multiple sets of base-pairing interactions that at times are mutually exclusive and, thus, act as switches between different functional states. The presence of such interactions underscores the highly complex evolutionary pressures under which the snRNAs have evolved. In addition, most of the snRNAs bind a number of proteins which play important functional roles. While in many cases these proteins complement the function of the snRNA to which they bind, emerging evidence suggests that they can impart a completely novel function on the snRNP. Research in the coming years is likely to provide additional instances of multi-functionality of the snRNPs and further elucidate their contribution to the highly complex network of interactions which regulate eukaryotic gene expression.

Summary

- The majority of cellular snRNAs (small nuclear RNAs) function in splicing, with their ability to form specific base-pairing interactions extensively utilized in recognition of functional sequence elements in primary transcripts. Examples include the interactions between U1 and U6 and the 5′ splice site, U2 and the branch site, and U5 and the exonic sequences.
- Another set of RNA–RNA interactions occur between snRNAs and plays important roles in regulation of the timing of their folding into their active structure (in the case of the interaction between U4 and U6) or form a structural scaffold for juxtaposition of the reactive groups in the splicing reaction (formation of the U2/U6 base-paired complex).
- In addition to forming RNA–RNA interactions, U6 snRNA seems to play a critical role in catalysis of the splicing reaction. Whether the spliceosomal proteins play a role in catalysis remains to be determined.
- Another aspect of the function of snRNAs is forming interactions with a set of proteins which play important roles in spliceosomal assembly, regulation and co-ordination of splicing and other steps of gene expression.

References