Deposited research article

Sequence complementarity of U2 snRNA and U2A' intron predicts intron function

Maria Lundin

Address: Dept. Life Sciences, Södertörn University College, SE-141 89 Huddinge, Sweden. E-mail: maria.lundin@sh.se

Posted: 29 March 2005

Genome ***Biology*** 2005, **6:**P6

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/4/P6

Received: 24 March 2005

This is the first version of this article to be made available publicly.

Sequence complementarity of U2 snRNA and U2A' intron predicts intron function

Maria Lundin

Dept. Life Sciences, Södertörn University College, SE-141 89 Huddinge, Sweden

Telephone: +46 8 6084712

Telefax: +46 8 6084510

Email: maria.lundin@sh.se

Key words: U2A' intron, U2 snRNA, U2 snRNP, *RNU2*

Abstract

Background: The human genome contains about 24 % introns and only 1-2 % exons. Why such large amount of intron RNA is produced is not known. This paper exemplifies a putative function of an intron RNA, the alternatively spliced intron 5, exon 6 and intron 6 (i5e6i6) of the U2 small nuclear ribonucleoprotein particle (U2 snRNP) A' specific protein (U2A') pre mRNA. The U2 snRNP is a central component of the spliceosomes and very abundant in human nucleus. The *U2 snRNA* genes are tandemly repeated in the *RNU2* locus which occasionally co-localize to Cajal bodies in a transcription dependent process not very well understood. We have earlier found that U2A' exon 6 that is skipped in alternative splicing, is highly conserved in its nucleotide sequence. In this paper I have searched for a possible function of the U2A'i5e6i6 RNA.

Results: The U2A'i5e6i6 contains conserved sequence cassettes that are complementary to cassettes of the U2 snRNA. A possible RNA-RNA structure, based on RNA helices that may form by these complementary sequences, is presented. The structure, which is conserved in vertebrates, suggests a role of U2A'i5e6i6 in the 3'end processing of U2 snRNA primary transcript.

Conclusion: I predict a function of the U2A' i5e6i6 RNA in the 3'end processing of the U2 snRNA primary transcripts, a process that most probably occur during the RNU co-localization to Cajal Bodies. The production of U2 snRNPs would, thus be autoregulated by coupling of splicing efficiency of one of its components (U2A') to transcription of another (U2 snRNA). Such autoregulatory function may well be a common feature of introns.

Background

The U2 small nuclear ribonucleoprotein particle (U2 snRNP) plays a central role during splicing. It recognizes and binds to the branch point and takes active part in catalysis of splicing. The U2 snRNP A' specific protein (U2A') is as the name interpret a protein specific to the U2 snRNP. In addition to the specific proteins the U2 snRNP contains a number of Sm proteins that are constituents also of the other snRNPs (U1, U4, U5 and U6) involved in splicing as well as other proteins, reviewed in [1-3]. In addition to non-specific and specific proteins the snRNPs also contain one specific RNA each. The U2 snRNP recognises and bind to the branch point of the intron that is to be spliced. In this process the U2 snRNA plays a role in the sequence recognition by base pairing between the intron branch site and the U2 snRNA, as well as in catalysis [4, 5].

The U2 snRNA sequences are very conserved (see the uRNA database [6]). Human U2 snRNAs are encoded by 10- 30 genes of tandemly repeated units of the *RNU2* locus located at chromosome 17. The U2 snRNAs are transcribed by RNA polymerase II. The genes are TATA-less, contain no introns and are not polyadenylated, however, the 5'-end is monomethyl capped. The primary U2 transcript contains a 3' extension of up to several hundred nucleotides including a 3´-box of 10-11 nucleotides right downstream of the mature U2 snRNA. The processing of the 3' end of the primary U2 transcript requires the 3' box, phosphorylated CTD (C-terminal domain) of pol II as well as certain snRNA specific promoter factors [7-9]. This process has been suggested to require "a specialised snRNA-specific transcription and processing complex" [9]. Some of the *RNU2* loci are associated with Cajal bodies during transcription [10-12]. This association has been proposed to depend on "guide RNPs" base pairing to the nascent

RNA transcript [10]. The processing of the primary U2 transcript probably occur before entry of the PreU2 containing the 3'box, into Cajal bodies [12].

The pre-U2 snRNAs are transported to the cytoplasm where the 5' mono methyl cap is tri methylated, the 3' box is cut off and Sm proteins are bound to the U2 snRNA and subsequently the snRNA is transported back to the nucleus [1]. Back in the nucleus the bases of U2 snRNA are heavily modified by pseudouridylations and methylations before the U2 snRNA is mature and assembled into functional snRNPs [13]. These modifications were suggested to be catalysed by proteins as well as guide RNAs [13]. Guide RNAs catalysing U2 snRNA modifications in vertebrates has been found [14-16]. The modifications of U1, U2, U4 and U5 snRNAs are believed to mainly occur in the Cajal bodies and the guide RNAs involved in these base modifications are called small cajal bodies RNAs (scaRNAs) [17, 18]. The scaRNAs sofar identified are structurally similar to the earlier characterized snoRNAs (small nucleolar RNAs) that are found in the nucleoli and function in the modification of rRNA, tRNAs as well as U6 snRNA, reviewed in [19, 20]. These guide RNAs belong to either the C/D or the H/ACA group of RNA involved in catalysis of methylation and pseudouridylation, respectively. The sofar known vertebrate proteins involved in these processes are fibrillarin and dyskerin. Yeast U2 snRNA is less modified. Two enzymes catalysing pseudouridylation of U2 snRNA, PUS1 [21] and PUS7 of yeast has been characterised. These enzymes need no guide RNAs [22].

The human U2A' protein cDNA was cloned in 1989 [23] and a large number of cDNA from various organisms have since been cloned. The U2A' protein is very conserved. In

the 234 amino terminal amino acid sequence there is 84 % identity between salmon and human, and 47 % between salmon and *Arabidopsis thaliana* [24]. A *Saccharomyces cerevisiae* ortholog has 29% identity and 52% similarity [25] and a *Trypanosoma brucei* ortholog has 31% identity and 57 % similarity [7] to the human U2A'protein.

The first vertebrate genomic sequence of the *U2A'* gene was revealed from *Salmo salar* by us [24]. We found that the *U2A'* gene is differently spliced in salmon as well as in human [24] and exon skipping of exon 6 occur. This exon skipping was observed also in a different transcript where exon 2 was skipped giving a truncated protein encoded from only exon 1 and parts of exon 2. In addition, we noted that the exon 6 nucleotide sequence is more conserved than required by the conserved amino acid sequence. These observations suggest that the skipping of exon 6 is done in order to produce a RNA that has some function to the cell. This RNA would contain intron 5, exon 6 and intron 6 of the *U2A'* pre mRNA.

In this paper I have investigated the human *U2A'* intron 5, exon 6 and intron 6 (U2Ai5e6i6) RNA sequence to see whether the sequence can reveal a putative function of the intron RNA. In addition the intron/exon pattern of *U2A'* genes of organisms with so far sequenced genomic sequences was compared. I found conserved sequences of U2Ai5e6i6 that are complementary to the U2 snRNA. Structures of the interactions between U2Ai5e6i6 and the primary transcript of U2 snRNA are presented for human, mouse, chicken and fugu. Interactions at the 5' end suggests a function in the process of localizing the U2 snDNA locus to Cajal bodies, a process that earlier has been proposed to involve an unknown guide RNA. Interactions at the region of the 3' box and

downstream of it suggest a function of the U2Ai5e6i6 in the catalysis of 3' end cleavage of the primary U2 snRNA transcript to yield the pre-transcript including the 3'box. The results suggest U2Ai5e6i6 as the missing link in the 3'end processing of primary U2 snRNA transcripts and the co-localization to Cajal bodies of *RNU2* loci.

Results

Introns 5 and 6 are both found only in vertebrates

In order to find a possible function of *U2A*' intron5 exon6 intron6 RNA I first investigated how conserved the introns of the *U2A'* gene are, and what other organisms possess these introns. The intron/exon pattern of sofar available genomic *U2A'* sequences were revealed and intron positions were indicated in the amino acid alignment (Fig 1). It can be noted that the U2A' protein is very conserved (Fig. 1 and Table I). The amino acid sequence identity between mammals and fish is about 73% and between mammals and the yeast *Schizosaccharomyces pombe* 38 %. There are 8 introns in the *U2A'* genes of vertebrates (Fig 1). These introns are, except for intron number 8 of salmon, found at exactly the same positions in the vertebrates. The other organisms have quite different intron /exon patterns. *Drosophila melanogaster* has got no introns at all. *Caenorhabditis elegans* has got one intron, which is positioned 3 codons upstream from intron 6 of vertebrates. *Arabidopsis thaliana* has got 7 introns. Of these, intron number 1, 4 and 5 are conserved with vertebrate introns 1, 5 and 7, respectively. However, intron 6 of vertebrates is lacking in *A. thaliana*. Interestingly, *S. pombe* has got one intron whose position is conserved with that of intron 1 of vertebrates. This intron thus, is conserved in position in a yeast, a plant and vertebrates,

whereas it does not appear in the fly or the nematode. This may suggest that intron 1 is an ancient intron that has during evolution been lost in flies and nematodes.

The intron sizes differ although intron positions are conserved (Table I). The smallest intron so far identified in the *U2A'* gene is the number 4 of fugu which is only 74 bases, whereas the largest is found in *A. thaliana* and is 9463 bases. Although, intron 4 is the smallest in all vertebrates, except human, it is not possible to rank the intron numbers for sizes. On the contrary the intron sizes are random among the species. In summary, introns 1-7 are conserved in position, but not size, among vertebrates. Although *A. thaliana* has got the intron 5 it has not got intron 6 conserved, thus introns 5 and 6 are both found conserved in position only among vertebrates.

U2Ai5e6i6 contain several conserved sequence cassettes

In order to see if introns 5 and 6 that are conserved in position within vertebrates, are also conserved in nucleotide sequence, the U2A' intron5-exon6-intron6 (U2Ai5e6i6) sequences were aligned (Fig. 2). Exon 6 is identical in human and mouse and between human and salmon the nucleotide sequence identity is 90 % (the amino acid sequence identity is 73 %). The over all identity of the U2Ai5e6i6 is about 69 % between human and mouse, the lengths being 1106 and 1061 bp respectively. The U2Ai5e6i6 of the fishes are the shortest, only 812 bp (salmon) and 722 bp (fugu), whereas the chicken sequence is the longest, 1593 bp. The long chicken sequence is caused by an insertion of about 500 bp within a hairpin loop of intron 6 (Fig 2, 3E and text below). The overall sequence identities between mammals, fish and chicken are quite low. Between salmon and human it is 27 %. However, the 100 bp at the 3'-end of the intron 5 are surprisingly

conserved. Between salmon and human it is 43 % identical, whereas within the mammals the identity is over 90 %. Although the overall sequence similarities are quite low there are several conserved sequence cassettes found in U2Ai5e6i6 (Fig 2).

Interactions between U2Ai5e6i6 and U2 snRNA

Intron 6 of U2Ai5e6i6 contains 3 sequence cassettes (light orange, red and orange in Fig 2) that are 100 % identical in the vertebrates, except fugu that lacks the light orange one. In addition there are conserved palindromes in between these sequence cassettes (Fig 2). This high sequence conservation within an intron suggests some specific function of these sequences. I therefore searched for complementary sequences in the human genome, and indeed found complementary sequences in the U2 snRNA. These cassettes are 100 % complementary in 7 + 12 + 7 base pairs of the 5' end of the U2 snRNA (Fig 3 and 4). It is well known that complementary RNA-RNA sequences may form basepairs and RNA helices. This suggests that RNA double helices may form between U2Ai5e6i6 and U2 snRNA. Detailed investigation of the human U2 snRNA and U2Ai5e6i6 RNA sequences identified 7 additional complementary sequence cassettes (pink, light pink, violet, blue-green, turquoise, beige and pale-green, Fig 2). Of these, the violet sequence of exon 6 is highly conserved, the light pink, blue-green and turquoise are conserved only among mammals, and the last ones exist only in human.

Intra-species co-evolved complementary sequence cassettes of U2Ai5e6i6 and U2 snRNA

Since the human sequence cassettes of U2Ai5e6i6 are complementary to human U2 snRNA it is interesting to investigate sequence complementarities between U2Ai5e6i6

and U2 snRNAs of other organisms, as well. Therefore the U2 snRNA sequences of various species were aligned (Fig. 4). The U2 snRNA sequence of salmon is not yet available, however, two zebrafish U2 snRNA sequences identical in the mRNA region as well as a fugu U2 snRNA sequence was found by Blast search to the zebrafish and fugu genomic databases (Fig. 4). It can be seen in Fig 4 that the orange, red, light orange and lavender sequence cassettes of U2Ai5e6i6 have got conserved complementary sequences in the U2 snRNAs of vertebrates. The blue-green and turquoise sequences that are less conserved between species nevertheless have perfect complementarities between their intra species U2Ai5e6i6 and U2 snRNAs. The high degree of conservation of intraspecies complementarities indicates that these RNAs have co-evolved to retain the complementarity. This in turn, shows that the U2Ai5e6i6 interaction with U2 snRNA is a conserved phenomenon, indicating a functional importance. The lack of conservation in the sequence cassettes in non-vertebrates is in agreement with only vertebrates possessing introns 5 and 6 (Fig. 1).

A predicted large RNA-RNA structure of U2Ai5e6i6 and the primary U2 snRNA transcript

In addition to conserved sequence cassettes with complementarities between U2Ai5e6i6 and U2 snRNA there are conserved sequence cassettes of U2Ai5e6i6 that are complementary to other parts of itself and may therefore form hairpin loops of various lengths. All these helices and hairpin loops that may form suggest that a large structure forms between the U2Ai5e6i6 and the U2 snRNA primary transcript, and a predicted, possible structure of this interaction is shown in figure 3. Except for palindromes that are all coloured yellow each sequence cassette is marked with “its own” colour that is

the same in figure 2 and 3. Sequences complementary to each other and that therefore may form helices are marked with the same colour. Consequently the complementary sequences of U2 snRNA are marked with the same colour as their counterparts of U2Ai5e6i6 (Fig 4).

A large hairpin loop in the middle of the structure

In intron 6 two complementary sequences may form a helix (violet) of 19 base pairs (in the human structure) that may loop out the major part of intron 6 (Fig 2 and 3). The loops differ quite a lot between species, both in sequence and in lengths, e. g. the chicken loop have an insertion of about 500 nucleotides compared to the mammals (Fig 2 and 3E). The nucleotide sequence of this helix is partly conserved between mammals and chicken, but is different in the fishes (Fig 2). However, the intron 6 of salmon and fugu have got a similar helix of complementary base pairs, and at least that of salmon is found at similar positions. The fugu hairpin loop differs somewhat. The downstream part of the helix is found closer to the red helix and the light orange helix is completely lost (Fig 2). The upstream part is found closer to the exon 6 and the green helix found in mammals and chicken is lost. In addition to the large hairpin loop, there are palindrome sequences that also can form hairpin loops (coloured yellow in Fig 2 and 3). Two of these are found between the conserved sequence cassettes of intron 6 in the mammals. (Fig 2). These palindromes, however, are not found in the fish sequences which may be explained by the lengths of the fish sequences between the complementary cassettes being shorter indicating that stabilizing hairpin loops in between them are not needed. In summary the hairpin loops are likely to contribute to a conserved 3D structure (Fig 3).

RNA helices surrounding the site of 3'end cleavage of the primary transcript of U2 snRNA suggest a catalytic function of U2Ai5e6i6

The sequence cassette found in the middle of exon 6 (lavender), is complementary to 8 bases of the very 3' end of the mature U2 snRNA and in addition one base of the 3' box which is cleaved off during maturation of the U2 snRNA (Fig 2 and 3). In addition, one sequence cassette of intron 5 (coloured turquoise in Fig 2 and 3) is complementary to 8 bases of the 3' box of pre-U2 snRNA. Complementary bases are also found downstream of the 3' box (Fig. 3). These sequence complementarities right upstream, within and downstream of the 3' box of U2 snRNA strongly suggest a function of U2Ai5e6i6 in the 3' end processing of the primary U2 snRNA transcript, in which the 3' end is cleaved off right downstream of the 3' box.

The interactions of U2Ai5e6i6 RNA and U2 snRNA are conserved among vertebrates

To investigate further the putatively conserved interaction between U2Ai5e6i6 and U2 snRNA I drew structures of the interactions also for mouse, chicken and fugu (Fig 3D, 3E and 3F). Comparison of the structures of human, mouse and chicken reveal that the overall structure is very conserved (Fig 3). To note is the central long helix (violet) forming a hairpin loop, the three conserved helices (orange, red and light orange) at the 5' end of the U2 snRNAs as well as the hairpins in between them (yellow) and the green intra molecular helix of the U2Ai5e6i6 RNAs which all are conserved and central to the structure. Right below the central hairpin loop (Fig 3), all four species have the blue-green helix, which holds the central part of the U2 snRNA to the U2Ai5e6i6 in a pseudo knot structure. Additionally, in all species, except fugu, two hairpins (yellow)

are flanking this blue-green helix. The hairpin to the left of the blue-green helix forms an additional pseudo knot by base pairings to the central large loop in both human and mouse. In chicken a similar pseudo knot is instead formed by base pairings between the large central loop and the sequence right beside the hairpin (Fig 3). The helices right upstream of the 3' end boxes of the U2 snRNA (lavender and turquoise) are both conserved. In addition, the green intra molecular helix of the U2Ai5e6i6 is conserved in mammals and chicken. That green helix are not conserved in sequence but in position in the structure, it is found 5-12 nucleotides upstream (to the right in Fig 3) of the violet large central helix and it is in all three organisms quite close to the turquoise helix and the nucleotides where the phosphodiester bond breakage will take place. This latter proximity is solved in different ways in the three organisms, in human the pink helix right beside the green helix keeps the green helix close to the turquoise one, whereas in mouse and chicken the green helix is kept close to the turquoise simply by the stretch of nucleotides in between them being very short and in addition hairpins can form to make the distance even shorter (Fig 3A, D and E). The distance, in nucleotides, from the turquoise helix to the central part of the structure, *i e*. the hairpin (yellow) to the right of the blue-green helix, is also conserved. It is 35 and 36 nucleotides in human and mouse, respectively, whereas in chicken it is only 23 nucleotides (Fig 3). The structure of fugu differs somewhat from the structures of mammals and chicken (Fig 3F). However, the main features of the structure are also found in fugu. The highly conserved red helix and the dark orange at the 5' end of the U2 snRNA are there, although the light orange helix is missing. The violet large helix of the hairpin loop is found much closer to the red helix in fugu compared to other organisms. The lavender and turquoise helices at the 3' end of U2 snRNA are found also in fugu as is the blue-green helix that holds the central

part of U2 snRNA to U2Ai5e6i6. Also in fugu, the central hairpin loop is kept in close contact to the blue-green helix holding the U2 snRNA, however, the helix doing this is formed at the right side of the blue-green helix and partly unwinds the blue-green helix as well as the yellow hairpin at the right. Similarly as in chicken fugu may form a helix between the 3'end of the U2 snRNA and the central loop (Fig 3).

Flexibility of the structure also suggests catalytic function

Right downstream of the 3' end of U2 snRNA there are ambiguities in possible helix formations. The pink helix may grow in length in the upstream direction of the U2 snRNA and towards the turquoise helix as indicated in Fig 3A and B by pink bases. As a consequence of the pink helix growing, the turquoise helix as well as the green intra molecular helix of U2Ai5e6i6 must partly unwind (Fig 3A, B and C). In addition, a second new helix may form right downstream of the pink one related to the U2 snRNA. The nucleotides of this helix are coloured blue in Fig 2, 3 and 4. The formation of this blue helix partly unwinds the green helix (Fig 3A, B and C) at the opposite side from the pink helix, leaving only two base pairs of the green helix. The formations of the pink and blue helices and the simultaneous disruption of the green helix suggests a three dimensional change of the structure. This structural rearrangement putatively plays a role in the breakage of the phosphodiester bond right downstream of the 3' box of the primary U2 snRNA transcript and would thus take place in a small loop of the pink helix (Fig 3A, B, C).

The sequence cassette found most to the 3' end of exon 6 in mammals (coloured light pink in Fig 2 and 3) is 8 base pairs long and is complementary to the orange cassette at

the very 5' end of the U2 snRNA. This U2 snRNA sequence is also complementary to the cassette found most to the 3'end of intron 6 (orange) of U2A'. Interestingly, these two cassettes (orange and light pink) of U2Ai5e6i6 are conserved between mammals (Fig 2 and 3). It is highly possible that the ambiguity of the light pink and orange helix formations found at the 5' end of the U2 snRNA are explained by this structural rearrangement and catalysis. Implicating that the pink helix may form while the orange breaks and conformational change takes place.

It is conceivable that the helices found further downstream at the 3' end of the primary U2 snRNA transcript are of stabilising importance for the structure (Fig 3A). The most proximal of them are only 69 bases away from the turquoise helix and the more distal helix will be only 50 bases away from the blue helix. The U2 snRNA sequences within and downstream of the 3'end box are not very conserved (Fig 4). There are 5 human genomic sequences available and sequence alignment of these shows that the 3' box can differ in 1 to 2 positions (Fig 4b). In addition the 3'end right upstream can differ by an A or a C (uRNA database). These differing positions of human sequences are located close to the breakage point of the primary U2 snRNA transcript (see the yellow explosion sign in Fig 3A and the arrow in Fig 3B and 3C). This sequence diversity around the breakage point can possibly be explained by a certain amount of flexibility allowed in the loop. However the nucleotides of the pink and blue helices downstream of the breakage point are completely conserved, which further strengthens the hypothesis that U2Ai5e6i6 plays a role in the catalysis of the phosphodiester bond cleavage that takes place right downstream of the 3' box of the U2 snRNA primary transcript.

Discussion

In this paper I have identified sequence complementarities between U2A i5e6i6 RNA and the U2 snRNA. The complementarities are found in the 5' end, in the middle, far downstream as well as in the region of the 3' box of the primary U2 snRNA transcript. Sequence complementarities between RNA molecules indicate RNA guide functions. The U2 snRNA may need guide RNAs in at least two types of processes proposed in the literature so far. One is in RNA modification reactions where nucleotides are pseudouridylated and methylated and the other is during transcription where the *RNU2* locus is colocalized with Cajal bodies in a RNA dependent and supposedly "guide RNA" dependent process.

Mammalian U2 snRNA is known to be heavily base modified [13]. The modified bases are mainly found in the 5' end of the RNA, indicated in Fig 4. The sequence complementarities between the U2Ai5e6i6 RNA and the 5' region of U2 snRNA found in this article, may therefore suggest a role in guiding of base modifications. To date only a few guide RNA's involved in some of these modifications have been isolated. A guide RNA involved in modification of ψ34 and ψ44 has been isolated from *Xenopus laevis* [14]. This H/ACA type of guide RNA is called pugU2-34/44 and has a human homolog called U92 identified [15]. A U93 guide RNA is involved in pseudouridylation of Ψ54 in human [17]. In yeast ψ35 (the homolog to human ψ34) is modified by the PUS7 protein without any guide RNA [22]. These identified guide RNAs for U2 snRNA modification are in their structure similar to H/ACA and C/D box types of

snoRNAs [15, 20]. Such structures cannot be found in the U2Ai5e6i6 RNA making it unlikely that its function is as guide RNA for base modification of U2 snRNA.

The second process involving U2 snRNA that may use guide RNA is its transcription occurring close to the Cajal bodies (CBs). The *RNU2* locus containing tandemly repeated U2RNA genes have been found to co-localise with CBs in a process whose frequency is dependent on transcription [10-12]. It was found that the frequency of co-localisation is correlated to nascent RNA transcript [10] and specifically the coding region [11]. These authors have proposed that a guide RNA may play a role in the interaction between the *RNU2* locus and the CB. I propose that the U2A'i5e6i6 is a good candidate as the guide RNA base pairing to nascent U2 snRNA and regulating its association to CBs. Frey and Matera [11], in addition, found that the *RNU2* - CB association requires U2 snRNPs. This was shown by using an antisense probe annealing to the 5'end of the U2 snRNA to target degradation of the U2 snRNA. It was found that cells microinjected with the antisense probe had a decreased frequency of *RNU2* - CB co-localization. Interestingly, the antisense probe used in that experiment is complementary to 11 of 12 bases of the red helix interaction between U2snRNA and U2Ai5e6i6 presented here, and would therefore disrupt the red helix. Consequently, an additional interpretation of the experiment of Frey and Matera is that the co-localization of *RNU2* and CBs is dependent on the red-box interaction between U2Ai5e6i6 and U2 snRNA.

What is the purpose of co-localization of CBs and *RNU2* loci and U2 RNA transcription? Frey and Matera [11] proposed that the pre U2 RNA joins export proteins

and assembles in export complexes within the CBs. Smith and Lawrence [12] indeed have shown that all CBs contain pre U2 RNAs. However, they also showed that no CBs contain the long primary transcript but this transcript appears outside the RNU2 DNA that was found in close contact to the surface of CBs. It has also been shown that the formation of the 3'end during transcription requires in addition to the specific promoter and a phosphorylated CTD (C-terminal domain) of polymerase II, also the 3' box [8, 9]. The U2Ai5e6i6 RNA may contribute a missing link in the explanation of these observations. Since U2Ai5e6i6 may form RNA double helices with the region of the 3' box of the U2 snRNA primary transcript as well as downstream thereof, I predict that it functions as a "guide RNA" in the cleavage process of the primary transcript to form the shorter pre U2RNA which includes the 3' box, but excludes the long 3' end of the primary transcript. This 3' end cleavage process may occur at the periphery of the CBs subsequently to transcription so that the pre U2RNAs can easily enter the CBs.

Frey and Matera [11] proposed that one additional reason of co-localisation of *RNU2* locus to CBs is autoregulation of transcription of U2 RNA by help of mature U2 RNPs that are known to accumulate in the CBs.

I predict that the U2Ai5e6i6 RNA is a link of autoregulation of U2 snRNP production. Assuming that splicing efficiency is dependent on amount of U2 snRNPs, the concentration of U2Ai5e6i6 constitutes a measure of splicing efficiency and thus amount of U2 snRNPs. When the amount of U2 snRNP is low the efficiency of the splicing machinery is less and the proper splicing of intron 5 and intron 6 of U2A' messenger fails and instead exon 6 skipping occurs and only one intron is achieved, the

U2Ai5e6i6. The U2Ai5e6i6 RNA then base pairs to the nascent U2 snRNA transcribed from the *RNU2* locus and guides (probably in association with proteins) the association to the CBs and the process of 3' cleavage of the primary transcript. When there is plentiful of U2 snRNP less U2 snRNA is needed, splicing of the U2A' gene is more effective and less U2Ai5e6i6 RNA is produced. The production of U2 snRNA would in this way be autoregulated in a feedback mechanism by the U2Ai5e6i6 RNA.

Conclusions

In this publication I have identified conserved sequences complementary between intron RNA of a transcript encoding one U2 snRNP protein, the U2A' protein, and the U2 snRNA. I have drawn putative structures of the RNA-RNA interactions between the intron RNA and U2 snRNA for human, mouse, chicken and fugu. Based on these structures I predict a conserved role of this interaction in the processing of the primary transcript of U2 snRNA and the co-localization of the *RNU2* locus to Cajal bodies. The U2A' intron RNA involved, is alternatively spliced and includes exon 6. The alternative splicing which in this case is exon skipping, can be considered as a measure of the cells splicing efficiency. Using such intron RNA for regulation of production of U2 snRNA, a central component of the splicing machinery, would be an efficient way for cells to autoregulate the splicing machinery. The results presented here and their interpretations exemplify a novel function of an intron autoregulating the complex of its gene's protein product. Such autoregulatory functions of introns may well be a common feature among introns in general.

Methods

Sequences were derived form GenBank; the zebrafish sequence database, uRNA database, Schizosaccharomyces pombe, and the Saccharomyces genome database (SGD) (see web site references). The MacVector program (Oxford Molecular Group, plc) was used for sequence alignments. RNA folding was tested by use of mfold on the Zuker Group homepage (see web site references)

Figure legends

Figure 1. Amino acid sequence alignment of the U2 snRNP specific A' proteins of selected organisms with their genomic sequences resolved. Intron positions of the corresponding genes are high-lighted in red. Above alignment the intron phase is indicated. One amino acid high-lighted means that intron is in phase 1 or phase 2. Two amino acids high-lighted means that intron is in phase 0. Amino acids identical to human are highlighted in yellow. Hsap – *Homo sapiens* (accession number mRNA: X13482, genomic: AC023024), Mmus – *Mus musculus* (F230356, AC124695.41), Rnor – *Rattus norwegicus* (XM_214963, NW_047560), Ssal – *Salmo salar* (AJ004824, AJ004823), Cele – *Caenorhabditis elegans* (NM_062362, AC006661) Dmel – *Drosophila melanogaster* (NM_136471, AC008259) *Arabidopsis thaliana* (BT000143, AC000132), Spom – *Schizosaccharomyces pombe* (SPBC1861.08c).

Figure 2. Nucleotide sequence alignment of intron5-exon6-intron6 of human (*Homo sapiens*, Hs), mouse (*Mus musculus*, Mm), rat (*Rattus norwegicus*, Rn), chicken (*Gallus gallus*, Gg), salmon (*Salmo salar*, ss) and pufferfish (*Fugu rubripes*, Fr) U2 snRNP specific A' protein gene. Base pair numbering starts from the 5'-ends of intron5. The 5'-ends of intron5 are excluded. Exon 6 is indicated by shading in grey. Sequence cassettes complementary to U2 snRNA are indicated by colours and text above the sequence. The same colours are used for complementary sequences in figure 3 and 4. Complementary sequences within intron 6 that form a large hairpin loop are highlighted with violet. Palindromic sequences are indicated by yellow high-lightening. BS means Branch sequence. Sequence accession numbers are: Hs - AC023024; Mm -

AC124695.41; Rn - NW_047560; Gg – Ensemble: WASHUC1:10:18302966:18309751; Ss - AJ004823; Fr - AC096845.

Figure 3. A. Possible interactions between the human U2A intron5exon6intron6 RNA (U2AiRNA) and human U2snRNA. Nucleotide numbering of U2Ai5e6i6 RNA as well as colours indicating complementary sequences are the same as in Figure 2. U2 snRNA is coloured light turquoise. U2AiRNA is coloured in tan, gold and pale yellow (three colours is used in order to better visualize the RNA where it is crossed by itself). Underlining and roman numbers indicate the hairpin loops I, IIA, IIB and III formed of mature U2 snRNA. Nucleotides of the mature U2 snRNA interacting with sm proteins are indicated and underlined. Exon 6 in the U2Ai5e6i6 sequence is shaded light grey. The 3'box of U2 snRNA is boxed. The blue and pink double arrows indicate a possible unwinding of the green helix and subsequent formation of the pink and blue helices. This shift in base pairings may be involved in the cleavage of the phosphodiester bond between the uridine and adenosine at the 3'-end of the 3'end-box of the pre U2 snRNA, indicated by yellow explosion sign. **B**. Elucidation of the putative subsequently formed blue and pink helices. Black lightning bolt point to the breakage point at the end of the 3'-end box of the pre U2 snRNA transcript. **C**. Illustration of the 3D structures and possible conformational change of the structure caused by the intact green helix (to the left) unwinding to form the blue and pink helices (to the right). Note the bending of the U2 snRNA (light turquoise) and the steric hindrance of the U2Ai5e6i6 (yellow part) in the vicinity of the end of the 3'-end box (black arrow). The formation of the extended pink helix and the blue helix probably induces conformational changes of the RNA structure that may be involved in the catalysis of the breakage of the phosphodiester

bond 3' of the 3'-end box of the U2 snRNA pre-transcript, **B and C**. For comparison the interactions between U2Ai5e6i6 and U2 snRNA, in mouse **D**, chicken **E** and fugu **F** are shown. Base parings between the central loop of chicken U2Ai5e6i6 and the 3'-end of the U2 snRNA are indicated in olive green.

Figure 4. Alignment of U2 snRNAs from various organisms. The complementary sequences are coloured in the same colours as in figure 2 and 3. Nucleotide numbering of the human sequence is found two lines above alignment. Base modifications of human and yeast are indicated above the sequences (m - methylation, p – pesudouridylation). Hs –U2 snRNA: *Homo sapiens* (accession number U57614, nucleotides 4882-5125) terminal part is from the genomic sequence, accession no: U57614; Mm - *Mus musculus* (accession number: X07913); Dr – *Danio rerio*, Dr1 (zC239J9, bases 75592-75778), Dr2 (zK85K7); Gg – *Gallus gallus* (uRNA database); Xl – *Xenopus laevis* (x00093); CeA – *Caenorhabditis elegans* variant A (x51372); SC – *Saccharomyces cerevisiae* (M14625); SpB – *Schizosaccharomyces pombe* (M23361).

Tables

Table1. Intron positions, phases and sizes of U2A' (U2 snRNP specific A' protein) genes of the species: Hs – *Homo sapiens*, Mm – *Mus musculus*; Rn – *Rattus norwegicus*; Ss – *Salmo salar*; Ce – *Caenorhabditis elegans*; At – *Arabidopsis thaliana*; Sp – *Schizosaccharomyces pombe*.

Intron			Intron size								
no	Pos (aa)[1]	Phase	Hs	Mm	Rn	Gg	Ss	Fr	Ce	At	Sp
1	28	1	1925	1431	1672	?	162	569	-	100	55
2	77	2	985	1495	1449	680	90	85	-	-	-
3	104	0	4260	5321	4758	826	237	343	-	-	-
4	119	2	646	210	214	96	177	74	-	-	-
5	154	0	615	697	572	529	388	177	-	500	-
6	180	2	413	391	383	985	346	465	-	-	-
7	206	0	666	566	604	388	431	94	-	120	-
8	237	1	3185	2996	2373	293	-	164	-	-	-
Ss	232	1	-	-	-	-	?[2]	-	-	-	-
At2	41	2	-	-	-	-	-	-	-	363	-
At3	84	0	-	-	-	-	-	-	-	100	-
At5	206	0	-	-	-	-	-	-	-	120	-
At6	223	0	-	-	-	-	-	-	-	99	-
At7	248	0	-	-	-	-	-	-	-	9463	-
Ce		1	-	-	-	-		-	1036	-	-

[1]Pos = position, aa means amino acid or codon split by intron or if intron is between codons, the codon before the intron.

[2]Unknown size (Lundin et al., 2000)

References

1. Will CL, Luhrmann R: **Spliceosomal UsnRNP biogenesis, structure and function**. *Curr Opin Cell Biol* 2001, **13**(3):290-301.
2. Nilsen TW: **The spliceosome: the most complex macromolecular machine in the cell?** *Bioessays* 2003, **25**(12):1147-1149.
3. Jurica MS, Moore MJ: **Pre-mRNA splicing: awash in a sea of proteins**. *Mol Cell* 2003, **12**(1):5-14.
4. Valadkhan S, Manley JL: **Splicing-related catalysis by protein-free snRNAs**. *Nature* 2001, **413**(6857):701-707.
5. Valadkhan S, Manley JL: **Characterization of the catalytic activity of U2 and U6 snRNAs**. *Rna* 2003, **9**(7):892-904.
6. Zwieb C: **The uRNA database**. *Nucleic Acids Res* 1997, **25**(1):102-103.
7. Cross M, Wieland B, Palfi Z, Gunzl A, Rothlisberger U, Lahm HW, Bindereif A: **The trans-spliceosomal U2 snRNP protein 40K of Trypanosoma brucei: cloning and analysis of functional domains reveals homology to a mammalian snRNP protein**. *Embo J* 1993, **12**(3):1239-1248.
8. Medlin JE, Uguen P, Taylor A, Bentley DL, Murphy S: **The C-terminal domain of pol II and a DRB-sensitive kinase are required for 3' processing of U2 snRNA**. *Embo J* 2003, **22**(4):925-934.
9. Jacobs EY, Ogiwara I, Weiner AM: **Role of the C-terminal domain of RNA polymerase II in U2 snRNA transcription and 3' processing**. *Mol Cell Biol* 2004, **24**(2):846-855.
10. Frey MR, Bailey AD, Weiner AM, Matera AG: **Association of snRNA genes with coiled bodies is mediated by nascent snRNA transcripts**. *Curr Biol* 1999, **9**(3):126-135.
11. Frey MR, Matera AG: **RNA-mediated interaction of Cajal bodies and U2 snRNA genes**. *J Cell Biol* 2001, **154**(3):499-509.
12. Smith KP, Lawrence JB: **Interactions of U2 gene loci and their nuclear transcripts with Cajal (coiled) bodies: evidence for PreU2 within Cajal bodies**. *Mol Biol Cell* 2000, **11**(9):2987-2998.
13. Yu YT, Shu MD, Steitz JA: **Modifications of U2 snRNA are required for snRNP assembly and pre-mRNA splicing**. *Embo J* 1998, **17**(19):5783-5795.
14. Zhao X, Li ZH, Terns RM, Terns MP, Yu YT: **An H/ACA guide RNA directs U2 pseudouridylation at two different sites in the branchpoint recognition region in Xenopus oocytes**. *Rna* 2002, **8**(12):1515-1525.
15. Darzacq X, Jady BE, Verheggen C, Kiss AM, Bertrand E, Kiss T: **Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNAs**. *Embo J* 2002, **21**(11):2746-2756.
16. Jady BE, Kiss T: **A small nucleolar guide RNA functions both in 2'-O-ribose methylation and pseudouridylation of the U5 spliceosomal RNA**. *Embo J* 2001, **20**(3):541-551.
17. Kiss AM, Jady BE, Darzacq X, Verheggen C, Bertrand E, Kiss T: **A Cajal body-specific pseudouridylation guide RNA is composed of two box H/ACA snoRNA-like domains**. *Nucleic Acids Res* 2002, **30**(21):4643-4649.

18. Jady BE, Darzacq X, Tucker KE, Matera AG, Bertrand E, Kiss T: **Modification of Sm small nuclear RNAs occurs in the nucleoplasmic Cajal body following import from the cytoplasm**. *Embo J* 2003, **22**(8):1878-1888.
19. Ferre-D'Amare AR: **RNA-modifying enzymes**. *Curr Opin Struct Biol* 2003, **13**(1):49-55.
20. Kiss T: **Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs**. *Embo J* 2001, **20**(14):3617-3622.
21. Massenet S, Motorin Y, Lafontaine DL, Hurt EC, Grosjean H, Branlant C: **Pseudouridine mapping in the Saccharomyces cerevisiae spliceosomal U small nuclear RNAs (snRNAs) reveals that pseudouridine synthase pus1p exhibits a dual substrate specificity for U2 snRNA and tRNA**. *Mol Cell Biol* 1999, **19**(3):2142-2154.
22. Ma X, Zhao X, Yu YT: **Pseudouridylation (Psi) of U2 snRNA in S. cerevisiae is catalyzed by an RNA-independent mechanism**. *Embo J* 2003, **22**(8):1889-1897.
23. Sillekens PT, Beijer RP, Habets WJ, van Verooij WJ: **Molecular cloning of the cDNA for the human U2 snRNA-specific A' protein**. *Nucleic Acids Res* 1989, **17**(5):1893-1906.
24. Lundin M, Mikkelsen B, Gudim M, Syed M: **Gene Structure of the U2 snRNP-Specific A' Protein Gene from Salmo salar: Alternative Transcripts Observed**. 2000, **2**(2):204-211.
25. Caspary F, Seraphin B: **The yeast U2A'/U2B complex is required for pre-spliceosome formation**. *Embo J* 1998, **17**(21):6348-6358.

Links:
GenBank [www.ncbi.nlm.nih.gov]
uRNA database [http://psyche.uthct.edu/dbs/uRNADB/uRNADB.html]
Saccharomyces Genome Database [www.yeastgenome.org]
Schizosaccharomyces pombe database [www.sanger.ac.uk/PostGenomics/S_pombe]
Zebrafish genomic database [www.ensemble.org/Danio_rerio]
Fugu genomic database [www.ensemble.org/Fugu_rubripes/]
The Zuker Group homepage [www.bioinfo.rpi.edu/~zukerm/rna/]

Figure 1

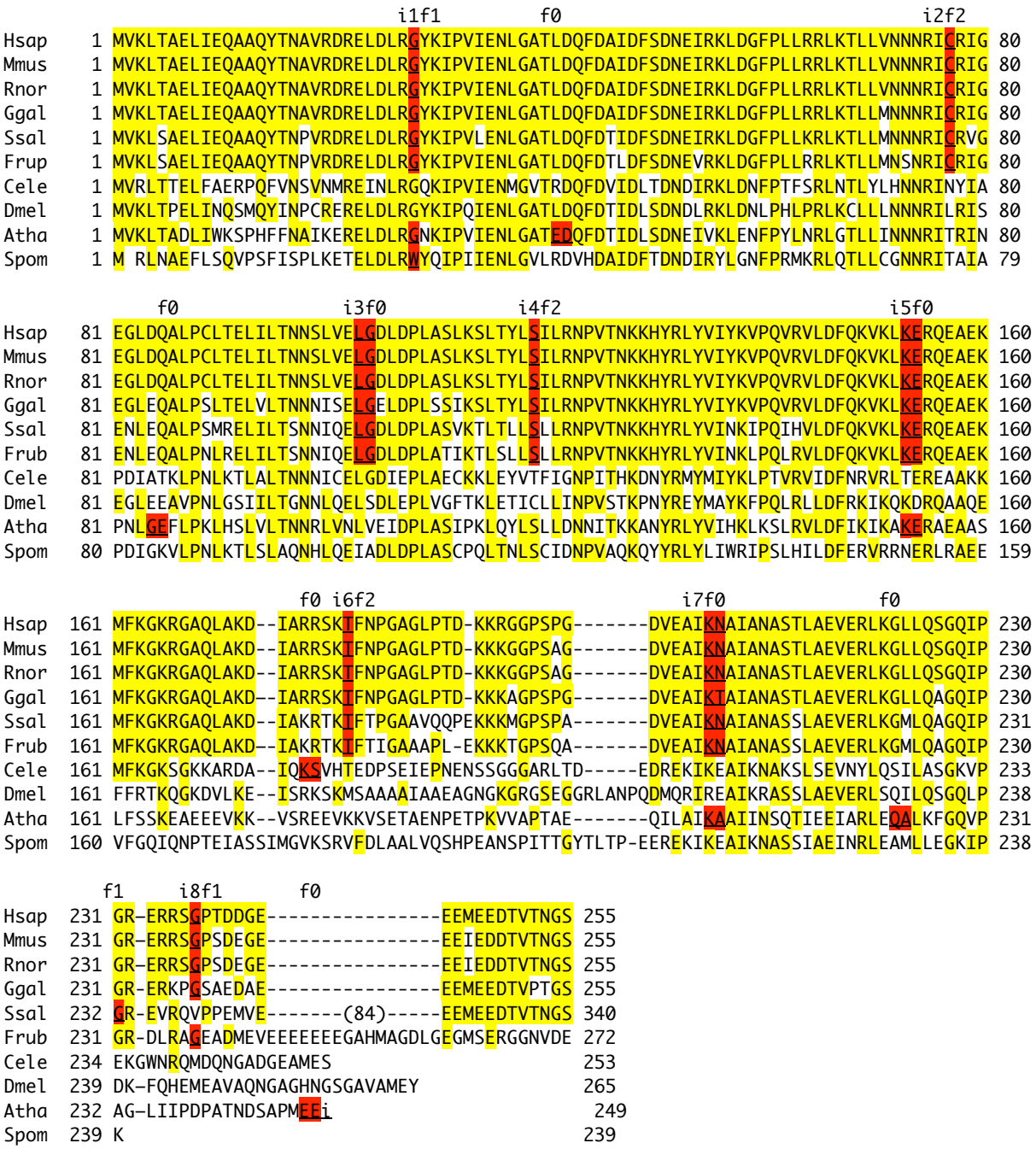
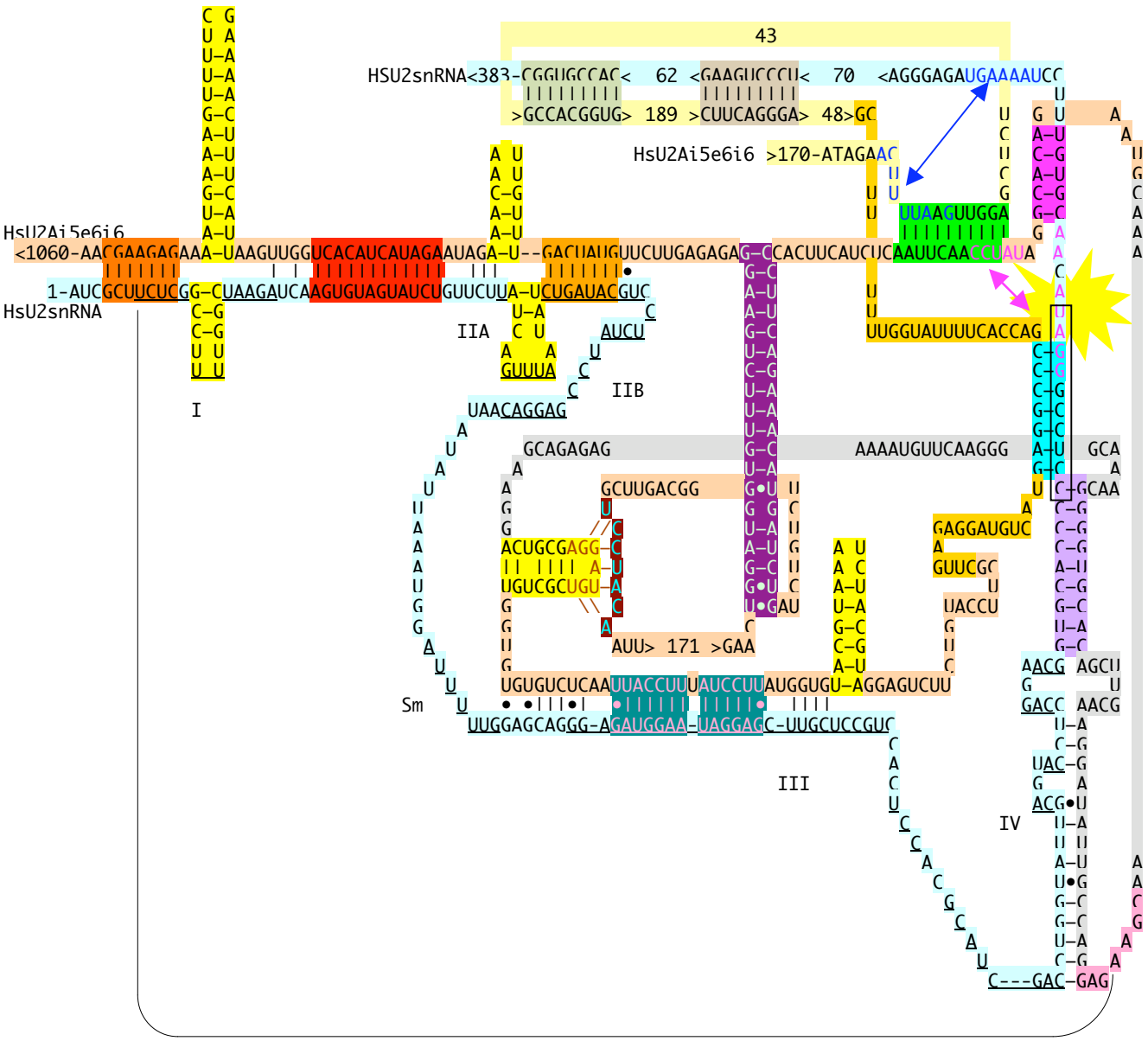
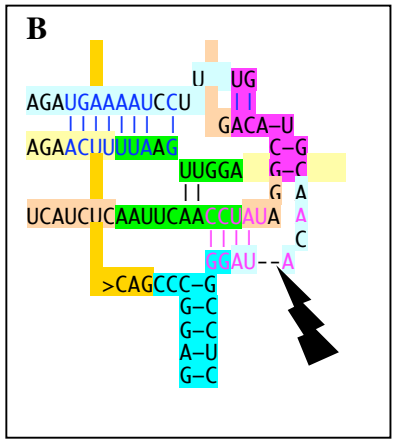

i1f1 f0 i2f2
Hsap 1 MVKLTAELIEQAAQYTNAVRDRELDLRGYKIPVIENLGATLDQFDAIDFSDNEIRKLDGFPLLRRLKTLLVNNNRICRIG 80
Mmus 1 MVKLTAELIEQAAQYTNAVRDRELDLRGYKIPVIENLGATLDQFDAIDFSDNEIRKLDGFPLLRRLKTLLVNNNRICRIG 80
Rnor 1 MVKLTAELIEQAAQYTNAVRDRELDLRGYKIPVIENLGATLDQFDAIDFSDNEIRKLDGFPLLRRLKTLLVNNNRICRIG 80
Ggal 1 MVKLTAELIEQAAQYTNAVRDRELDLRGYKIPVIENLGATLDQFDAIDFSDNEIRKLDGFPLLRRLKTLLMNNNRICRIG 80
Ssal 1 MVKLSAELIEQAAQYTNPVRDRELDLRGYKIPVLENLGATLDQFDTIDFSDNEIRKLDGFPLLKRLKTLLMNNNRICRVG 80
Frup 1 MVKLSAELIEQAAQYTNPVRDRELDLRGYKIPVIENLGATLDQFDTLDFSDNEVRKLDGFPLLRRLKTLLMNSNRICRIG 80
Cele 1 MVRLTTELFAERPQFVNSVNMREINLRGQKIPVIENMGVTRDQFDVIDLTDNDIRKLDNFPTFSRLNTLYLHNNRINYIA 80
Dmel 1 MVKLTPELINQSMQYINPCRERELDLRGYKIPQIENLGATLDQFDTIDLSDNDLRKLDNLPHLPRLKCLLLNNNRILRIS 80
Atha 1 MVKLTADLIWKSPHFFNAIKERELDLRGNKIPVIENLGATEDQFDTIDLSDNEIVKLENFPYLNRLGTLLINNNRITRIN 80
Spom 1 M RLNAEFLSQVPSFISPLKETELDLRWYQIPIIENLGVLRDVHDAIDFTDNDIRYLGNFPRMKRLQTLLCGNNRITAIA 79
f0 i3f0 i4f2 i5f0
Hsap 81 EGLDQALPCLTELILTNNSLVELGDLDPLASLKSLTYLSILRNPVTNKKHYRLYVIYKVPQVRVLDFQKVKLKERQEAEK 160
Mmus 81 EGLDQALPCLTELILTNNSLVELGDLDPLASLKSLTYLSILRNPVTNKKHYRLYVIYKVPQVRVLDFQKVKLKERQEAEK 160
Rnor 81 EGLDQALPCLTELILTNNSLVELGDLDPLASLKSLTYLSILRNPVTNKKHYRLYVIYKVPQVRVLDFQKVKLKERQEAEK 160
Ggal 81 EGLEQALPSLTELVLTNNNISELGELDPLSSIKSLTYLSILRNPVTNKKHYRLYVIYKVPQVRVLDFQKVKLKERQEAEK 160
Ssal 81 ENLEQALPSMRELILTSNNIQELGDLDPLASVKTLTLLSLLRNPVTNKKHYRLYVINKIPQIHVLDFQKVKLKERQEAEK 160
Frub 81 ENLEQALPNLRELILTSNNIQELGDLDPLATIKTLSLLSLLRNPVTNKKHYRLYVINKLPQLRVLDFQKVKLKERQEAEK 160
Cele 81 PDIATKLPNLKTLALTNNNICELGDIEPLAECKKLEYVTFIGNPITHKDNYRMYMIYKLPTVRVIDFNRVRLTEREAAKK 160
Dmel 81 EGLEEAVPNLGSIILTGNNLQELSDLEPLVGFTKLETICLLINPVSTKPNYREYMAYKFPQLRLLDFRKIKQKDRQAAQE 160
Atha 81 PNLGEFLPKLHSLVLTNNRLVNLVEIDPLASIPKLQYLSLLDNNITKKANYRLYVIHKLKSLRVLDFIKIKAKERAEAAS 160
Spom 80 PDIGKVLPNLKTLSLAQNHLQEIADLDPLASCPQLTNLSCIDNPVAQKQYYRLYLIWRIPSLHILDFERVRRNERLRAEE 159
f0 i6f2 i7f0 f0
Hsap 161 MFKGKRGAQLAKD--IARRSKIFNPGAGLPTD-KKRGGPSPG-------DVEAIKNAIANASTLAEVERLKGLLQSGQIP 230
Mmus 161 MFKGKRGAQLAKD--IARRSKIFNPGAGLPTD-KKKGGPSAG-------DVEAIKNAIANASTLAEVERLKGLLQSGQIP 230
Rnor 161 MFKGKRGAQLAKD--IARRSKIFNPGAGLPTD-KKKGGPSAG-------DVEAIKNAIANASTLAEVERLKGLLQSGQIP 230
Ggal 161 MFKGKRGAQLAKD--IARRSKIFNPGAGLPTD-KKKAGPSPG-------DVEAIKTAIANASTLAEVERLKGLLQAGQIP 230
Ssal 161 MFKGKRGAQLAKD--IAKRTKIFTPGAAVQQPEKKKMGPSPA-------DVEAIKNAIANASSLAEVERLKGMLQAGQIP 231
Frub 161 MFKGKRGAQLAKD--IAKRTKIFTIGAAAPL-EKKKTGPSQA-------DVEAIKNAIANASSLAEVERLKGMLQAGQIP 230
Cele 161 MFKGKSGKKARDA--IQKSVHTEDPSEIEPNENSSGGGARLTD-----EDREKIKEAIKNAKSLSEVNYLQSILASGKVP 233
Dmel 161 FFRTKQGKDVLKE--ISRKSKMSAAAAIAAEAGNGKGRGSEGGRLANPQDMQRIREAIKRASSLAEVERLSQILQSGQLP 238
Atha 161 LFSSKEAEEEVKK--VSREEVKKVSETAENPETPKVVAPTAE-------QILAIKAAIINSQTIEEIARLEQALKFGQVP 231
Spom 160 VFGQIQNPTEIASSIMGVKSRVFDLAALVQSHPEANSPITTGYTLTP-EEREKIKEAIKNASSIAEINRLEAMLLEGKIP 238
f1 i8f1 f0
Hsap 231 GR-ERRSGPTDDGE----------------EEMEEDTVTNGS 255
Mmus 231 GR-ERRSGPSDEGE----------------EEIEDDTVTNGS 255
Rnor 231 GR-ERRSGPSDEGE----------------EEIEDDTVTNGS 255
Ggal 231 GR-ERKPGSAEDAE----------------EEMEEDTVPTGS 255
Ssal 232 GR-EVRQVPPEMVE-------(84)-----EEMEEDTVTNGS 340
Frub 231 GR-DLRAGEADMEVEEEEEEEGAHMAGDLGEGMSERGGNVDE 272
Cele 234 EKGWNRQMDQNGADGEAMES 253
Dmel 239 DK-FQHEMEAVAQNGAGHNGSGAVAMEY 265
Atha 232 AG-LIIPDPATNDSAPMEEi 249
Spom 239 K 239

Figure 2


basepairing to U2Ai6
U2 snRNA 3’ terminal
Hs 172 AGAACTTTTAAG----------TTGGAGCTCTCTTATGTAA---TGCTGGCTATTCCGAAGTTCGCAAAGTGCTGAGGCCACGGTGAGCAGTCTTAACAGTGGATTGTCTGCCTAGATACGTTC 282
Mm 157 GTTCTTTTTCAGTGTTTTGTTTTGTTCTCTTAGCAGTTTATGGTAGGGCTTTGTTGGCCATTCCAAAGTTGGCCAAACTTTGTAGTTG---------ACCCCCACAGTGAATTGTCAGCTTAGATGTATTT 278
Rn 169 GTTTTGTTTTA---------------TTCTTAGCAGTTTGTGGTAGGGCTTTGTTGGTCATTCCAAAGTTGGCCAAACTTTGTAGTTG---------AACTACACAGTGAATTGTCTGCTTAGATGGATTT 275
Gg 163 ACACTGTTAAGG---ATGTAATCTACTTCATGATTTTTAATGATGTTTTACTGCC--------------------GTTGTTTGTTTAGTCTGATTC 235
Ss 63 AGGTCCAAAGGG---TGGTCGTCTATAACACAGGTCAGGTTGTATGTTGTATGTT--------------------TTTTTGTAATGATGTAGGCCT 135
U2 snRNA 3’ terminal
INTRON5
Hs 283 CCAGCGGAGAACTTAATGTAACCTACTTTCGTGCAGCTCCGTGTGTTAGGATGATTTTTCCCTTTATT-AGGAATTCATTATTTTATTATTCATTTAATTGCCTTCTTAGTTCCCTGGGATCCCCTCCTTG 412
Mm 279 CCAACTGGGAGCGAACTATAACCCTGCTGTGAGCAGCTCTGCCTGTTGGAATGGCCTTTCCATGTTGG-ACGTGCTTCTTCTTCTCTGCAGTCTCTTGGGAAT---------------GAATCCCTCTTTG 393
Rn 276 CCAATGGGGAGCATACTGTAATCCAGCTGTGAGCAGCCCTGCCTGTTAGAATGGCATTTCCATGTTG--ACGTGCTAATTCTTCTCTG-----------------------------GGACCACTCTTTG 374
Gg 236 TTTTCTGGCTGGCCAGTGTTACTGGTGGTAAGACAACACCCAGAGGAAGGAAGGAGATCACAGCAGAGCCCTCTGCTTTCCCTGTAGAAGTGTGG----------------------G---GTCATTCCTTT 341
Ss 136 TTAGTCAGTGCAAACAGTCACATACCGTTGGACTTGGCCTACGAAGGCCCTAGTTATTAA-----------TCTTTGCAGTGCAATGATAAAG----------------------A---AGCATGCAGAA 230
U2 snRNA 3’terminal
basepairing to U2Ai6
U2snRNA 3’-box
Hs 413 GAGCTAGGAGAATGAACTTTCTTCAGGGAGCTCTACCAAAGAAAGAAGA-CCACTGTGTCTTCTCTTAATTTGAAATTGCTTTTTGGTATTTTCACCAGCCCGGAGTACTGTAGGAGAGTTCGCTTCCAT 540
Mm 394 AAACCAGG---AGGGCGAGCTTTTTAACAACTTTGCCCAAGAATGTTGGACCACTGTGTCTTCT---AATTTAAAGTTGGCTTTTGGTGTTTTTACCAACCCAGAATACTTCAGGAGAATTCTCTTCCAT 516
Rn 375 AAACTGGGGGGAGGGGAGGCTTTTTGAGAACTTTGCCCAAGAATGTTGG-CCAATGTGTCGTCT---AATTTAAAGTTGGCTTTTGGTGTTTTTACCAACCCAGAATACTTCAG-AGAATTCTCTTCCAT 499
Gg 342 TAAGAGCAC---------TAAAGTGATATTCCTCGAGTCCGAGTGTTTGGGCAATGTGTCTGTG-----CTTGTTCTTACATCTGGTCTCCTCGTGCCAGTGTTCACTGGAGATGAGCCAAAGCATCTTT 458
Ss 230 TAGAAGGAAA----TGTATAGGAGGAGAGCTCTGTCTTTCTAGACATGCTAAGAGAGTGCTATA-----TTTGTATGTTTCTCCCAAACATCTGTCCG----------------------------TCCT 322
Fr 1 GTATGTGAGCAGGAAGGATGCCGACCTCAT----------GAGAATGTGTTCCTCCGTGTATCAGAACTGAAACTAGAGTTTTCCATTGCTAA-ACTCGTCCAGCCTGGAGAGTT 114
U2snRNA BS
EXON6
U2snRNA upstream of 3’end box
Hs 541 GTCTTCTGAGGAATGCATCTAAATGCATGTGGTATTCCTATTTCCATTAACTCTGTGTGTGGTGTCGCTGTAGGAGCGTCAGGAAGCAGAGAAAATGTTCAAGGGCAAACGGGGTGCACAGCTTGCAAAG 672
Mm 517 GTTTTCTGAGGAATGCATCTAAATGCATATGGTATTCCTATTTCAATTAACTCTGTGTGTGGTGTTGCTGTAGGAGCGTCAGGAAGCAGAGAAAATGTTCAAGGGCAAACGGGGTGCACAGCTTGCAAAG 648
Rn 500 GTTTTCTGAGGAATGCATCTAAATGCATGTGGTATTCCTATTTCCATTAACTCTGTGTGTGGTGTTGCTGTAGGAGCGTCAGGAAGCAGAGAAAATGTTCAAGGGCAAACGGGGTGCACAGCTTGCAAAG 630
Gg 459 CTTTGCAAAGACAGCAGCTTTGCAAAATGCCGTGTTCTTAAATTAACTTTGT--GTGTGTGGTATCGCTGTAGGAGCGACAGGAGGCAGAGAAAATGTTCAAGGGCAAACGGGGTGCACAGCTTGCAAAG 586
Ss 323 TCTCGTGAAGGTGTCGTTGTCT--GATTGGAAAAGCTCTCACTTTAACT-------GTGATGTATTACCATAGGAGCGTCAGGAGGCGGAGAAAATGTTCAAGGGCAAACGAGGTGCTCAGCTTGCAAAG 444
Fr 115 CTGTCCTTGCCTCCCTCTTGGATCAAAGCTCGCGGTTTTCACCCTGTTTTTTCACATGAATTATATCACCATAGGAACGGCAGGAGGCGGAGAAAATGTTCAAGGGCAAACGAGGTGCTCAGCTTGCAAAG 235
baseparing to U2Ai5
central hairpin loop
Hs 673 GATATTGCCAGGAGAAGCAAAACGTAAGACACGGATATCCAACTTAACTCTACTTCACCTTCAGAAACATGATCTG>-----LOOP (218 BP) ---->TGGATGGTGTTTCTGAAGGAGAGAGTTC---- 979
Mm 649 GATATTGCCAGGAGAAGCAAAACGTAAGATACGGATATCCAACTTAACTCTACTTCACCTTCAGAAACACTGTC>-------LOOP (202 BP) ------>GGTGGTGTTTCTGAAGGAGAGAGTTC---- 939
Rn 631 GATATTGCCAGGAGAAGCAAAACGTAAGATACGGATATCCAACCTAACTCTCCTCC-CCTTCAGAAACACTGTC>-------LOOP (199 BP) ------>GGTGGTGTTTCTGAAGGAGAGAGTTC---- 915
Gg 587 GATATTGCCAGGAGATCAAAAACGTAAGATGAAACAAACACTTC-CACTGGCATTCTCCTTCAGAAACACCA>--------LOOP (798 BP) -------->TGGTGTTTCTGAAGGAGGCAATTCATCA 1460
Ss 445 GATATTGCCAAGCGGACCAAAACGTAAGATGAAAGTCCCCGTCG--GACGGCAGCTCCT>----------LOOP (175 BP)------->AGGAGC-GCCGTCTCAATGGCATTTCCTGT 703
Fr 236 GATATTGCCAAGCGAACCAAAACGTAAGATAAAAAGTCCAC>----------LOOP (354 BP) ---------- 622
U2snRNA
palindrome
U2snRNA 5’ region
palindrome
U2snRNA
BS
Hs 980 TTGTATCAGTTTGTT-----AACAAAGATAAGATACTACACTGG-TTGAATTACTTTCAAAAGCTTTTGAAAGTAA---AAGAGAAGCAACTTCTGTTGGTCTAGGTCTAACAGTTATGTTCT--TCCCCCTCTTAAG 1106
Mm 940 TTGTATCAGTTTGTT-----AACAAAGATAAGATACTACACTGG-TTGAATTACTTTCAAAAGCTTTTGAAAGTAA-----GAGAAGCAGCTTGCCCCTTTCAGAATGTAACAATTATATTCT-----CCCTCCCCAG 1061
Rn 916 TTGTATCAGTTTGTT-----AACAAAGATAAGATACTACACTGG-TTGAATTACTTTCAAAAGCTTTTGAAAGTAA-----GAGAAGCAGCTTGCACAGT-CAGAATATAACAATTATATTCT------CCTCCCCAG 1035
Gg 1461 CTGTATCAGTTTGTT-----AACAAAGATAAGATACTACACTTGATTGAATGAAATGCTTCCAAACCTGGTTTGGAAGTAAGAGAAGCACACACTGCTCCCCGACAGTAACGATACCTCTTTGCTTTTTCCTCTGAAG 1593
Ss 704 -CGTATCAGTTTATTTGACTCAATAAGAAAAGATACTACACTGGTTGAGAGGCTA----------GAGAAGCACCCATCCCCTCCATGTTCTCTCTAACTCTGCTGG--TGTGCGTTCCCAG 812
Fr 623 ------->GTGGACTGTTTGACCTGAGAAAAGATACTACACTGGTTCTAGAGGTCA----------GAGGAGCTGGTGCGGGTCCGTG---------TCACCGACGTGCC---TTTCTTTCAG 722

Figure 3A. HUMAN





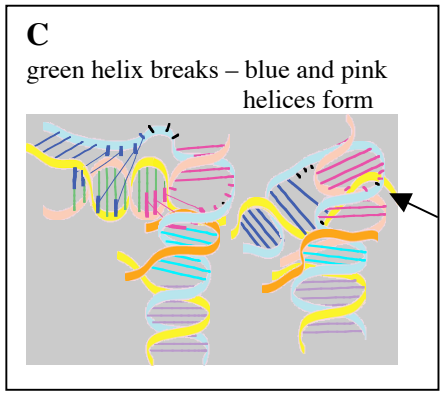


Figure 3D. MOUSE

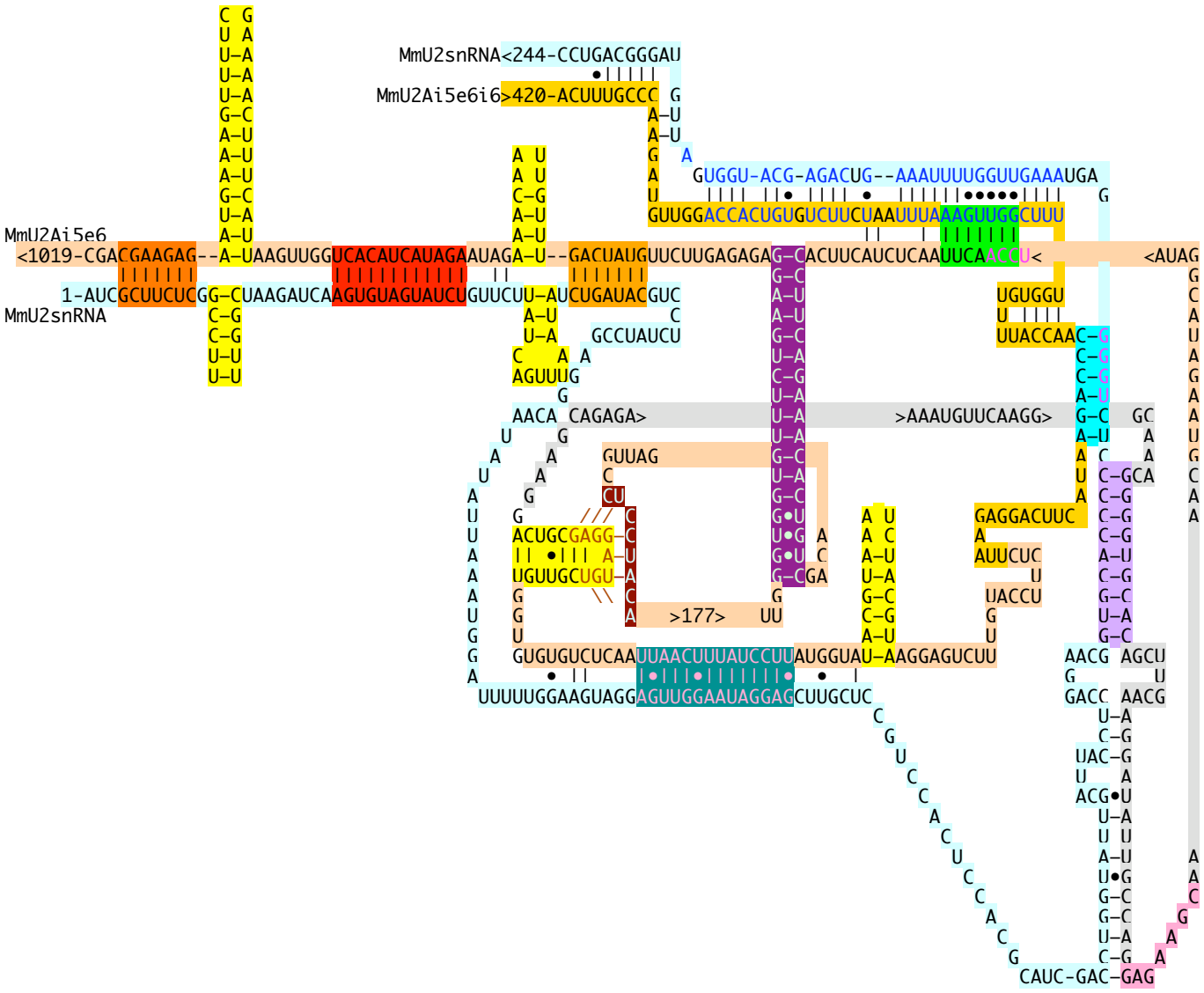


Figure 3E. CHICKEN

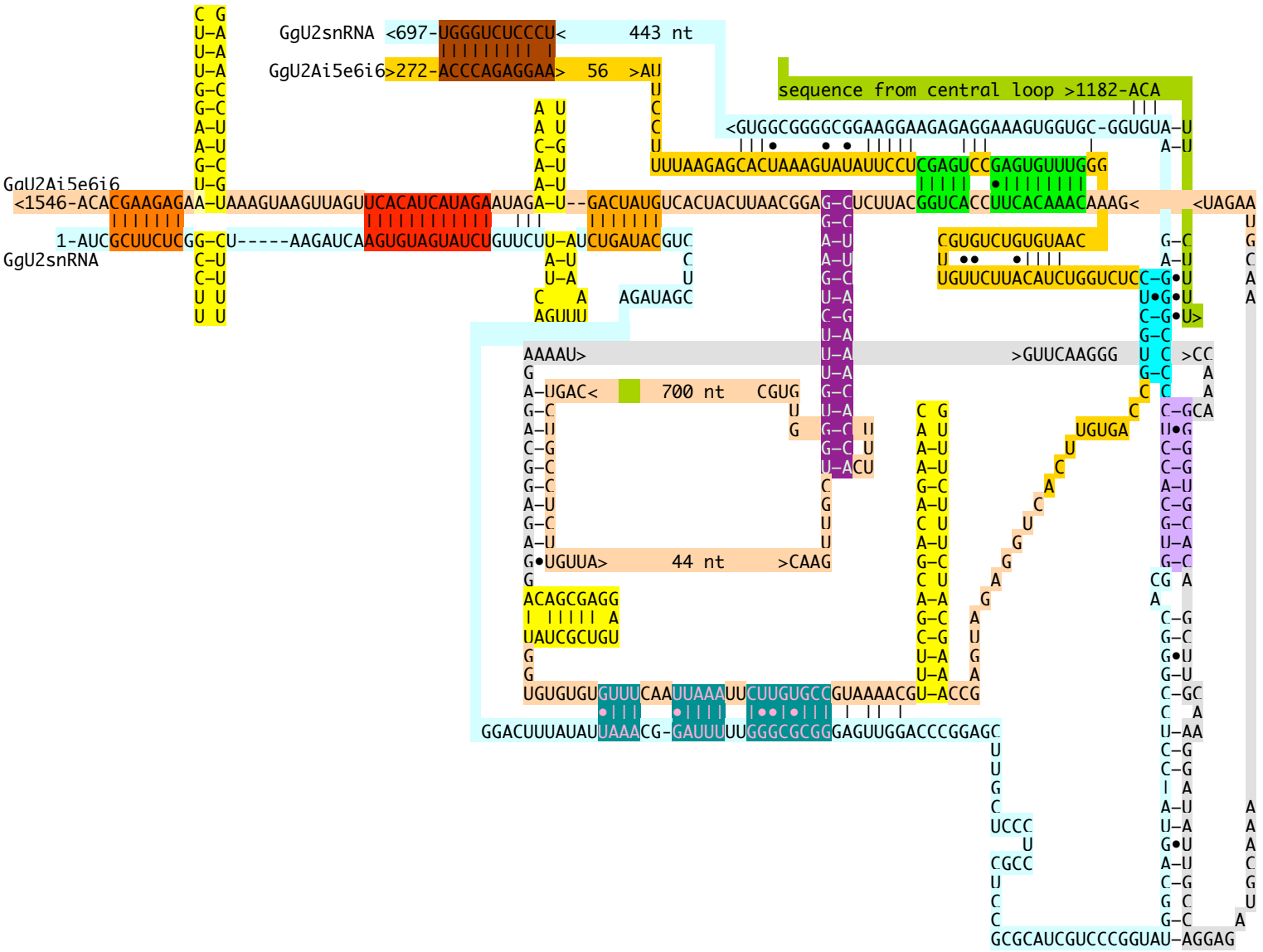


Figure 3F. FUGU

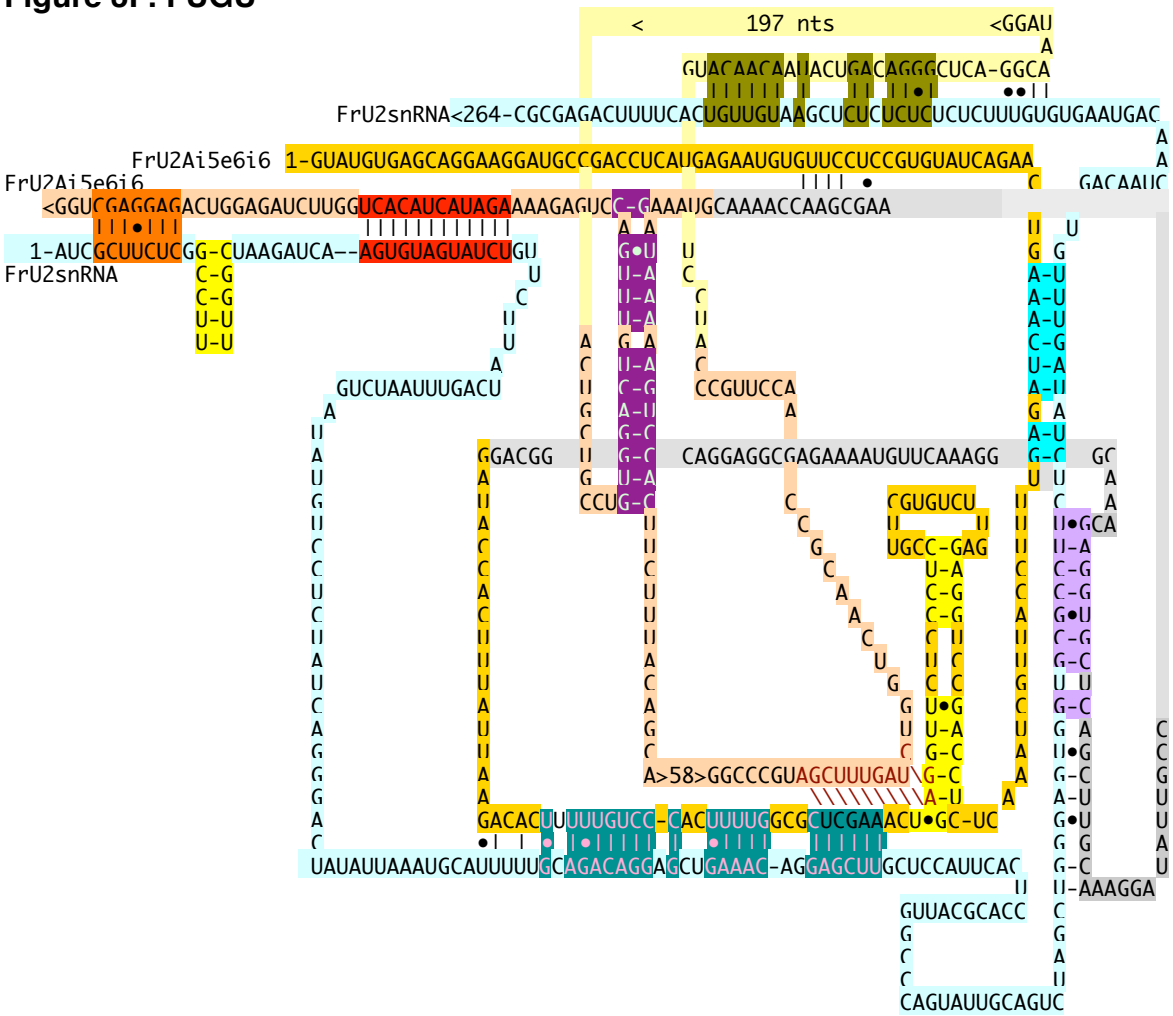
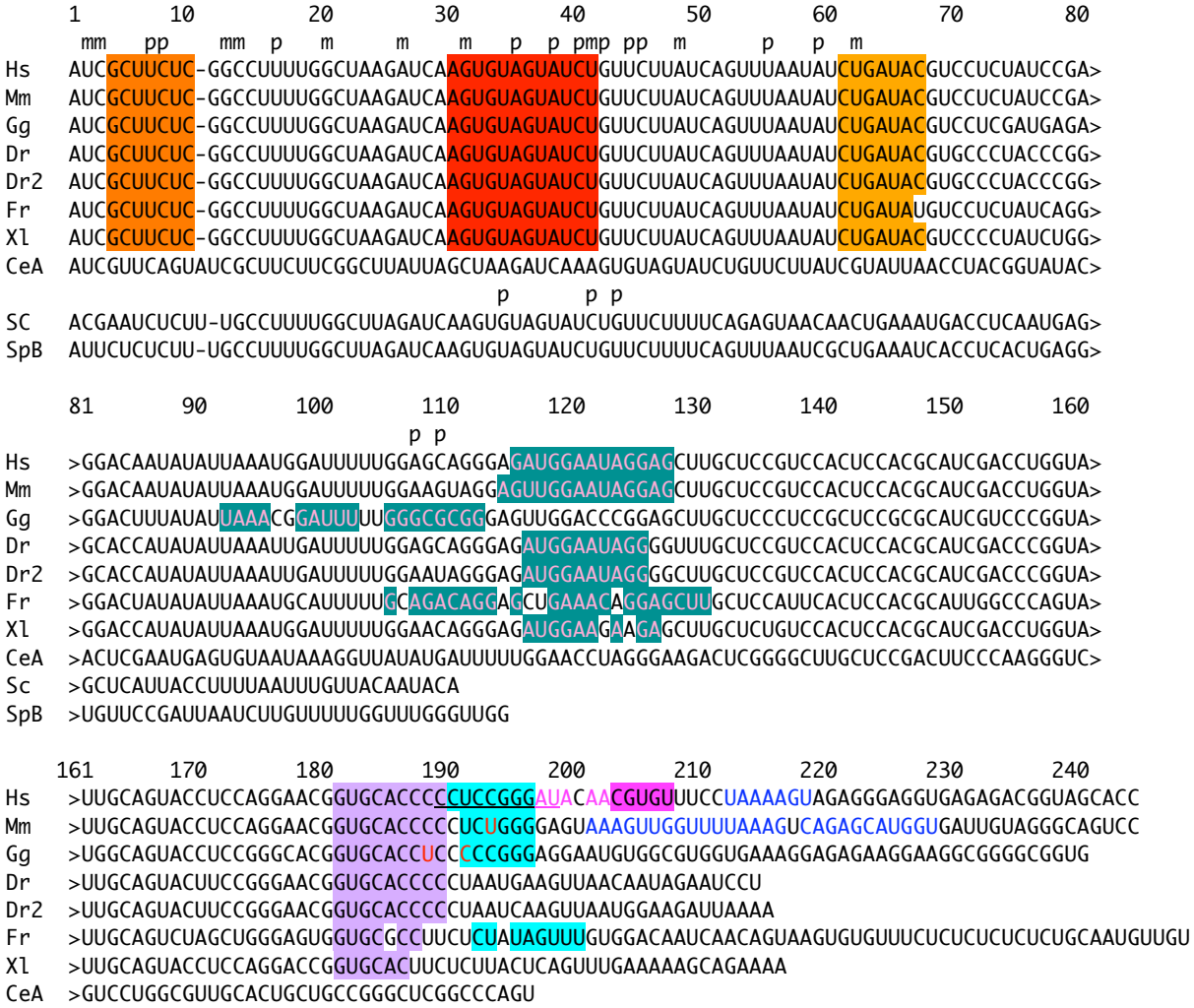
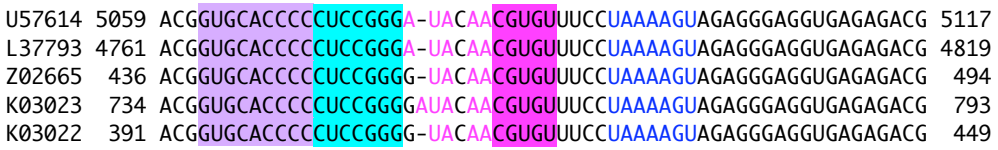


Figure 4A


1 10 20 30 40 50 60 70 80
Hs AUCGCUUCUC-GGCCUUUUGGCUAAGAUCAAGUGUAGUAUCUGUUCUUAUCAGUUUAAUAUCUGAUACGUCCUCUAUCCGA>
Mm AUCGCUUCUC-GGCCUUUUGGCUAAGAUCAAGUGUAGUAUCUGUUCUUAUCAGUUUAAUAUCUGAUACGUCCUCUAUCCGA>
Gg AUCGCUUCUC-GGCCUUUUGGCUAAGAUCAAGUGUAGUAUCUGUUCUUAUCAGUUUAAUAUCUGAUACGUCCUCGAUGAGA>
Dr AUCGCUUCUC-GGCCUUUUGGCUAAGAUCAAGUGUAGUAUCUGUUCUUAUCAGUUUAAUAUCUGAUACGUGCCCUACCCGG>
Dr2 AUCGCUUCUC-GGCCUUUUGGCUAAGAUCAAGUGUAGUAUCUGUUCUUAUCAGUUUAAUAUCUGAUACGUGCCCUACCCGG>
Fr AUCGCUUCUC-GGCCUUUUGGCUAAGAUCAAGUGUAGUAUCUGUUCUUAUCAGUUUAAUAUCUGAUAUGUCCUCUAUCAGG>
Xl AUCGCUUCUC-GGCCUUUUGGCUAAGAUCAAGUGUAGUAUCUGUUCUUAUCAGUUUAAUAUCUGAUACGUCCCCUAUCUGG>
CeA AUCGUUCAGUAUCGCUUCUUCGGCUUAUUAGCUAAGAUCAAAGUGUAGUAUCUGUUCUUAUCGUAUUAACCUACGGUAUAC>
SC ACGAAUCUCUU-UGCCUUUUGGCUUAGAUCAAGUGUAGUAUCUGUUCUUUUCAGAGUAACAACUGAAAUGACCUCAAUGAG>
SpB AUUCUCUCUU-UGCCUUUUGGCUUAGAUCAAGUGUAGUAUCUGUUCUUUUCAGUUUAAUCGCUGAAAUCACCUCACUGAGG>
81 90 100 110 120 130 140 150 160
Hs >GGACAAUAUAUUAAAUGGAUUUUUGGAGCAGGGAGAUGGAAUAGGAGCUUGCUCCGUCCACUCCACGCAUCGACCUGGUA>
Mm >GGACAAUAUAUUAAAUGGAUUUUUGGAAGUAGGAGUUGGAAUAGGAGCUUGCUCCGUCCACUCCACGCAUCGACCUGGUA>
Gg >GGACUUUAUAUUAAACGGAUUUUUGGGCGCGGGAGUUGGACCCGGAGCUUGCUCCCUCCGCUCCGCGCAUCGUCCCGGUA>
Dr >GCACCAUAUAUUAAAUUGAUUUUUGGAGCAGGGAGAUGGAAUAGGGGUUUGCUCCGUCCACUCCACGCAUCGACCCGGUA>
Dr2 >GCACCAUAUAUUAAAUUGAUUUUUGGAAUAGGGAGAUGGAAUAGGGGCUUGCUCCGUCCACUCCACGCAUCGACCCGGUA>
Fr >GGACUAUAUAUUAAAUGCAUUUUUGCAGACAGGAGCUGAAACAGGAGCUUGCUCCAUUCACUCCACGCAUUGGCCCAGUA>
Xl >GGACCAUAUAUUAAAUGGAUUUUUGGAACAGGGAGAUGGAAGAAGAGCUUGCUCUGUCCACUCCACGCAUCGACCUGGUA>
CeA >ACUCGAAUGAGUGUAAUAAAGGUUAUAUGAUUUUUGGAACCUAGGGAAGACUCGGGGCUUGCUCCGACUUCCCAAGGGUC>
Sc >GCUCAUUACCUUUUAAUUUGUUACAAUACA
SpB >UGUUCCGAUUAAUCUUGUUUUUGGUUUGGGUUGG
161 170 180 190 200 210 220 230 240
Hs >UUGCAGUACCUCCAGGAACGGUGCACCCCCUCCGGGAUACAACGUGUUUCCUAAAAGUAGAGGGAGGUGAGAGACGGUAGCACC
Mm >UUGCAGUACCUCCAGGAACGGUGCACCCCCUCUGGGGAGUAAAGUUGGUUUUAAAGUCAGAGCAUGGUGAUUGUAGGGCAGUCC
Gg >UGGCAGUACCUCCGGGCACGGUGCACCUCCCCCGGGAGGAAUGUGGCGUGGUGAAAGGAGAGAAGGAAGGCGGGGCGGUG
Dr >UUGCAGUACUUCCGGGAACGGUGCACCCCCUAAUGAAGUUAACAAUAGAAUCCU
Dr2 >UUGCAGUACUUCCGGGAACGGUGCACCCCUAAUCAAGUUAAUGGAAGAUUAAAA
Fr >UUGCAGUCUAGCUGGGAGUGGUGCGCCUUCUCUAUAGUUUGUGGACAAUCAACAGUAAGUGUGUUUCUCUCUCUCUCUGCAAUGUUGU
Xl >UUGCAGUACCUCCAGGACCGGUGCACUUCUCUUACUCAGUUUGAAAAAGCAGAAAA
CeA >GUCCUGGCGUUGCACUGCUGCCGGGCUCGGCCCAGU

B


U57614 5059 ACGGUGCACCCCCUCCGGGA-UACAACGUGUUUCCUAAAAGUAGAGGGAGGUGAGAGACG 5117
L37793 4761 ACGGUGCACCCCCUCCGGGA-UACAACGUGUUUCCUAAAAGUAGAGGGAGGUGAGAGACG 4819
Z02665 436 ACGGUGCACCCCCUCCGGGG-UACAACGUGUUUCCUAAAAGUAGAGGGAGGUGAGAGACG 494
K03023 734 ACGGUGCACCCCCUCCGGGGAUACAACGUGUUUCCUAAAAGUAGAGGGAGGUGAGAGACG 793
K03022 391 ACGGUGCACCCCCUCCGGGG-UACAACGUGUUUCCUAAAAGUAGAGGGAGGUGAGAGACG 449