THE REGULATION OF U12-DEPENDENT SPlicing 
AND ITS SIGNIFICANCE TO mRNA STABILITY

Elina Niemelä

Institute of Biotechnology
Division of Biochemistry
Department of Biosciences
Faculty of Biological and Environmental Sciences
Integrative Life Science Doctoral Program
University of Helsinki

ACADEMIC DISSERTATION

To be presented for public examination with the permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki, in auditorium 1041 in Biocenter 2 (Viikinkaari 5, Helsinki) on October 10th 2014, at 12 noon.

HELSINKI 2014
“But nature is always more subtle, more intricate, more elegant than what we are able to imagine.”

Carl Sagan
# TABLE OF CONTENTS

List of original publications........................................................................................................ vi
Abbreviations.................................................................................................................................. vi
Summary.......................................................................................................................................... vii
Tiivistelmä......................................................................................................................................... viii

1 Review of the literature ................................................................................................................... 1
   1.1 Introns and exons...................................................................................................................... 1
      1.1.1 Evolutionary costs and benefits of introns ................................................................... 1
      1.1.2 Prevalence and origins of introns ................................................................................. 2
      1.1.3 Self-splicing introns ...................................................................................................... 3
      1.1.4 Splice sites ................................................................................................................... 4
      1.1.5 Intron types in different organisms ............................................................................. 5
   1.2 Spliceosome components ....................................................................................................... 7
      1.2.1 Components of the U2-dependent spliceosome ......................................................... 7
      1.2.2 Components of the U12-dependent spliceosome .................................................... 10
   1.3 Spliceosome assembly and catalysis ...................................................................................... 14
      1.3.1 Splicing catalysis ....................................................................................................... 14
      1.3.2 Assembly of the U2-dependent spliceosome ........................................................... 15
      1.3.3 Intron recognition by the U12-dependent spliceosome ........................................... 16
      1.3.4 Assembly of the U12-dependent spliceosome ........................................................... 19
   1.4 Splice site selection ................................................................................................................. 19
      1.4.1 Exon definition ........................................................................................................... 19
      1.4.2 Commitment complex formation .............................................................................. 20
      1.4.3 Alternative splicing ................................................................................................... 21
      1.4.4 Splicing regulatory elements ................................................................................... 23
      1.4.5 Splicing factors ......................................................................................................... 23
      1.4.6 SnRNPs as processing factors .................................................................................. 25
      1.4.7 Other factors affecting alternative splicing ............................................................ 26
   1.5 Interconnectivity in nuclear RNA-processing pathways ...................................................... 27
      1.5.1 Cotranscriptional processing of pre-mRNA .............................................................. 27
      1.5.2 Cotranscriptionality of splicing ............................................................................... 28
1.5.3 Kinetics of splicing and export ................................................................. 31

1.6 RNA quality control ......................................................................................... 32
  1.6.1 Nuclear quality control ............................................................................ 32
  1.6.2 Exosome ................................................................................................. 33
  1.6.3 Exosome targets ...................................................................................... 33
  1.6.4 Other nuclear degradation pathways ....................................................... 34
  1.6.5 Kinetic proofreading hypothesis .............................................................. 35
  1.6.6 Nonsense mediated decay ...................................................................... 35
  1.6.7 Regulation via NMD ................................................................................ 36

2 Aims of the study .................................................................................................... 37

3 Materials and methods ........................................................................................... 38

4 Results and discussion ............................................................................................. 39
  4.1 The levels of U11/U12 di-snRNP are regulated through a conserved feedback mechanism (I, II) ............................................................................... 39
    4.1.1 Genes encoding U11-48K and U11/U12-65K contain highly conserved sequence elements (I, II) ......................................................... 39
    4.1.2 Both USSE motifs bind U11/U12 di-snRNP (I, II) ............................... 41
    4.1.3 Feedback mechanism relies on alternative splicing and exon definition interactions (I, II) .......................................................... 43
    4.1.4 USSE activation regulates stability and localization of 48K and 65K mRNAs (I)........................................................................ 45
    4.1.5 The function of the USSE and evolutionary implications ..................... 46
  4.2 U12-dependent transcripts are spliced less efficiently and targeted to nuclear degradation (III) .......................................................................... 49
    4.2.1 Exosome inactivation leads to stabilization of intron-retaining transcripts...................................................................................... 49
    4.2.2 U12-type intron retention on global scale.............................................. 49
    4.2.3 U12- and U2-type introns show differential decay rates..................... 51

5 Concluding remarks ................................................................................................ 53

6 Acknowledgements ................................................................................................ 54

7 References .............................................................................................................. 57
LIST OF ORIGINAL PUBLICATIONS

The thesis is based on the following articles, which are referred to in the text by their Roman numerals:


* Equal contribution

The author’s contribution to each publication:

I EHN planned and performed the experiments characterizing the U11-48K alternative splicing regulation and nonsense mediated decay, cellular distribution of U11-48K and U11/U12-65K transcripts, and wrote the article together with the other authors.

II EHN planned and performed the characterization of U11-35K in splicing assays, immunofluorescence, and wrote the manuscript.

III EHN planned and performed the majority of experiments, participated in the analysis of RNA sequencing data, and wrote the article together with the other authors.
### ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>3’ss</td>
<td>3’ splice site</td>
</tr>
<tr>
<td>5’ss</td>
<td>5’ splice site</td>
</tr>
<tr>
<td>AS</td>
<td>alternative splicing</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>BP(S)</td>
<td>branch point (sequence)</td>
</tr>
<tr>
<td>CHX</td>
<td>cycloheximide</td>
</tr>
<tr>
<td>CTD</td>
<td>C-terminal domain (of RNA polymerase II)</td>
</tr>
<tr>
<td>DRB</td>
<td>5,6-dichloro-1-β-D-ribofuranosylbenzimidazole</td>
</tr>
<tr>
<td>EJC</td>
<td>exon junction complex</td>
</tr>
<tr>
<td>EST</td>
<td>expressed sequence tag</td>
</tr>
<tr>
<td>exoRNase</td>
<td>exoribonuclease</td>
</tr>
<tr>
<td>hnRNP</td>
<td>heterogenous nuclear RNP</td>
</tr>
<tr>
<td>miRNA</td>
<td>microRNA</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>ncRNA</td>
<td>noncoding RNA</td>
</tr>
<tr>
<td>NMD</td>
<td>nonsense mediated decay</td>
</tr>
<tr>
<td>nt</td>
<td>nucleotide</td>
</tr>
<tr>
<td>PF</td>
<td>perichromatin fibril</td>
</tr>
<tr>
<td>Pol II</td>
<td>RNA polymerase II</td>
</tr>
<tr>
<td>PPT</td>
<td>polypyrimidine tract</td>
</tr>
<tr>
<td>pre-mRNA</td>
<td>precursor mRNA</td>
</tr>
<tr>
<td>PTC</td>
<td>premature termination codon</td>
</tr>
<tr>
<td>qPCR</td>
<td>quantitative PCR</td>
</tr>
<tr>
<td>RNAi</td>
<td>RNA interference</td>
</tr>
<tr>
<td>RRM</td>
<td>RNA recognition motif</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>RS</td>
<td>arginine-serine rich (domain)</td>
</tr>
<tr>
<td>snoRNA</td>
<td>small nucleolar RNA</td>
</tr>
<tr>
<td>snRNA</td>
<td>small nuclear RNA</td>
</tr>
<tr>
<td>snRNP</td>
<td>small nuclear ribonucleoprotein</td>
</tr>
<tr>
<td>SR</td>
<td>serine-arginine rich (protein)</td>
</tr>
<tr>
<td>SRE</td>
<td>splicing regulatory element</td>
</tr>
<tr>
<td>UTR</td>
<td>untranslated region</td>
</tr>
</tbody>
</table>
SUMMARY

The removal of noncoding sequences, or introns, from eukaryotic messenger RNAs is an essential step in the expression of genetic information. Two types of introns are removed by dedicated spliceosomes. The majority of introns are spliced by the U2-dependent spliceosome, whereas a small fraction of introns, less than 0.5%, are removed by a dedicated U12-dependent spliceosome. Although the two splicing pathways are highly similar, differences exist in the initial step of intron recognition. The rationality of maintaining parallel spliceosomes has remained unclear and is thought to stem from the suggested regulatory functions of the minor type introns. The U12-dependent spliceosome removes introns at a slower rate than the major spliceosome, and therefore the minor introns may constitute a rate-limiting step in the processing of the mRNA.

In this work, I have investigated the regulation of the minor spliceosome via a negative feedback loop that relies on alternative splicing and nonsense mediated decay. Pre-mRNAs encoding U11-48K and U11/U12-65K proteins of the minor intron recognition particle, U11/U12 small nuclear RNP, are alternatively spliced. Alternative splice site usage is activated upon binding of the U11/U12 di-snRNP on regulatory motifs resembling bona fide spliceosomal 5’ splice sites. The alternative isoforms contain premature termination codons leading to destabilization of the transcript in question, or display differential cellular localization. I have further characterized the splicing regulatory element and the role of U11-35K protein required for its activation. The findings point to the existence of a considerable selection pressure maintaining the important motifs of the regulatory element, as both the motifs and their location relative to each other are highly conserved in distant organisms and can be found both in animals and plants, highlighting the importance of regulating the intron recognition step of the U12-dependent splicing pathway.

The rate-limiting characteristics of U12-type introns are well known but earlier works have investigated only a small number of introns. In the final part of this work, RNA sequencing was performed and the results confirm that U12-type introns are less efficiently removed on a transcriptome-wide scale. I also investigated a hypothesis suggesting that slowly processed mRNAs are targeted by nuclear decay. Following the inactivation of RNA exosome components, transcripts with unspliced U12-type introns are stabilized while U2-type introns in the same transcripts remain unaffected. This supports the results from previous work on the delayed processing of U12-type introns and indicates a mechanism for their role in the regulation of eukaryotic gene expression.


1 REVIEW OF THE LITERATURE

1.1 INTRONS AND EXONS

1.1.1 Evolutionary costs and benefits of introns

The majority of genes in eukaryotic genomes are encoded in a discontinuous manner, that is, coding sequences are found as fragments dispersed within long stretches of noncoding sequence. Expression of the correct gene products requires accurate joining of the protein-coding sequences known as exons, and the removal of noncoding intron sequences. Considered “junk DNA” for decades, it has now become apparent that introns play many roles in eukaryotic gene expression (reviewed in Le Hir et al., 2003). Removal of intronic sequences, a process called RNA splicing, is one of the central steps in eukaryotic gene expression pathway. Intron removal is highly regulated, mainly through a layer of information which is located both in the intronic and exonic sequences and often overlaps with protein-coding information needed for translation.

The presence of introns can be beneficial for cells or organisms in many ways. In a process called alternative splicing (see section 1.4.3) a single gene can code for multiple gene products, thus increasing protein diversity and therefore the coding capacity of eukaryotic genomes. Exon shuffling and duplication/deletion events accelerate evolution by producing raw material for evolutionary selection processes. Intron sequences can provide sites for and increase the likelihood of recombination processes, but they can also buffer imprecisions in recombination, thus facilitating the evolutionary emergence of new exons or even entire genes (reviewed by Patthy, 1999).

Introns also have more immediate effects on many steps of gene expression. Their presence enables transcriptional regulation because, as regions free of selection pressures imposed by protein coding information, introns are a common location for transcriptional enhancer (or suppressor) elements (Rossi and de Crombrugghe, 1987; Stergachis et al., 2013). More directly, introns stimulate RNA polymerase II (Pol II) initiation and elongation rate (Brinster et al., 1988; Fong and Zhou, 2001; Furger et al., 2002; Kwek et al., 2002). Also, pre-mRNA splicing is linked to subsequent steps of gene expression as a consequence of exon junction complexes (EJC) deposited on the mRNA during splicing, which stimulates many subsequent processes such as nuclear export, mRNA translation, and mRNA quality control (Le Hir et al., 2001; Nott et al., 2004).

Introns contain many RNA genes that do not code for proteins (noncoding RNA or ncRNA), and therefore not all excised introns are completely discarded after splicing (Rearick et al., 2011). The classification to coding and non-coding sequences is, in fact,
an oversimplification as it has become evident that for a growing number of ncRNA genes RNA is the final gene product. Furthermore, in many instances introns are processed separately to give rise to stable RNA products such as microRNAs and small nucleolar RNAs (miRNAs and snoRNAs; Kiss and Filipowicz, 1995; Rodriguez et al., 2004). There are also extreme cases where introns code for stable RNA products while the exonic sequences do not have any apparent function (Tycowski et al., 1996). It has also been argued that the complexity of eukaryotic gene expression is an intricate strategy to ensure the genuineness of expression, that is, to defend against genomic parasites, such as transposable elements and viruses which do not necessarily participate in such complex processes (Madhani, 2013).

On the other hand, the transcription of such additional lengths of RNA comes with a cost. This is evident from the genomic organization of housekeeping genes and highly expressed genes which have shorter and fewer introns (Castillo-Davis et al., 2002; Eisenberg and Levanon, 2003). This may be because the presence of introns causes a delay in the synthesis of a transcript, both as an increased transcription time and the time spent in the processing steps. Such delays can sometimes be beneficial, for example in controlling the timing of mRNA expression and even in establishing oscillating expression patterns (Takashima et al., 2011). Another drawback involving introns and splicing is that owing to its complexity, the system is sensitive to mutations within the splicing factors or within the necessary sequence elements on pre-mRNAs, which can cause serious misregulation of gene expression and lead to various diseases, particularly in neurological diseases and cancer (Ward and Cooper, 2010; Abdel-Wahab and Levine, 2011). And finally, while the eukaryotic gene expression may have developed to a high degree of complexity in order to avoid the expression of parasitic sequences, certain viruses exploit this complexity to their own advantage to produce highly regulated gene expression programs (for example HIV; Stoltzfus and Madsen, 2006).

1.1.2 Prevalence and origins of introns

Intron density and size varies considerably between different eukaryotic genomes. Unicellular eukaryotes such as the baker’s yeast *Saccharomyces cerevisiae* tend to have few introns. Only 4% of *S. cerevisiae* genes have introns, mostly only one per gene with a short average length (256 bp, range 50-1000 bp; Kupfer et al., 2004). In humans, on the other hand, mean intron length is 3300 bp and on average there are 7.8 introns per gene (Lander et al., 2001). Clearly, in more complex eukaryotes the opportunities provided by enhanced regulatory capacity outweigh the costs of maintaining longer introns, or at least the longer introns do not pose a significant reduction of fitness during the long generation time of these organisms (Rogozin et al., 2012).
The origin of spliceosomal introns has been debated extensively, with the discussion centering around two hypotheses called “introns-early” and “introns-late” (Roy and Gilbert, 2006). The introns-early theory postulates that introns were already present in the early genomes, and that the present-day genes were originally evolved from short coding fragments surrounded by noncoding sequences. According to this theory, present-day streamlined genomes (found in Bacteria, Archaea and some unicellular eukaryotes) that show only a few or no introns, have emerged through a process of intron loss during evolution (Darnell, 1978; Doolittle, 1978). In contrast, the introns-late theory postulates that prokaryotic genomes represent ancestral genome organization, and that present-day intron-containing eukaryotic genomes emerged through insertion or invasion of introns to the ancestral genomes (Logsdon, 1998). A variant of introns-early theory, called “introns-first”, proposes that the origin of introns lies in the RNA world (Gilbert and de Souza, 1999). It postulates that introns have evolved from ribozymes, and protein-coding genes from exonic fragments that accumulated to the RNA genome over time (Jeffares et al., 1998).

Comparison of a large number of eukaryotic genomes has revealed that intron-poor genomes have undergone extensive intron loss, which supports the introns-early theory (Csuros et al., 2011). Only one sequenced eukaryotic genome, that of Hemiselmis andersenii, has been found to be completely intronless (Lane et al., 2007). However, H. andersenii is an endosymbiont with a heavily reduced genome, and even as such rather exceptional, as other reduced genomes of parasitic and deep-branching eukaryotes do contain introns (Nixon et al., 2002; Simpson et al., 2002; Vanácová et al., 2005). Intron gain has occurred in some lineages, especially the Metazoans, but only on a limited number of occasions (Csuros et al., 2011). Thus, it seems evident now that the genome of the last eukaryotic common ancestor was intron-rich, and genome streamlining has led to intron loss in many lineages.

1.1.3 Self-splicing introns

Introns in eukaryotic genes are thought to have evolved from self-splicing RNA elements. The latter are capable of autocatalytic excision from host RNA practically without the help of proteins, thanks to a complex secondary structure. They may also contain an intron-encoded reverse transcriptase and homing endonuclease that enables retrotransposition into new genomic sites. Intron-encoded RNA maturase is also required for splicing in vivo but not in vitro. Self-splicing introns can be found in bacteria and organelles of various eukaryotes, such as fungi and plants, and are classified according to their catalytic mechanism as group I and group II introns. (Bonen and Vogel, 2001; Fedorova and Zingler, 2007). Both self-splicing intron types employ two sequential transesterification reactions initiated by a nucleotide, but differ in the source of the nucleotide (external or internal) and the secondary structure the intron
folds into. Group I self-splicing introns utilize an external GTP molecule to initiate the first nucleophilic reaction. Group II self-splicing introns resemble spliceosomal introns in their splicing mechanism, initiated by a bulged adenosine residue within the intron itself, and resulting in spliced exons and an excised lariat intron (see section 1.3.1). Group II self-splicing introns are thought to be the evolutionary ancestors of spliceosomal small nuclear RNAs (snRNAs) and introns due to similarities in secondary structure and the catalytic mechanisms of intron removal. Their mobility due to retrotransposition also presents an attractive hypothesis on the mechanism of intron gain in very early organisms. However, no conclusive evidence exists to resolve the origin of spliceosomal snRNAs and introns, i.e. whether they evolved from group II introns or whether the two groups share a common ancestor (Penny et al., 2009). Alternatively, they may also have risen not through common ancestry but via convergent chemical evolution (Weiner, 1993).

1.1.4 Splice sites
Spliceosomal introns and group II introns contain three short conserved sequence elements. Intron ends are known as the 5’ splice site (5’ss) and 3’ splice site (3’ss) sequences, while the branch point sequence (BPS) lies slightly upstream of the 3’ss (Figure 1). Most introns are processed by the canonical U2-dependent spliceosome, which recognizes a 5’ splice site of the consensus sequence AG/GTAAG (where / denotes exon-intron boundary). The consensus sequence for the 3’ss is shorter, consisting of only three nucleotides, YAG, at the end of the intron, preceded by the polypyrimidine tract (PPT). In close proximity to the 3’ss, typically 20-40 nt upstream, is the branch point sequence (Gao et al., 2008; Corvelo et al., 2010), containing the branch point adenosine, which is used in the first step of catalysis. BP sequences are highly degenerate in multicellular organisms, only in yeasts and other fungi do they follow a strict consensus sequence (Tolstrup et al., 1997; Kupfer et al., 2004; Gao et al., 2008). BP can be accurately predicted, however, using an algorithm that predicts U2 snRNA binding affinity combined with PPT strength and 3’ terminal dinucleotide (Corvelo et al., 2010). U2-type intron splice sites are overall highly degenerate, only the terminal dinucleotides GT-AG of the intron are 99 % invariant (Sheth et al., 2006). Typically, in intron-poor genomes the adherence to consensus splice sites is more stringent, whereas in genomes with high intron density, more deviation is tolerated (Irimia et al., 2007).
A parallel type of introns exists in diverse eukaryotic lineages. It was originally discovered due to unusual splice sites that diverged significantly from the canonical GT-AG rule found in the majority of introns. Therefore, this group of introns was originally known as AT-AC introns (Jackson, 1991). Shortly afterwards the group was renamed as U12-type introns according to an snRNA component of the novel spliceosome (called U12-dependent or minor spliceosome) involved in the removal of these introns (see section 1.2.2.; Sharp and Burge, 1997). Renaming was further prompted by the discoveries that the normal, or U2-type, introns can also have AT-AC termini and that an AT-AC intron with terminal nucleotides mutated to GT-AG was still spliced by the minor spliceosome (Dietrich et al., 1997; Wu and Krainer, 1997). The 5’ splice site consensus for U12-type introns is /RTATCCTTT (Burge et al., 1998; see Figure 1). The polypyrrimidine tract is missing in these introns; instead, the branch point sequence is highly conserved among U12-type introns (TTCCCTAAC; see Figure 1). Dinucleotides at the 3’ss splice site are typically AG/ for introns with GT 5’ss and AC/ for introns with AT 5’ss, although a few introns with atypical terminal dinucleotides have been experimentally shown to be spliced by the U12-dependent spliceosome (Dietrich et al., 1997; Levine and Durbin, 2001; Hastings et al., 2005; Lin et al., 2010).

1.1.5 Intron types in different organisms

While U2-dependent introns are ubiquitous in virtually all eukaryotic genomes (see above), the U12-dependent introns show more restricted distribution. They can be found in the majority of plant and animal species and some fungi as well as a few deep-branching single-celled eukaryotes. However, they are missing from a number of species, including the common model organisms S. cerevisiae and Caenorhabditis elegans (Burge et al., 1998; Levine and Durbin, 2001; Russell et al., 2006; Sheth et al., 2006; Bartschat and Samuelsson, 2010). Their absence in many organisms whose relatives still retain them suggests that the U12-dependent spliceosome has been lost on several occasions during evolution following extensive intron loss and conversion of U12-type introns to U2-type (Turunen et al., 2013a).
Conversion of U12-type introns to U2-type introns has been suggested to be more common than in the opposite direction (Burge et al., 1998). This is because the consensus sequence for U2-type splice sites is more relaxed, causing a higher likelihood of a small number of mutations leading to the loss of recognition by the U12-dependent spliceosome and switch to the U2-dependent spliceosome. Also, the cooperativity in 5'ss and BPS recognition (see section 1.3.3; Frilander and Steitz, 1999) requires that both of these sequences should be compatible with the U12-dependent spliceosome. Thus a single point mutation in either site can lead to conversion of the intron type. Inspection of orthologous U12-type introns in several genomes has indeed confirmed that U12-type to U2-type conversion is a frequent process; however, loss of U12-type introns (instead of replacement with U2-type introns) is noted to be even more common (Lin et al., 2010). Loss of introns is more common than intron gain for introns in general (Coulombe-Huntington and Majewski, 2007); however, the overall loss of U12-type introns is greater than of U2-type introns given that the former is a sum of both genuine loss and intron conversion events.

If U12-dependent introns are more easily converted to U2-dependent introns, they must have been more common earlier in evolution. As discussed above (section 1.1.2), general purging of introns has occurred in organisms that have short generation times and large populations. In this light it is not surprising that genomic streamlining has led to the complete loss of the parallel minor spliceosome in some lineages (Dávila López et al., 2008; Bartschat and Samuelsson, 2010). Possible advantages of maintaining two splicing systems are discussed in section 1.5.3.

The origin of the U12-dependent spliceosome is as perplexing as its perseverance. The two systems are similar and share components (see section 1.2.2) but are clearly specialized. Their degree of similarity suggests a common ancestor, yet this may be the result of convergent evolution guided by a shared pool of protein factors (Burge et al., 1999). The ancestral group II self-splicing intron that possibly invaded the genome of a eukaryotic ancestor may have duplicated early, and the two splicing systems may have diverged from there. A fission/fusion model suggests that the genetic drift was facilitated by a split into separate lineages, which then fused to became the ancestor of modern eukaryotes (Burge et al., 1998). However, it has been argued that such genomic drift would have made the diverged genomes incompatible to fusion, and therefore another scenario explaining the present day situation is the seeding of a genome with two similar but nonidentical group II self-splicing introns (Lynch and Richardson, 2002).
1.2 SPliceosome Components

1.2.1 Components of the U2-dependent spliceosome

The two parallel types of introns are excised in the nucleus by two parallel spliceosomes. The U2-dependent spliceosome consists of five small nuclear RNAs, U1, U2, U4, U5 and U6, and together with protein components, they form the small nuclear ribonucleoprotein particles (snRNPs). The snRNAs fold into conserved secondary structures (Figure 2). U1, U2, U4, and U5 snRNAs contain a 2,2,7-trimethylguanosine (m\textsubscript{3}G) cap at their 5’ end, and a uridine-rich binding site for Sm proteins (listed in Table 1), which are assembled into a ring structure around the snRNAs during their maturation (Patel and Bellini, 2008; Pomeranz Krummel et al., 2009; Leung et al., 2011). U6 has a similar structure composed of Sm-like (Lsm2-8) proteins binding to the U-rich 3’ end and is capped by a γ-monomethyl cap structure (Singh and Reddy, 1989; Zhou et al., 2014). The Sm-class snRNAs are transported to the cytoplasm after transcription to be loaded with the Sm ring and hypermethylated at the cap, which triggers nuclear import (Fischer et al., 1993). The final steps of the Sm-class snRNPs takes place in the Cajal bodies (see section 1.5.1; Sleeman and Lamond, 1999). In contrast, the biogenesis of U6 is entirely nuclear (Will and Lührmann, 2001).

In addition to common proteins, each snRNP contains unique proteins (reviewed by Will and Lührmann, 2011). U1-specific proteins U1A and U1-70K bind directly to U1 snRNA, while U1C associates via protein-protein interactions (Scherly et al., 1989; Surowy et al., 1989; Nelissen et al., 1991; Pomeranz Krummel et al., 2009). U2 snRNP contains the specific proteins U2A’ and U2B” (Scherly et al., 1990; Price et al., 1998) in addition to the multiprotein complexes named SF3a and SF3b (Brosi et al., 1993; Krämer et al., 1999). U4 and U6 snRNAs associate with each other via extensive base pairing, and together with specific proteins (listed in Table 1) form the U4/U6 di-snRNP (Teigelkamp et al., 1998; Nottrott et al., 1999; Makarova et al., 2002; Nottrott et al., 2002). U5 is the largest of the spliceosomal snRNPs, and contains nine specific proteins listed in Table 1 (Behrens and Lührmann, 1991; Mougin et al., 2002). Many U5-specific proteins perform critical functions in dynamic rearrangements of the spliceosome and splicing catalysis. The most notable are the large multifunctional Prp8/U5-220K which makes contacts with the 5’ exon, BPS and 3’ss, providing a scaffold for the catalytic core; RNA helicases Brr2/U5-200K and Prp28/U5-100K; and the GTPase Snu114/U5-116K (Grainger and Beggs, 2005; Small et al., 2006; Bessonov et al., 2008; Häcker et al., 2008; Valadkhan and Jaladat, 2010). U5 associates with U4/U6 di-snRNP to form a ternary complex termed the U4/U6.U5 tri-snRNP (Black and Pinto, 1989). There is no evidence of RNA-RNA base-pairing interactions bridging U4/U6 to U5; instead, the tri-snRNP is held together by protein-protein interactions (Liu et al., 2007). The assembled U4/U6.U5 tri-snRNP contains several proteins not present in the individual snRNP.
complexes, which are proposed to account for the bridging interactions (Bordonné et al., 1990; Makarov et al., 2002; Mougin et al., 2002; Will and Lührmann, 2011).

A large number of non-snRNP proteins associate with the spliceosome, and in fact, the total count of spliceosomal proteins has been placed as high as 150-300 proteins, as identified in purification and mass spectrometry analyses of different spliceosomal complexes (reviewed by Valadkhan and Jaladat, 2010). Comparison of spliceosomes from different eukaryotes revealed that a core set of proteins is conserved between yeast, Drosophila melanogaster and human spliceosomes, but metazoan complexes contain considerably more proteins (Fabrizio et al., 2009). The dynamic changes that the spliceosome undergoes during the splicing steps (see 1.3.2) are reflected in the protein composition as well; proteins enter and leave at specific steps (Fabrizio et al., 2009), when their enzymatic activities as peptidyl-prolyl cis/trans-isomerases, helicases, kinases and other protein and RNA-modifying enzymes are employed. Consequently, there are some discrepancies in reported numbers of bona fide spliceosomal components. These arise from the dynamic nature of the spliceosome combined with differences in the experimental settings and cutoff criteria.
Figure 2. Secondary structures of human spliceosomal snRNAs.
The sequences interacting with 5'ss or BPS are marked with a black line, the binding sites for Sm proteins are boxed, and the sequences involved with U2/U6 or U12/U6atac interactions are highlighted in grey. Nucleotide modifications are omitted. Adapted with modifications from Turunen (2012). Structures are as published by Yu et al. (1999) for U1, U2, and U5, Nottrott et al. (2002) for U4, U6, U4atac, and U6atac, Tarn and Steitz (1997) for U11, and Sikand and Shukla (2011) for U12.
1.2.2 Components of the U12-dependent spliceosome

Analogous to the major spliceosome, the U12-dependent spliceosome is composed of five snRNPs. U11, U12, U4atac and U6atac replace the U1, U2, U4 and U6, respectively, whereas U5 is shared between both spliceosomes (Hall and Padgett, 1996; Tarn and Steitz, 1996a, b). At sequence level the U12-type snRNAs are not similar to U2-type snRNAs; U11 and U12 share no sequence similarity with their counterparts U1 and U2 (Montzka and Steitz, 1988), and U4atac and U6atac show only partial homology to U4 and U6, with 40% similarity (Tarn and Steitz, 1996a). However, all four snRNAs fold to highly similar secondary structures when compared with major snRNAs, and U4atac base pairs with U6atac in a manner similar to U4 and U6 (see Figure 2; Padgett and Shukla, 2002). U5 associates with U4atac and U6atac to form the minor tri-snRNP U4atac/U6atac.U5 (Tarn and Steitz, 1996a). In contrast to the U2-type snRNPs, U11 and U12 form a stable di-snRNP where the individual snRNPs are connected via protein-protein interactions, but the cell also contains a pool of a free U11 snRNP (Wassarman and Steitz, 1992b; Benecke et al., 2005; Turunen et al., 2008).

U12-type snRNAs associate with many of the same proteins as major snRNAs. U11, U12 and U4atac are complexed with the Sm heptameric protein ring, and U6atac with the Lsm counterpart (Montzka and Steitz, 1988; Tarn and Steitz, 1996a; Will et al., 1999; Schneider et al., 2002). Indeed, most (if not all) of the minor U4atac/U6atac.U5 tri-snRNP proteins are shared with the major tri-snRNP, and U4 snRNA can even substitute for U4atac if it is mutated so that it base-pairs with U6atac (Luo et al., 1999; Schneider et al., 2002; Shukla and Padgett, 2004). In contrast, the U11/U12 di-snRNP contains a number of proteins unique to the U12-dependent spliceosome. SF3a protein complex is not present in the U11/U12 di-snRNP, while SF3b complex is (Will et al., 2004). Other missing proteins are the U1-specific proteins U1A, U1-70K and U1C. Instead, the U11/U12 di-snRNP contains seven proteins designated 65K, 59K, 48K, 35K, 31K, 25K, and 20K (Will et al., 2004). Of these, 48K has a critical role in the sequence specific recognition of the 5’ss, and also contributes to the stability of the di-snRNP via interactions with the 59K protein (Turunen et al., 2008). Due to a similar domain structure, 35K has been suggested to act as a functional analog of U1-70K, and likewise, structures of the 20K and 65K proteins are reminiscent of the U1C and U1A/U2B” proteins, respectively (Will et al., 2004). 65K binds U12 snRNA and interacts with the 59K protein of U11 snRNP, bringing the di-snRNP together (Benecke et al., 2005). Consistently, mutations of the human 65K gene cause the dissociation of the di-snRNP, aberrant U12-dependent splicing, and developmental defects in patients (Argente et al., 2014).

U12-type snRNPs are less abundant than their major snRNP counterparts, numbering in mammalian cells approximately 2 x 10<sup>3</sup> to 10<sup>4</sup> particles per cell, or ~100-fold less
than major snRNPs (Montzka and Steitz, 1988; Tarn and Steitz, 1996a). This is not surprising as the introns excised by the U12-dependent spliceosome are also much less common; however, the reports about the effect of minor snRNP abundance on U12-dependent splicing rate are conflicting (Pessa et al., 2006; Younis et al., 2013, see also section 1.5.3). Even so, defects in the snRNPs are not tolerated in cells, as mutations of the human minor snRNP components cause severe developmental defects that result from disruption of RNA-protein interactions (Edery et al., 2011; He et al., 2011; Argente et al., 2014; Jafarifar et al., 2014). Defective minor snRNP assembly and splicing defects resulting therefrom may also play a role in the degeneration of motor neurons in spinal muscular atrophy (SMA) patients (Boulisfane et al., 2011). Minor spliceosome components are also essential for development in Arabidopsis thaliana, D. melanogaster, and Danio rerio (Otake et al., 2002; Kim et al., 2010; Pessa et al., 2010; Markmiller et al., 2014).
Table 1. Human integral spliceosomal snRNP proteins.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sm1</td>
<td>Sm1</td>
<td>Sm1</td>
<td>Sm1 Sm1</td>
<td>Sm1</td>
<td>Sm1</td>
<td>snRNP core components</td>
<td>Pomeranz Krummel et al. (2009)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lsm2 Lsm2</td>
<td></td>
<td>Lsm2</td>
<td>snRNP core components</td>
<td>Zhou et al. (2014)</td>
</tr>
<tr>
<td>U1-70K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cho et al. (2011)</td>
</tr>
<tr>
<td>U1A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U1C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U2A'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U2B&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF3a³</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF3b³</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRP43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>220K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>116K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>102K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Sm: snRNP core components
- Lsm: snRNP core components
- Snp1: Structural; SR protein interactions
- Mud1: Structural; RNA-binding
- Yhc1: 5'ss recognition
- Lea1: Structural; RNA-binding
- Msl1: Structural; RNA-binding
- SF3a³: BPS binding
- SF3b³: BPS binding
- PRP43: Disassembly factor, proofreading
- Prp8: Stabilization of catalytic core, contacts many spliceosomal and pre-mRNA sites
- Brr2: RNA helicase, unwinding of U4/U6
- Snu14: GTPase
- Prp6: Tri-snRNP bridging
- Prp28: RNA helicase, exchange of U1 with U6 at 5'ss
- Snu40: Tri-snRNP assembly
- Structural

Selected references:

- Pomeranz Krummel et al. (2009)
- Zhou et al. (2014)
- Cho et al. (2011)
- Price et al. (1998)
- Gozani et al. (1996)
- Gozani et al. (1996); Will et al. (1999)
- Will et al. (2004); Mayas et al. (2010)
- Pena et al. (2008); Galej et al. (2013)
- Santos et al. (2012)
- Small et al. (2006)
- Schneider et al. (2010a)
- Staley and Guthrie (1999)
- Laggerbauer et al. (1998)
- Achsel et al. (1998)
<table>
<thead>
<tr>
<th>15K</th>
<th>15K</th>
<th>Dib1</th>
<th>Structural</th>
<th>Laggerbauer et al. (1998)</th>
</tr>
</thead>
<tbody>
<tr>
<td>90K</td>
<td>90K</td>
<td>Prp3</td>
<td>Tri-snRNP bridging</td>
<td>Horowitz et al. (1997)</td>
</tr>
<tr>
<td>61K</td>
<td>61K</td>
<td>Prp31</td>
<td>Tri-snRNP bridging</td>
<td>Schneider et al. (2010a)</td>
</tr>
<tr>
<td>60K</td>
<td>60K</td>
<td>Prp4</td>
<td>Kinase, tri-snRNP entry into spliceosome</td>
<td>Schneider et al. (2010a)</td>
</tr>
<tr>
<td>20K</td>
<td>20K</td>
<td>CypH</td>
<td>Peptidyl-prolyl cis/trans isomerase</td>
<td>Teigelkamp et al. (1998)</td>
</tr>
<tr>
<td>15.5K</td>
<td>15.5K</td>
<td>Snu13</td>
<td>Tri-snRNP bridging</td>
<td>Nottrott et al. (2002)</td>
</tr>
<tr>
<td></td>
<td>110K</td>
<td>Snu66</td>
<td>Recruitment of tri-snRNP to the spliceosome</td>
<td>Makarova et al. (2001); Schneider et al. (2002)</td>
</tr>
<tr>
<td>65K</td>
<td></td>
<td>Sad1</td>
<td>Recruitment of tri-snRNP to the spliceosome</td>
<td>Makarova et al. (2001); Schneider et al. (2002)</td>
</tr>
<tr>
<td>27K</td>
<td></td>
<td>-</td>
<td>Recruitment of tri-snRNP to the spliceosome</td>
<td>Fetzer et al. (1997)</td>
</tr>
<tr>
<td>65K</td>
<td></td>
<td>-</td>
<td>Structural, binds U12 snRNA</td>
<td>Benecke et al. (2005)</td>
</tr>
<tr>
<td>59K</td>
<td></td>
<td>-</td>
<td>Structural, binds 48K and 65K</td>
<td>Benecke et al. (2005); Turunen et al. (2008)</td>
</tr>
<tr>
<td>48K</td>
<td></td>
<td>-</td>
<td>5’ss recognition</td>
<td>Turunen et al. (2008)</td>
</tr>
<tr>
<td>35K</td>
<td></td>
<td>-</td>
<td>SR protein interactions, homology to U1-70K</td>
<td>Lorković et al. (2004)</td>
</tr>
<tr>
<td>31K</td>
<td></td>
<td>-</td>
<td>Unknown; RNA-binding</td>
<td>Kim et al. (2010)</td>
</tr>
<tr>
<td>20K</td>
<td></td>
<td>-</td>
<td>Unknown; homology to U1C</td>
<td>Will et al. (2004)</td>
</tr>
<tr>
<td>Urp</td>
<td></td>
<td>-</td>
<td>3’ss recognition</td>
<td>Shen et al. (2010)</td>
</tr>
</tbody>
</table>

After Will et al. (2004); Will and Lührmann (2011); and Turunen et al. (2013a).

1Sm proteins B/B’, D1, D2, D3, E, F and G.
2Lsm2-8.
3Multisubunit complexes.
1.3 SPLICEOSOME ASSEMBLY AND CATALYSIS

1.3.1 Splicing catalysis
Pre-mRNA splicing is achieved by two consecutive transesterification reactions (reviewed by Valadkhan and Jaladat, 2010), the first step being the nucleophilic attack of the 2' hydroxyl group of the branch point nucleotide on the phosphate group at the 5' splice site. As a result, the upstream exon is left with a free 3' hydroxyl group, and the intron together with 3' exon forms a looped structure known as the lariat intermediate. The liberated 3' hydroxyl group at 5'ss then performs the second nucleophilic attack on the 3'ss phosphate group, resulting in the ligation of the two exons and the release of the excised intron as a lariat. As such, the reaction is energetically neutral, but ATP is required in many proofreading and rearrangement steps during spliceosome assembly. A similar two-step reaction mechanism is employed by both spliceosomes and also by group II self-splicing introns (Lambowitz and Zimmerly, 2011).

As self-splicing introns are capable of splicing catalysis without the help of proteins, it has been a long-standing belief that the spliceosomal catalysis, too, is performed by RNA. U6 snRNA, the most conserved snRNA component of the spliceosome (Guthrie and Patterson, 1988), resides at the catalytic center of the spliceosome and coordinates the catalytic Mg$^{2+}$ ions (Yean et al., 2000; Fica et al., 2013). Even more concrete evidence for snRNA catalysis comes from the observations that U2 and U6 snRNAs can catalyze a two-step splicing-like reaction in the absence of proteins (Valadkhan and Manley, 2001). Furthermore, with the minor spliceosome it has been shown that replacing the U6atac stem-loop structure with a group II intron domain D5 stem-loop, which is also involved in coordination of divalent cations in the catalysis of self-splicing introns, supports splicing of U12-type introns (Shukla and Padgett, 2002).

In addition to snRNAs, the Prp8 protein of the U5 snRNP has been, for a long time, the other candidate for catalysis, as it makes many contacts within the catalytic core, including U6, U5, 5'ss, 3'ss, and the branch point (reviewed by Abelson, 2008). However, in a recent crystal structure of Prp8, it was shown that critical residues of the protein, corresponding to active site residues of a related protein fold, do not coordinate magnesium and are thus unable to perform catalysis (Galej et al., 2013). Still, it is clear that Prp8 has important functions in the positioning of the substrate and snRNPs (Pena et al., 2008), and numerous protein components are necessary to perform splicing efficiently and precisely in vivo.
1.3.2 Assembly of the U2-dependent spliceosome

Spliceosome assembly has been studied extensively in vitro. It is assembled by the stepwise addition of the spliceosomal snRNPs onto the splicing substrate (see Figure 3). Using biochemical methods, assembly intermediates have been observed and are usually labeled E, A, B, B\textsubscript{act}, B\textsuperscript{*}, and C complexes, in the order of appearance (reviewed by Will and Lührmann, 2011).

The assembly pathway begins with the association of the U1 snRNP with the 5' splice site and the U2AF protein dimer with the 3' end of the intron (Zhuang and Weiner, 1986; Zamore and Green, 1991). U1 snRNA base-pairs with the 5'ss, and the interaction is stabilized by protein-RNA interactions, especially by the U1C protein (Heinrichs \textit{et al.}, 1990; Du and Rosbash, 2002; Pomeranz Krummel \textit{et al.}, 2009). At the same time, protein factors SF1 (splicing factor 1) and U2AF recognize the branch point sequence and the 3' end of the intron, respectively (Berglund \textit{et al.}, 1997; Liu \textit{et al.}, 2001). The branch point A residue used as the nucleophile in the first catalytic step is already recognized and bulged out by the SF1 (Liu \textit{et al.}, 2001). The 65K subunit of U2AF is responsible for binding to the polypyrimidine tract while the 35K subunit binds to the 3'ss (Zamore and Green, 1991; Guth \textit{et al.}, 1999; Wu \textit{et al.}, 1999; Zorio and Blumenthal, 1999). These interactions constitute the early (E), or commitment complex.

Prespliceosome, or A complex, emerges when U2 snRNP displaces SF1 at the branch point (Bindereif and Green, 1987). U2 is already present in the E complex in close proximity to U1, but remains loosely associated (Das \textit{et al.}, 2000; Dönmez \textit{et al.}, 2007). ATP is needed for the stable binding of U2 snRNP but this requirement is relaxed when the 5' end of U1 is mutated, or when a checkpoint protein Cus2 is deleted, suggesting that some rearrangement is necessary to accommodate U2 stably (Liao \textit{et al.}, 1992; Perriman and Ares, 2000). U2AF65 binds to the polypyrimidine tract and with the help of its RS domain recruits U2 snRNP via the SF3b complex (Valcárcel \textit{et al.}, 1996; Gozani \textit{et al.}, 1998; Selenko \textit{et al.}, 2003; Spadaccini \textit{et al.}, 2006). U2 snRNA base pairs with the branch point sequence (Wu and Manley, 1989; Zhuang and Weiner, 1989). The same branch point adenosine that was already recognized by SF1 (see above) is also bulged out from the U2/BPS base-pairing helix (Parker \textit{et al.}, 1987; Query \textit{et al.}, 1994).

In the next step of assembly, U4/U6.U5 tri-snRNP enters the nascent spliceosome, forming the precatalytic B complex. Rearrangements of base-pairing helices follow, transforming the spliceosome to the activated B\textsubscript{act} complex. In these rearrangements, new base-pairing interactions are formed between U6 snRNA and 5'ss, and the interaction between U1 and 5'ss is unwound by the action of Prp28 helicase (U5-100K in humans; Kandels-Lewis and Séraphin, 1993; Staley and Guthrie, 1999). Similarly U4 unwinds from U6, catalyzed by the Brr2 (U5-200K) helicase (Konarska and Sharp, 1987; Wassarman and Steitz, 1992a; Raghunathan and Guthrie, 1998; Mozaffari-Jovin \textit{et al.}, 2000).
As a consequence, U2 snRNA base-pairs with U6 (Hausner et al., 1990; Wu and Manley, 1991; Madhani and Guthrie, 1992), which requires rearrangements of the internal stem loops of both snRNAs (reviewed in Staley and Guthrie, 1998). A "catalytic core" structure is formed, where U2 and U6 snRNA bound together also bind to the BPS and 5’ss, thus bringing them close in space for the first step of the catalysis (Burke et al., 2012). U5 helps to align the exons for accurate cleavage (Newman and Norman, 1991; Sontheimer and Steitz, 1993). Significant remodeling of spliceosomal protein content also occurs during B complex activation, with U1 and U4-associated proteins leaving and a considerable number of proteins joining, including the Prp19/CDC5 complex (abbreviated 19C or NTC; Makarov et al., 2002; Bessonov et al., 2008; Fabrizio et al., 2009; Agafonov et al., 2011). Consequently, the B act complex is converted to the catalytically active spliceosome (B*) by the DExH/D-box helicase Prp2 which remodels the active site by removing SF3a and SF3b complexes and exposing the branch point adenosine (Warkocki et al., 2009; Lardelli et al., 2010; but see also Wlodaver and Staley, 2014). The first step of splicing is then performed. Further conformational changes, assisted by helicases and peptidyl-prolyl cis/trans isomerases, accompany the progress to complex C, which catalyses the second transesterification step.

After the splicing reaction has taken place, the postcatalytic complex is disassembled, releasing ligated exons and a lariat intron, which still retains U2, U6 and U5 snRNPs. These, along with other spliceosomal components, are recycled for the next splicing reaction (Tsai et al., 2005). Recycling factors, such as Prp43, can also be used for the rejection of suboptimal splicing intermediates, improving the fidelity of the reaction (Mayas et al., 2010).

1.3.3 Intron recognition by the U12-dependent spliceosome

The overall assembly pathway of the U12-dependent spliceosome is likely to resemble the major splicing pathway (Turunen et al., 2013a). However, a fundamental difference exists in the recognition and commitment of splice sites. No commitment complex can be detected on minor introns, and protein factors U2AF and SF1 do not participate in their recognition. Instead, the 5’ splice site and branch point sequences are recognized cooperatively by the pre-formed U11/U12 di-snRNP (Tarn and Steitz, 1996b; Frilander and Steitz, 1999). Analogously to the major spliceosome, the 5’ splice site is recognized by U11 snRNA through sequence-specific base-pairing, but the base-pairing spans only nucleotides +4 to +8 compared to U2-type interaction spanning the exon-intron junction (Kolossova and Padgett, 1997; Yu and Steitz, 1997). A protein component unique to the minor spliceosome, U11-48K, recognizes the first three nucleotides of the intron (Turunen et al., 2008; Tidow et al., 2009). U12 snRNA performs the same function as the U2 snRNA in the major spliceosome in recognizing and base-pairing
with the BPS, and bulging out the branch site A residue for the first step of catalysis (Hall and Padgett, 1996; Tarn and Steitz, 1996b). The 3’ splice site is initially bound by the Urp protein, a protein component of the U11/U12 di-snRNP, which is related to U2AF35 and required for both U12 and U2-dependent splicing pathways, but, interestingly, at different stages of spliceosome assembly (Will et al., 2004; Shen et al., 2010).

The higher splice site sequence conservation, the lack of the polypyrimidine tract, and the non-involvement of U2AF imply that initial intron recognition by the U12-dependent spliceosome relies more on sequence-specific snRNA-mRNA interactions than the major type spliceosome (Brock et al., 2008). Still, exon definition interactions, assisted by protein splicing factors, are important also for the U12-dependent splicing (Hastings and Krainer, 2001).
**Figure 3.** Assembly pathways of the U2-dependent and U12-dependent spliceosomes. Adapted with modifications from Turunen et al. (2013a). See text for details. Question mark reflects that the association of the Nineteen complex (19C) to the U12-dependent spliceosome has not been directly shown, but is inferred from the major spliceosome.
1.3.4 Assembly of the U12-dependent spliceosome
Due to the differences in early assembly steps, described in the previous section (1.3.3), the first splicing complex that can be detected on U12-dependent introns is the A complex, in which the U11/U12 di-snRNP has bound the intron’s 5’ss and BPS in a cooperative manner (Tarn and Steitz, 1996b; Frilander and Steitz, 1999). During prespliceosome formation, these catalytic sites are therefore brought to proximity, implying that less rearrangements in spliceosome conformation are needed later on (Frilander and Meng, 2005). A similar arrangement of catalytic sites in the early spliceosome has also been observed with the major spliceosome (Dönmez et al., 2007), suggesting that the prespliceosomal architectural constraints may also be conserved between the two spliceosomes. Consistently, overall the assembly pathways of the two spliceosomes are thought to resemble each other remarkably well. As noted above (section 1.2.2), the snRNA components of the U12-dependent spliceosome, although divergent in sequence, fold into similar secondary structures, and associate with an almost identical protein complement. Also, bulging of the reactive adenosine is achieved analogously during base-pairing between U12 and BPS (Hall and Padgett, 1996; Tarn and Steitz, 1996b). The U4atac/U6atac/U5 tri-snRNP behaves as their major tri-snRNP counterpart, its entry marking the formation of complex B, where U6atac unwinds from U4atac and replaces U11 at the 5’ splice site (Tarn and Steitz, 1996a; Yu and Steitz, 1997; Incorvaia and Padgett, 1998). The base-pairing that occurs between U12 and U6atac closely resembles that between U2 and U6, and they are thought to constitute the catalytic core in a manner similar to the major spliceosome and group II self-splicing introns (Tarn and Steitz, 1996a; Frilander and Steitz, 2001; Shukla and Padgett, 2001; Shukla and Padgett, 2002). It has been noted, however, that in the U12-dependent splicing pathway, the catalytic core formed by U12/U6atac base-pairing may appear before U11 has been released from the 5’ splice site, implying that proofreading of the 5’ splice site is not strictly required for the formation of the catalytic core (Frilander and Steitz, 2001). Finally, the two transesterification reactions are identical between the two spliceosomes and yield the same final products, the ligated exons and a lariat intron, and are likely followed by a similar disassembly pathway and recycling of snRNPs for the next round of splicing as in the major spliceosome (Damianov et al., 2004). It has also been noted that the U11/U12 di-snRNP contains the disassembly factor PRP43, which in the major spliceosome associates with U2 snRNP (Will et al., 2004).

1.4 SPlice Site SELECTION

1.4.1 Exon definition
As stated earlier, the structure of eukaryotic genes, particularly in vertebrates, is such that short exons are found within long stretches of intronic sequences. Combined with
the degeneracy of typical major type splice sites, it poses a problem for the splicing machinery to accurately identify exons and disregard those sequences that only resemble *bona fide* splice sites by chance.

The exon definition hypothesis resolves the difficulty of pairing exons correctly over long introns by postulating that the splice sites across exons are paired first, and then a switch occurs to pair those sites that will be spliced together. The hypothesis was formulated when it was observed that a downstream 5′ splice site enhanced the splicing of an upstream intron (Robberson *et al.*, 1990). Since then substantial evidence has accumulated in support of the hypothesis. For instance, strengthening the 5′ss (leading to increased U1 binding) will enhance the binding of U2AF65 to an upstream 3′ss (Hoffman and Grabowski, 1992). Also, terminal exons that do not have upstream (first exon) or downstream splice sites (last exon) are enhanced by the communication to mRNA terminal structures, the cap binding complex in the 5′ end and the polyadenylation site in the 3′end of the transcript (Niwa and Berget, 1991; Lewis *et al.*, 1996). The alternative hypothesis, intron definition, has received support for those transcripts that have relatively large exons and short introns, as is typically the case in yeast (Talerico and Berget, 1994; Fox-Walsh *et al.*, 2005). However, in vertebrates it appears that the length of a typical exon is shorter than 300 nt, and one reason may be the requirement to support exon definition interactions (Berget, 1995; Lander *et al.*, 2001). Further support to exon-bridging interactions playing an important role comes from observations that steric hindrance can restrict the efficiency of snRNP binding across very short exons (Black, 1991; Dominski and Kole, 1991).

In human hereditary diseases that involve mutations in the vicinity of splice junctions, mutations often lead to exon skipping, but only very rarely to intron retention, arguing for initial pairing of exons rather than introns (Krawczak *et al.*, 1992; Nakai and Sakamoto, 1994). Compensatory evolution of mammalian splice sites bordering an exon has also been detected, meaning that if one splice site next to an exon is mutated, compensatory mutations at the other side of the exon may alleviate the splicing defect, but no compensation occurs across introns (Xiao *et al.*, 2007). This further implies that exons are the mammalian evolutionary units that selection acts upon.

### 1.4.2 Commitment complex formation

Splice sites become committed to splicing at the E complex stage of spliceosome assembly. Based on the early work on spliceosome assembly that has utilized minigene constructs containing only one intron surrounded by two exons, it has been assumed that intron definition occurs at this stage. However, in a more natural context the spliceosome is often presented with a task of pairing many consecutive exons and a choice of alternative splice sites (for a discussion of alternative splicing, see section below). The choice of pairing specific splice sites has been convincingly shown to occur
only at A complex stage (Lim and Hertel, 2004). This leaves the opportunity for interactions to occur first across exons, committing splice sites for usage, and then switching the interaction to pair intron ends that are to be spliced together, committing the pairing of exons. Even more persuasively, an A-like complex containing the tri-snRNP assembles across an exon substrate and is converted to a B-like complex capable of splicing when offered a 5’ splice site oligonucleotide in trans, implying that interactions connecting intron ends are not required for B complex formation, and that disruption of cross-exon interactions is not required for the formation of cross-intron interactions (Schneider et al., 2010b). Furthermore, splicing factors that affect alternative splicing may do so at the level of inhibiting the conversion of exon-defined complexes into intron-bridging complexes (Sharma et al., 2008).

### 1.4.3 Alternative splicing

Recent transcriptome-wide surveys have revealed that the majority, a staggering 95%, of mammalian transcripts display alternative splicing, in which the processing of a pre-mRNA transcript results in the inclusion of different parts of that transcript in the final mRNA product (Pan et al., 2008; Wang et al., 2008). Such extensive alternative processing increases variation in the expressed proteome and provides ample opportunity for cell-type, tissue and time point dependent regulation. When comparing different organisms, alternative splicing is found to be most common in mammals and particularly in primates (Barbosa-Morais et al., 2012).

![Figure 4. Types of alternative splicing.](image)

White boxes, constitutive exons; grey boxes, alternative exons. Straight lines, introns; bent lines, splicing patterns. A, Constitutive splicing; B, Alternative 5’ss; C, Alternative 3’ss; D, Alternative cassette exon (i.e. exon skipping or inclusion); E, Intron retention; F, Mutually exclusive exons.
Different kinds of alternative splicing patterns exist, including alternative 5’ and 3’ splice site usage, alternative exon skipping or inclusion (the exon in question is often referred to as a cassette exon), intron retention, and mutually exclusive alternative exons (see Figure 4). Of these, a large fraction of investigations have focused on cassette exon splicing. Remarkable diversity can be achieved with such a restricted number of categories. For instance, the Drosophila Dscam (Down syndrome cell adhesion molecule) gene contains three clusters of mutually exclusive exons, the combination of which results in over 38 000 different mRNA products and changes the protein-protein binding specificity of Dscam (Schmucker et al., 2000). The Dscam adhesion molecules show homophilic binding, meaning that the products of one isoform bind to each other but not to any other isoform type (Wojtowicz et al., 2004). It appears that isoform choice is probabilistic (i.e. not regulated in cell-dependent or temporal manner), and that one neuron expresses on average 4 different types (Miura et al., 2013). In effect, the Dscam adhesion proteins provide an identity code for individual neurons that causes the dendrites of one neuron to repel each other and make contacts only with the dendrites of another neuron (Hattori et al., 2009).

Alternative splicing is rare in U12-type introns (Turunen et al., 2013a). Splice site selection in the minor splicing pathway is less flexible due to longer consensus sequences and more strict distance requirements between splicing signals (Levine and Durbin, 2001). Exon skipping is also absent for U12-dependent introns because U12-type introns usually occur only once in a given gene, and the splice sites are incompatible with U2-type sites. Alternative 3’ splice site usage has been reported, but may represent only splicing noise as the reported alternative 3’ss sequences are typically in very close proximity of the bona fide 3’ss (Levine and Durbin, 2001; Zhu and Brendel, 2003; Hastings et al., 2005).

The incompatibility of U12-type and U2-type splice sites is demonstrated by the alternative splicing pattern found in the vertebrate JNK gene family, in which two alternative mutually exclusive exons utilize the two different types of spliceosomes (Chang et al., 2007). The splicing choice is made in a hybrid intron, which contains a U12-type 5’ splice site and a U2-type 3’ splice site. These can only splice to the compatible splice sites downstream or upstream, respectively. Another case of alternative splicing involving the competition of the two spliceosomes can be found in the Drosophila gene prospero, which codes for a protein involved in axon guidance (Scamborova et al., 2004). This gene contains a U2 intron nested within a U12 intron, a so-called ‘twintron’. If the U12-type sites are used for splicing instead of U2-type, the resulting mRNA codes for a protein that lacks 29 amino acids in the homeodomain region. A similar twintron can also be found in the Drosophila splicing factor Urp (Lin et al., 2010).
1.4.4 Splicing regulatory elements

Splice site signals themselves are too degenerate to contain sufficient information for unequivocal specification of the intron/exon boundary (Burge et al., 1999). This is true even for the U12-type splice site sequences which are more conserved than the U2-type sites (Burge et al., 1999). Therefore, sequence elements termed splicing regulatory elements (SREs) are needed to provide additional information for the identification of correct splice sites from the abundant pseudo sites (Zhang et al., 2003). In addition to helping in specifying constitutively spliced exons, the same elements also regulate alternative splicing. These sequence elements are present in both exons and introns of the transcript, and can function either as inhibitory or enhancing signals. A sequence element can even have opposite effects on splicing, depending on its location and other context-dependent factors (Yeo et al., 2007). The effect of an individual SRE is highly context-dependent and often several SREs exert a concerted effect on a particular splicing event. Theoretically, the combinatorial evaluation of sequence elements, collectively named the “splicing code”, on a given transcript should predict the outcome of splice site choices for that transcript (Wang and Burge, 2008). Until recently, however, the complexity and context dependence of the splicing code have made these predictions rather imprecise. Bioinformatic approaches combined with tissue- and developmental stage specific splicing pattern data have resulted in advances in compiling a splicing code that can be used to predict splicing outcome for cassette exons with improved probability (Barash et al., 2010; Barbosa-Morais et al., 2012). However, for splicing choices other than alternative cassette exons, such combinatorial understanding is still missing.

1.4.5 Splicing factors

The SREs described in the previous section mostly function by recruiting trans-acting factors that aid in the selection of splice sites or assembly of the spliceosome. The splicing factors perform this task by binding to their SREs via RNA-binding domains, such as RNA recognition motifs (RRMs), hnRNPK homology domains (KH) and zinc-finger domains (Castello et al., 2012). Subsequently they recruit spliceosome components to the nascent transcripts via association to the Pol II (Das et al., 2007), acting positively by stabilizing interactions between spliceosome components and the pre-mRNA or other splicing factors (Fu and Maniatis, 1992; Boukis et al., 2004), or negatively by blocking such interactions (Kanopka et al., 1996; McNally and McNally, 1996). Splicing factors are important in both constitutive and alternative splicing (Wu and Maniatis, 1993; Zhang et al., 2003). The best known splicing regulatory factors are the proteins belonging to SR (serine-arginine rich) and hnRNP (heterogeneous nuclear RNP) protein families. Originally identified as positive and negative regulators of splicing, respectively, it is now known that their effect on a particular splicing event is heavily context-dependent (Witten and Ule, 2011).
SR splicing factors typically contain one or two RRM domains and an arginine-serine (RS) rich domain (Manley and Krainer, 2010). The SR family comprises twelve proteins including some of the originally described and well-studied members, SRSF1 and SRSF2 (previously known as SF2/ASF and SC35, respectively). There are also SR proteins without RNA recognition domains that can link other SR proteins bound on the RNA (Boucher et al., 2001). Other SR-like proteins include a number of more integral spliceosomal proteins, for instance U1-70K, which enhances the binding of U1 snRNP to the 5’ splice site (Kohtz et al., 1994; Cho et al., 2011), U2AF proteins and Urp, which interact with the 3’ splice site (Shen et al., 2010), and SR proteins that are involved in recruiting the U4/U6.U5 tri-snRNP to the spliceosome (Roscigno and Garcia-Blanco, 1995; Makarova et al., 2001). Interestingly, SR proteins have functions that range beyond splicing regulation and are now recognized to act also in chromatin remodeling, mRNA export, miRNA processing, translation, and mRNA decay (reviewed by Twyffels et al., 2011; Änkö, 2014).

The RRM domains found in SR proteins function by binding to single-stranded RNA while the RS domain typically participates in protein-protein interactions (Mayeda et al., 1992). The length of the RS domain, or more specifically, the number of RS dipeptide repeats, correlates with the strength of activation (Philipps et al., 2003). The serine residues in the RS domain can be phosphorylated and this modulates the activity of the SR protein, and ultimately, provides the means to decode transduction pathway signals to alternative splicing outcomes (Lynch, 2007). For example, phosphorylation of the RS domain can alter the conformation of a SR protein; the unphosphorylated RS domain of SRSF1 is involved in intramolecular interactions to the RRM domain of the same protein, while phosphorylation leads to an open conformation and a switch to intermolecular interactions with the RRM of U1-70K, required for the formation of the E complex (Cho et al., 2011).

HnRNP proteins, like the SR proteins, display a modular structure of RNA-binding domains. They bind to many kinds of RNA and single-stranded DNA sequences, with binding interactions described with mRNAs, miRNAs, and telomerase RNA, but with a particular preference for intronic sequences. They participate in the packaging of nuclear mRNAs and regulate their processing and export, and also shuttle to cytoplasm where they regulate the translation and localization of mRNAs (Han et al., 2010). Their most studied function is to regulate alternative splicing (Huelga et al., 2012). For instance, hnRNP A1, which is one of the best characterized members of the hnRNP group, often antagonizes the effect of SR proteins in splicing regulation. HnRNP C, on the other hand, is involved in the packaging of the majority of nuclear pre-mRNAs and packs the intronic RNA around protein tetramers, much reminiscent of DNA nucleosomes (Huang et al., 1994; König et al., 2010). In doing so it may affect the
choice of splice sites, and indeed, hnRNP proteins also have context-dependent influences on alternative exons, resulting in either enhanced inclusion or silencing (Blanchette et al., 2009).

Splicing factors involved in U2-type splicing are also required in the U12-dependent pathway, where they are involved in exon definition interactions between spliceosomes and also contact the pre-mRNA directly (Wu and Krainer, 1996; Hastings and Krainer, 2001; Shen and Green, 2007). Both SR proteins and hnRNP proteins have been shown to occupy binding sites in U12-type intron-containing transcripts that affect the efficiency of splicing (Lewandowska et al., 2004; McNally et al., 2006; Borah et al., 2009). In addition, the minor spliceosome specific protein U11-35K has a similar domain structure to U1-70K, including an N-terminal RRM domain and a C-terminal RS domain, suggesting that it is involved in stabilization of 5’ splice site interaction and/or bridging interactions (Will et al., 1999; Lorković et al., 2004). Importantly, since there are usually only one, and rarely two or three, U12-type introns in a given transcript in conjunction with several U2-type introns, the U12-dependent spliceosome must be able to form compatible exon definition interactions with the major type spliceosome for efficient processing.

1.4.6 SnRNPs as processing factors
In addition to splicing regulatory proteins, core spliceosomal components can also have a role in splice site selection that is more complex than a mere binding to the splice sites. The levels of several core spliceosomal proteins, including U1, U2, and U4/U6 snRNP-associated proteins, have been shown to affect the outcome of splicing with Drosophila transcripts (Park et al., 2004). In human cells, the depletion of SmB/B’ has been shown to change alternative splicing outcomes (Saltzman et al., 2011). A U1C-deficient zebrafish retains a surprisingly high level of functional splicing but shows misregulation of alternative splicing patterns (Rösel et al., 2011). These experiments suggest that the levels or activity of key snRNP proteins affect the intron recognition process, especially for weak splice sites.

Binding of snRNPs to pseudo U2-type 5’ splice sites that do not function as splicing donor sites can cause splicing inhibition and these pseudo sites can have alternative roles both as exonic and intronic splicing silencers, as suggested by large-scale screens for SREs (Wang et al., 2004; Yeo et al., 2007). For example, U1 snRNP, when bound to an exonic pseudo splice site, has been shown to inhibit the recognition of a close-by bona fide splice site, possibly by direct steric hindrance or by destabilization of the binding of other splicing components (Cunningham et al., 1991; Siebel et al., 1992; Cloutier et al., 2008). In Rous sarcoma virus such unproductive binding of U1 snRNP has been harnessed to suppress all splicing of the pre-mRNA, which is necessary for the
formation of viral genomic RNAs during the retroviral lifecycle (McNally and McNally, 1999).

In addition to U2-type splicing signals, also U12-type sequences, particularly the 5’ss consensus sequence but also U12-type BPS and 3’ss sequences, were discovered in the above-mentioned SRE screens (Yeo et al., 2007). Remarkably, deletion of a functional U1 snRNP binding site that represses a pseudoexon inclusion can be rescued by U11 snRNP binding site in an in vitro splicing reaction (Dhir et al., 2010). This supports the notion that exon definition interactions between the two spliceosomes are similar and often interexchangable.

Interestingly, U1 snRNP binding sites can also act as splicing enhancers. G nucleotide triplets are enriched in the vicinity of genuine exons and can bind U1 snRNP directly, and their deletion can be rescued by replacing them with 5’ss consensus sequences (McCullough and Berget, 2000). In another example, U1 snRNP binding enhances the use of cryptic upstream 3’ss in U1-70K mRNA (see section 1.6.7; Rösel-Hillgärtner et al., 2013). SREs recognized by snRNPs can recruit other spliceosomal snRNPs and, remarkably, even induce assembly of complete spliceosome-like complexes (Giles and Beemon, 2005; Dhir et al., 2010). Importantly, these complexes differ from genuine assembly intermediates either in composition or conformation, as they do not activate splicing but instead form dead-end complexes.

U1 snRNP extends its influence on the mRNA processing pathway even further at the level of polyadenylation. Evidence of this was first discovered with the observation that the U1A protein can bind to the 3’UTR of its own mRNA and inhibit polyadenylation (Boelens et al., 1993; Gunderson et al., 1994). U1 can also bind to intronic pseudo 5’ss sequences and suppress premature cleavage and polyadenylation, which would otherwise lead to formation of truncated transcripts. In fact there is a need for such machinery as the polyadenylation signal (AAUAAA or AUUAAA and their variants) occurs on average every 2000 nt at genomic scale. This mechanism is most likely ubiquitous, working at the level of the whole transcriptome (Kaida et al., 2010).

1.4.7 Other factors affecting alternative splicing

A number of other aspects can also influence splice site selection. For instance, pre-mRNA folding may bring distant splice sites closer together in space and so enhance their pairing (Charpentier and Rosbash, 1996). Pre-mRNA secondary structure may also sequester splice sites and lead to the exclusion of cassette exons (Clouet d’Orval et al., 1991). In the Drosophila Dscam gene, which can form up to 38 000 different splice variants (see section 1.4.3), alternative cassette selection occurs after the formation of secondary structure between docking and selector sequences surrounding alternative cassette exons. This mechanism ensures that only one of the many alternative cassette
Review of the literature

exons is included in the mRNA (Graveley, 2005). In other instances intronic secondary
structure elements may even function as riboswitches that bind to small molecule
metabolites and activate or repress splicing in response to the concentration of the
metabolite (Cheah et al., 2007).

Additionally, a substantial amount of evidence has accumulated demonstrating that
transcription rate and chromatin structure can affect alternative splicing. This topic will
be discussed in the following section.

1.5 INTERCONNECTIVITY IN NUCLEAR RNA-PROCESSING PATHWAYS

1.5.1 Cotranscriptional processing of pre-mRNA

Pre-mRNA molecules are transcribed from the chromatin template by RNA polymerase
II. The transcription cycle goes through preinitiation, initiation, promoter clearance,
promoter-proximal pausing, elongation, and termination steps. A pre-initiation
complex containing general transcription factors and a hypophosphorylated Pol II
assembles on the promoter, followed by opening of the DNA duplex at the promoter
("promoter melting"), RNA phosphodiester bond formation and the transit of Pol II
away from promoter ("promoter clearance").

Progression through the consecutive steps is regulated via the reversible
phosphorylation of the C-terminal domain (CTD) of Rbp1, the largest subunit of Pol II.
The CTD is a tail-like extension consisting of up to 52 repeats of the heptapeptide
sequence YSPTSPS (reviewed by Lenasi and Barboric, 2013). During initiation and early
elongation, mainly Ser5 is phosphorylated (Komarnitsky et al., 2000). This is needed for
the association of processing factors, and processing begins as soon as the pre-mRNA
emerges from the RNA polymerase, due to the targeting of capping enzymes to the
CTD (McCracken et al., 1997a; Schwer and Shuman, 2011). Later during elongation,
phosphorylation emphasis switches to Ser2, and cleavage and polyadenylation factors
are recruited, leading to transcription termination and release of the newly
synthesized transcript (McCracken et al., 1997b). Thus, the CTD functions as an
organization hub or a "landing pad" for many processing factors, and its
phosphorylation status controls the affinity of different processing factors, ensuring
timely recruitment.

The nascent transcripts go through numerous processing steps before they are ready
to leave the nucleus, including the above-mentioned capping, cleavage and
polyadenylation, and also splicing, editing or other mRNA modifications. Once the
mRNA has matured through these processing steps, it is translocated through the
nuclear pore to the cytoplasm. While many of these processes have been studied
extensively using various in vitro systems, typically in isolation, it is important to bear
in mind that in the nucleus they occur in close proximity and often concurrently (reviewed by Lee and Tarn, 2013).

Nuclear contents are not uniformly distributed; instead, the nucleus contains different microenvironments. Areas of transcriptional repression, known as heterochromatin, tend to localize to nuclear periphery in complex eukaryotes, whereas areas of active transcription, euchromatin, tend to localize to the interior, with channels of active chromatin linking to the nuclear pores (reviewed by Carmo-Fonseca and Carvalho, 2007). In yeast, however, the situation is reversed, and association of active chromatin with nuclear pores is a more preferred arrangement (Casolari et al., 2004). Distinct nuclear bodies, including nucleoli, speckles, paraspeckles, Cajal bodies, histone locus bodies, nuclear stress bodies and promyelocytic leukemia bodies, can also be visualized with high resolution microscopy (reviewed by Dundr and Misteli, 2010). These bodies are thought to facilitate nuclear events by locally increasing the concentration of essential components, and by coordinating separate steps. They are highly dynamic in the sense that components are constantly entering and leaving the compartment, and are thought to form stochastically due to the self-associative properties of the components involved (Hebert and Matera, 2000; Kaiser et al., 2008).

The relationship of transcription and splicing with these nuclear compartments is still somewhat under debate, as the structures are not unambiguously defined. Speckle cores, also known as interchromatin granules, are devoid of active transcription and thus represent sites of snRNP and splicing factor storage and recycling (Misteli and Spector, 1997; Spector and Lamond, 2011). Nascent RNAs are located to the periphery of speckles in areas called perichromatin fibrils (PFs), and snRNPs and splicing factors move away from speckles to PFs to perform splicing. High transcriptional activity at a chromatin locus can also attract a large number of snRNPs, which can appear as a speckle-like body under the microscope, or active chromatin can loop in such a manner that it overlaps with the speckle (Han et al., 2011). Furthermore, a subset of introns can splice post-transcriptionally (see section 1.5.2), and this may cause some transcripts to localize to the vicinity of speckles for completion of splicing.

1.5.2 Cotranscriptionality of splicing
Several interdependencies between transcription and splicing have been described. For a long time it has been known that pre-mRNAs subjected to splicing are associated with chromatin, which suggested that processing steps can occur co-transcriptionally (Baurén and Wieslander, 1994). As truncation of the RNA polymerase CTD leads to the inhibition of splicing (McCracken et al., 1997a), it was conceivable that processing factors, including spliceosome components and SR proteins, can be recruited to the nascent transcript via interactions with the CTD (Mortillaro et al., 1996; Kim et al., 1997). Consistently, it has been shown that U2AF65 copurifies with the CTD and
crosslinks to the very first nucleotides of the nascent RNA emerging from the polymerase (Robert et al., 2002; Ujvári and Luse, 2004). Furthermore, the presence of hyperphosphorylated CTD enhances spliceosome complex assembly (Hirose et al., 1999). Finally, positive feedback with recruitment of splicing factors to the transcription machinery has been described, so that the presence of introns in pre-mRNA stimulated pre-mRNA synthesis (Brinster et al., 1988; Furger et al., 2002; Damgaard et al., 2008). Reciprocally, inhibition of splicing with spliceostatin A, which binds to SF3b inhibiting branch point recognition by U2 snRNP, leads to inhibition of transcription elongation, suggesting a close cooperativity between elongation and processing (Koga et al., 2014).

The recruitment of splicing factors to CTD is largely dependent on the presence of introns in the pre-mRNA. Chromatin immunoprecipitation (ChIP) experiments have revealed that while recruitment of splicing factors takes place cotranscriptionally for transcripts containing introns, such recruitment is not observed with intronless transcripts (Görnemann et al., 2005; Listerman et al., 2006; Moore et al., 2006). An exception to this is U2AF65, which has been observed to be recruited also to intronless genes (Görnemann et al., 2005). As the detected interactions are sensitive to RNase treatment (Moore et al., 2006), it seems that the CTD does not directly recruit splicing factors, but requires concomitant interaction with introns.

It is likely that the majority of splicing is cotranscriptional. Splicing kinetics of newly synthesized pre-mRNAs have been mapped using quantitative PCR (qPCR) and, more recently, with RNAseq (Singh and Padgett, 2009; Khodor et al., 2011), which both reveal that splicing takes place while transcription is ongoing, and that spliced mRNAs associate with the chromatin fraction during cellular fractionation (Pandya-Jones and Black, 2009). Furthermore, the localization of the splicing process has been also investigated using immunological (Girard et al., 2012) and single-molecule techniques (Vargas et al., 2011). These works readily confirm that the majority, but not all, splicing is cotranscriptional. While most of this data arises from studies concentrating on U2-type introns, essentially the same conclusions have been confirmed with U12-type introns as well (Singh and Padgett, 2009). Given that both types of spliceosomes share most of their protein components, and interact with the same protein components known to associate with the CTD (such as SR proteins), this conclusion is hardly surprising. This general conclusion has been challenged in a study claiming a cytoplasmic localization for U12-dependent spliceosome (König et al., 2007). However, subsequent studies on the localization of the minor spliceosome and splicing of U12-type introns, and the demonstration of cotranscriptional nature of the splicing of minor introns, have refuted this single claim and demonstrated that on this level both
spliceosomes function similarly (Friend et al., 2008; Pessa et al., 2008; Singh and Padgett, 2009).

While there is a consensus that splicing is mostly cotranscriptional, there is also evidence of post-transcriptional splicing. For example, there are conflicting reports concerning yeast, finding either that the majority of transcripts are cotranscriptionally spliced (Carrillo Oesterreich et al., 2010), or that transcription is finished before splicing completes, especially in genes whose last exons are shorter than 1 kb (Tardiff et al., 2006). Similarly, it has been suggested that in mammalian cells at least some of the introns, particularly those that are alternatively spliced, can be spliced post-transcriptionally (Vargas et al., 2011; Tilgner et al., 2012).

Cotranscriptionality of splicing is also underscored by the results that link transcription elongation rate to the choice of alternative splice sites. In constitutive splicing there is a general tendency to follow a “first come, first served” principle, so that the first intron to emerge from the RNA polymerase is engaged by splicing factors first (Beyer and Osheim, 1988). However, this is not always the case, and the order of removal may instead reflect the rate of intron recognition or exon definition (Kessler et al., 1993). It has been observed that pausing or slowly elongating RNA polymerase II mutants affect the inclusion of cassette exons (Roberts et al., 1998; de la Mata et al., 2003). Slow elongation favors the use of weak splice sites, because at a slow elongation rate there is more time for spliceosome assembly before a strong site emerges from the RNA polymerase.

Analogously, the chromatin landscape that affects the rate of elongation also has an impact on splicing. Chromatin remodeling by the SWI/SNF complex affects alternative exon choice (Batsché et al., 2006). Global relaxation of chromatin structure after concerted histone acetylation causes relocation of splicing factors to nuclear speckles away from chromatin (Schor et al., 2012). Spliceosome components may be recruited directly to the chromatin via, for instance, the H3K4me3-recognizing chromodomain protein CHD1 (Sims et al., 2007). Interestingly, the influences seem to work in both directions – H3K36me3 histone methylation marks and the positioning of nucleosomes localize preferentially to exons instead of introns (Kolasinska-Zwierz et al., 2009; Tilgner et al., 2009). It is not yet known whether exon-enriched chromatin marks are deposited as a result of splicing activity, or whether splicing is regulated by pre-deposited chromatin marks. The latter view is supported by the observation that even non-transcribed genes retain exon-enriched marks (Tilgner et al., 2009); on the other hand, no mechanism has yet been proposed that would explain the circular argument of regulating splicing with chromatin marks and chromatin marks with splicing.
1.5.3 Kinetics of splicing and export

The removal rate of different introns on a given transcript can vary, and some introns are retained even after transcription has completed and cleavage and polyadenylation have occurred. In general, splicing rate is modulated by the strength of splice sites and presence or absence of splicing enhancers (Hertel and Maniatis, 1998; Hicks et al., 2010). In the case of U2-dependent splicing, the snRNPs have been shown to diffuse freely in the nucleoplasm, and therefore their availability does not restrict the rate of splicing (Rino et al., 2007). Neither are differences in splicing rate explained by the binding rates of snRNPs (U1, U2, U5 or the NTC) onto yeast pre-mRNA (Shcherbakova et al., 2013). As stated above, alternative splicing is especially prone to delayed splicing (Vargas et al., 2011). Still, unspliced introns typically remain associated with chromatin in a DNA-dependent manner, and are released to the nucleoplasm only after splicing is completed (Pandya-Jones et al., 2013). However, unspliced transcripts may also be retained at the nuclear pore, but this has only been observed in yeast (Legrain and Rosbash, 1989; Dziembowski et al., 2004). After release from chromatin, export to the cytoplasm is efficient in mammals (Custódio et al., 1999; Audibert et al., 2002).

Therefore, splicing rate is typically a major determinant of how quickly an induced transcript is available for the translation machinery, if there are no other regulatory steps causing delays (i.e. mRNA intracellular localization or silencing at the translational level).

Analysis of a subset of U12-type introns has revealed that these introns are often processed more slowly than the major introns. In vitro splicing experiments revealed at least three to five-fold reduction in the rate of splicing compared to major introns (Tarn and Steitz, 1996b; Frilander and Steitz, 1999). This observation is supported by in vivo experiments that also showed slower accumulation of spliced products both with a reporter construct containing U12- and U2-type introns in the same context, and with endogenous genes (Patel et al., 2002). Similar results were obtained with quantification of transcription and splicing kinetics by RT-qPCR (Singh and Padgett, 2009). Furthermore, in vivo experiments on the steady state situation also reveal that U12-type introns remain partially unspliced, both in the case of splicing reporters and with endogenous U12-type introns (Bozzoni et al., 1984; Santoro et al., 1994; Patel et al., 2002; Pessa et al., 2006). Finally, it has been found that the presence of a U12-type intron can lead to reduction in the amount of protein produced, as described with premRNA injected to Xenopus oocytes (Bozzoni et al., 1984). Conversely, protein yields from a reporter have been reported to increase 6 to 8-fold when a U12-type intron was changed to U2-type (Patel et al., 2002).

The mechanistic reasons for the slower rate of splicing are currently unknown. Although minor spliceosome components are approximately 100-fold less abundant in
the nucleus, this is not the limiting factor as splicing efficiency increases linearly with increasing transfection efficiency of an exogenous reporter (Patel et al., 2002). Neither does an additional 10-fold reduction in U4atac levels further impair splicing efficiency for endogenous introns (Pessa et al., 2006). Spliceosome assembly may be less flexible in the U12-dependent spliceosome, given that the intron recognition occurs cooperatively (Frilander and Steitz, 1999). Alternatively, tri-snRNP recruitment may also be a rate-limiting step, as the tri-snRNP proteins are shared between major and minor tri-snRNPs, so unproductive recruitment of major tri-snRNP to the U12-dependent spliceosome may occur (Schneider et al., 2002).

Regardless of the underlying mechanism, the slow removal of U12-dependent introns has been proposed to constitute a rate-limiting step in the processing of transcripts containing these introns (Patel et al., 2002). In the cellular context, this represents a point of regulation whereby the expression of U12-type intron-containing genes could be regulated by modulating the efficiency of splicing. For example, it is known that several protein kinase inhibitors and histone acetyltransferase and deacetylase inhibitors have an inhibitory effect on splicing (Parker and Steitz, 1997; Kuhn et al., 2009). Similarly, it has been also shown that signaling via the stress-activated p38MAPK pathway can regulate U6atac expression levels, suggesting that the splicing of U12-type introns may be connected to cellular growth conditions (Younis et al., 2013).

Regulation of gene expression by U12-dependent spliceosome could explain why a parallel spliceosome is maintained in many eukaryotic genomes. This hypothesis is supported by the conservation of U12-type introns in certain gene families, and more strikingly, the presence of U12-type introns in nonhomologous positions in paralogous genes (Burge et al., 1998).

1.6 RNA QUALITY CONTROL

1.6.1 Nuclear quality control

Pre-mRNA processing in the nucleus is inherently noisy, that is, there is variation in processing and splicing not subject to regulation or conservation between species (Melamud and Moult, 2009; Pickrell et al., 2010). Such aberrant transcripts should be eliminated before they are translated to defective proteins, which is wasteful energy use by the cell and may even produce harmful truncated proteins lacking critical domains. While nonsense-mediated decay (NMD) targets frameshift transcripts during the pioneer round of translation (see section 1.6.6), RNA quality control targeting aberrant mRNAs begins already in the nucleus. The main mechanisms for quality control are nuclear decay and/or retention at the transcription site (Schmid and Jensen, 2010). Both the export and decay machineries are linked physically to processing and are recruited to the elongating polymerase, ready to monitor the processing as it
occurs on the chromatin (Andrulis et al., 2002; Kim et al., 2004; Moore et al., 2006; Hessle et al., 2009). Inhibition of splicing by mutated splicing factors or mutation of splicing reporters causes accumulation of the pre-mRNAs if decay is also inhibited (Bousquet-Antonelli et al., 2000). Capping and polyadenylation reactions are also monitored, and failure to perform these processing steps leads to retention and/or degradation (Hilleren et al., 2001; Jimeno-González et al., 2010). The decay can proceed both from 5’ to 3’ or 3’ to 5’, and the exoribonuclease (exoRNase) activities of both Xrn2/Rat1p and the RNA exosome complex are involved in nuclear quality control (Bousquet-Antonelli et al., 2000; de Almeida et al., 2010; Davidson et al., 2012).

1.6.2 Exosome

The exosome is a multiprotein complex with ribonuclease activities and consists of a core of nine subunits, which are present in both in the nucleus and cytoplasm of eukaryotic cells. The structure is highly conserved from archaea to metazoans, and is also very similar to the eubacterial polynucleotide phosphorylase (PNPase) complex (Schilders et al., 2006). The core subunits of the eukaryotic exosome have lost their catalytic activity and the ribonuclease activity is instead provided by the associated active subunits Rrp6 (also known as PM/SCL-100 in mammalian system) or Rrp44 (DIS3; Liu et al., 2006; Dziembowski et al., 2007). In humans, two additional homologs of DIS3 have been discovered, called DIS3L1 and DIS3L2, both of which localize to the cytoplasm but show differential association with the exosome core, as DIS3L1 associates with it while DIS3L2 does not (Staals et al., 2010; Tomecki et al., 2010; Astuti et al., 2012; Malecki et al., 2013). RRP6 and DIS3L1 exoRNases degrade RNA only in the 3’ to 5’ direction, whereas DIS3 has both 3’ to 5’ exonucleolytic and endonucleolytic activities (Lebreton et al., 2008).

The nine subunits of the core exosome form a barrel-like structure containing six RNase PH-like proteins in a ring (Rrp41, Rrp45, Rrp46, Rrp43, Mtr3 and Rrp42) and three S1/KH proteins as a cap on one side of the ring (Csl4, Rrp4 and Rrp40). Rrp44/DIS3 is located on the opposite side of the ring from the S1/KH proteins. The RNA takes a path through the barrel from the S1/KH side, through the barrel and into the exonucleolytic active site of Rrp44/DIS3 (Makino et al., 2013). The exosome is able to process structured RNAs because they are unwound during entry to the exosome channel (Bonneau et al., 2009).

1.6.3 Exosome targets

The exosome targets a multitude of RNA species in the cell. It is involved in the maturation of rRNA (ribosomal RNA), snRNAs and snoRNAs, degradation of RNAs that are generated by pervasive transcription such as PROMPTs (promoter upstream transcripts), CUTs (cryptic unstable transcripts), SUTs (stable uncharacterized
transcripts) and XUTs (Xrn1-dependent unstable transcripts), different types of mRNA decay such as nonsense-mediated, nonstop and no-go decay as well as AU-rich element (ARE-) mediated decay (reviewed by Chlebowski et al., 2013). The exosome is also involved in a checkpoint of nascent pre-mRNAs, as pre-mRNAs that fail to gain a poly-A tail are retained at the transcription site in an exosome-dependent manner and are further degraded by the exosome at the same location (Bousquet-Antonelli et al., 2000; Hilleren et al., 2001). Recent transcriptome-wide work has indicated that the exosome targets more than half of the yeast intron-containing genes (Gudipati et al., 2012). Similarly, crosslinking and cDNA sequencing (CRAC) experiments in yeast revealed that DIS3 binds to intron-containing pre-mRNAs (Schneider et al., 2012).

It is difficult to conceive how an enzyme complex is able to target such variety of substrates, yet do so in a controlled manner. One solution to this problem is that the inactive core restricts the ribonuclease activities; another is that there are several cofactor complexes that help in selecting targets. Of these, the Trf4/5p-Air1/2p-Mtr4p polyadenylation (TRAMP) complex is an important auxiliary complex located in the nucleolus and facilitates exosome activity by the helicase activity of Mtr4p and by the addition of unstructured oligo-A stretches to RNA 3’ ends (Jia et al., 2011). In humans, MTR4 also associates to the nuclear exosome-targeting (NEXT) complex (Lubas et al., 2011). MPP6 and C1D are additional factors that associate with the human and yeast exosomes and are required for exosome targeting to rRNA (Schilders et al., 2005; Schilders et al., 2007).

### 1.6.4 Other nuclear degradation pathways

In addition to the 3’ to 5’ decay pathway, RNAs can be degraded in the 5’ to 3’ direction. In the nucleus, this activity is provided by the XRN2 exonuclease (in yeast Rat1p). This activity requires a free 5’end. Because the 5’ ends of the nascent pre-mRNA transcripts are capped with an m7G cap as soon as they emerge from the RNA polymerase and are bound by the cap-binding complex (CBC), they are protected from XRN2 degradation. XRN2 can target pre-mRNAs if the cap is removed by decapping enzymes such as DCP2. This can take place in the nucleus during pausing of Pol II -mediated transcription (Brannan et al., 2012). This leads to premature termination by the so-called torpedo model, where the pre-mRNA is degraded in the 5’ to 3’ direction until the exoRNase reaches the polymerase, and, together with other protein factors, causes the polymerase to detach from the template (Luo et al., 2006; Brannan et al., 2012). Decapping is also subject to quality control as defectively capped pre-mRNAs accumulate in decapping deletion mutant yeast strains (Jiao et al., 2010). In mammals, it is unclear whether XRN2 participates in quality control of capping. However, XRN2 has been shown to play a role in transcription regulation (Brannan et al., 2012; Eberle
and Visa, 2014), and to monitor and degrade aberrantly spliced transcripts (Davidson et al., 2012).

1.6.5 Kinetic proofreading hypothesis
As it has been difficult to pinpoint the mechanisms that would directly target aberrant or unstable transcripts to the exosome, an alternative hypothesis on the selection of exosome targets has been proposed (Burgess and Guthrie, 1993; Doma and Parker, 2007). According to the kinetic proofreading hypothesis, the fate of a transcript is determined by competition between the rate of processing reactions and the rate of decay reaction. Therefore, any transcripts showing delayed processing are more likely to be degraded, and vice versa, in order to escape the kinetic surveillance, the rate of the normal processing reaction should be optimal. In the case of pre-mRNA processing, aberrant transcripts are retained in the nucleus, and correspondingly, splicing enhances nuclear export (Legrain and Rosbash, 1989; Custódio et al., 1999; Zhou et al., 2000; Audibert et al., 2002; Cheng et al., 2006). A clear advantage from such a targeting mechanism is that many kinds of defects can be recognized in this manner, and no specific features of the transcript need to be recognized. This seems to be applicable to the diverse nature of exosome targets.

1.6.6 Nonsense mediated decay
Cytoplasmic quality control represents a second checkpoint for monitoring the accuracy of processed mRNAs. The NMD pathway targets transcripts that contain premature termination codons (PTCs; Leeds et al., 1991; Lykke-Andersen et al., 2000). In most cases it relies on the deposition of exon junction complexes during the process of splicing (Zhang et al., 1998; Ishigaki et al., 2001). An EJC marks the spot on mRNA where an intron has been excised from. After export from the nucleus, mRNAs associate with ribosomes in a primary round of translation. This process clears EJCs from the mRNA, up to the point when the ribosome reaches a stop codon. At a normal termination codon, release factors eRF1 and eRF3 together with poly-A binding protein PABP (PABPC1 in mammals) stimulate efficient translation termination. If more EJCs remain on the mRNA, that is, downstream of the stop codon, the release factors interact with EJC proteins Upf2 and Upf3b to recruit Upf1, which is thought to trigger degradation (Kashima et al., 2006). The distance between PTC and a downstream, triggering EJC must be more than ~50 nt due to the size of the translation termination complex and EJC. Termination at a PTC has also been found to be inefficient due to lack of interaction with the poly-A tail and may in such case be independent of an EJC (Amrani et al., 2004; Bühler et al., 2006; Ivanov et al., 2008). Instead, the activation is determined by the length of the 3′UTR and the escape of transcripts with short 3′UTR from NMD requires the binding of poly-A binding protein PABPC1 (Eberle et al., 2008; Singh et al., 2008).
1.6.7 Regulation via NMD

Besides targeting transcripts that contain premature termination codons due to splicing errors, NMD is also activated in a regulated manner to control the level of mRNAs. In this instance it often occurs in conjunction with alternative splicing (AS) and is known by the acronym AS-NMD. Initial computational estimates suggested that for genes with EST evidence for alternative splicing, as many as 35% could be potential NMD targets (Lewis et al., 2003). Examination of the conservation of PTC+ isoforms between mouse and human confirmed the importance of some AS-NMD events but placed the estimate of frequency slightly lower, at 21% (Baek and Green, 2005). Large-scale experimental evidence from yeast and mammalian cells suggests that 3-10% of transcripts are targeted by NMD (He et al., 2003; Mendell et al., 2004).

Alternative splicing coupled to NMD appears to be especially common in the regulation of splicing factor expression. This phenomenon was discovered in the SR splicing factor family, which contain ultraconserved regulatory elements that lead to expression of alternative splice forms, targeted by NMD (Lareau et al., 2007). Similar autoregulation depending on AS-NMD has been described for PTB and other genes (Wollerton et al., 2004; Ni et al., 2007). Overexpression of the protein in question then leads to increased inclusion of the PTC-containing exon, degradation by NMD, and a reduction of splicing factor expression in order to restore homeostasis (Saltzman et al., 2008).

An snRNP can also participate in a negative feedback loop. Work by Rösel-Hillgärtnet al. (2013) has shown that the U1 snRNP regulates its own proteins by binding to cryptic 5′ splice sites in an alternative cassette exon and enhancing the recognition of an upstream 3′ splice site. Three cryptic sites in the U1-70K gene exert an additive effect on the inclusion of an NMD-sensitive cassette exon, and the binding is influenced by the levels of U1C and U1-70K proteins. As a result, the splicing choice regulates U1-70K protein levels in a U1C-dependent manner.
2 AIMS OF THE STUDY

This project began with the identification of a conserved sequence element in the genes SNRNP48 and RNPC3, encoding two proteins of the minor spliceosome, U11-48K and U11/U12-65K, respectively. The aims were to characterize the alternative splicing regulated by this element, its activation in cells and the consequences for transcript stability and localization. As the sequence element contained two matches to the U12-type 5’ splice site consensus sequence, we hypothesized that there exists an autoregulatory loop controlling the homeostasis of the U11/U12 di-snRNP.

Following this, the aim was to further characterize the distance requirements between the two 5’ss-like motifs and the splice sites they control. Since the alternative splice site activation requires exon definition interactions to form between U12- and U2-type spliceosomes, I investigated the communication between the spliceosomes and specifically, the role of the U11-35K protein in it.

A long-standing question in the minor spliceosome field has been to pinpoint a role for the maintenance of a parallel spliceosome in the genomes of many organisms, especially since it has been lost in certain lineages. The final aim of this study was to test the possibility that rate-limiting splicing by the minor spliceosome plays a role in the nuclear processing of U12-intron containing transcripts on a transcriptome-wide scale. This was done by studying the exosome-mediated degradation of pre-mRNAs.
3 MATERIALS AND METHODS

Methods used in this study are listed in Table 2. For a detailed description of methods, see the original publications.

Table 2. Methods used in this study.

<table>
<thead>
<tr>
<th>Method</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioinformatic analysis</td>
<td>III</td>
</tr>
<tr>
<td>Cell fractionation</td>
<td>I, III</td>
</tr>
<tr>
<td>Cell lines and culture</td>
<td>I, II, III</td>
</tr>
<tr>
<td>Immunofluorescence</td>
<td>II</td>
</tr>
<tr>
<td>Immunoprecipitation</td>
<td>I, II</td>
</tr>
<tr>
<td><em>In vitro</em> splicing assays and spliceosome assembly</td>
<td>I</td>
</tr>
<tr>
<td><em>In vivo</em> splicing reporter assays</td>
<td>I, II</td>
</tr>
<tr>
<td>Luciferase expression assays</td>
<td>I</td>
</tr>
<tr>
<td>Northern blotting</td>
<td>I, III</td>
</tr>
<tr>
<td>Protein overexpression in mammalian cells</td>
<td>I, II, III</td>
</tr>
<tr>
<td>Quantitative RT-PCR</td>
<td>I, III</td>
</tr>
<tr>
<td>RNA sequencing</td>
<td>III</td>
</tr>
<tr>
<td>RNA-RNA crosslinking</td>
<td>I</td>
</tr>
<tr>
<td>RNAi knockdown</td>
<td>I, III</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>I, II, III</td>
</tr>
<tr>
<td>Sequencing library construction</td>
<td>III</td>
</tr>
<tr>
<td>Site-directed mutagenesis</td>
<td>II, III</td>
</tr>
<tr>
<td>Streptavidin pulldown assays</td>
<td>I</td>
</tr>
<tr>
<td>Transfection</td>
<td>I, II, III</td>
</tr>
<tr>
<td>Western blotting</td>
<td>I, II, III</td>
</tr>
</tbody>
</table>
4 RESULTS AND DISCUSSION

4.1 THE LEVELS OF U11/U12 DI-SNRNP ARE REGULATED THROUGH A CONSERVED FEEDBACK MECHANISM (I, II)

4.1.1 Genes encoding U11-48K and U11/U12-65K contain highly conserved sequence elements (I, II)

The recognition of U12-type introns relies on the cooperative binding of U11/U12 di-snRNP on the 5′ splice site and branch point sequences (Frilander and Steitz, 1999). Sequence-specific binding is achieved by both RNA-RNA base-pairing and protein-RNA interactions, with U11-48K playing an important role in the 5′ss recognition (Turunen et al., 2008; Tidow et al., 2009). In the course of our study into the 48K gene, SNRNP48, we discovered a duplication of the U12-type 5′ss sequence within its fourth intron. EST evidence suggested that neither site functions as a splice donor.

Remarkably, the region containing the 5′ss duplication within the 48K intron 4 was conserved in evolutionarily very distant organisms. A closer inspection revealed that the duplicated 5′ss motif was present in the same intronic location of the 48K gene in mammals, fishes and many insect species (see Figure 1A in I). In mammals the high level of conservation is not limited to the duplicated 5′ss motif, but also includes a ~110 bp region upstream of this motif. This region is nearly 100% conserved and contains a novel 8 nt exon, designated as exon 4i, which is surrounded by near-consensus PPT, 3′ss, and 5′ss elements located ca. 50 bp upstream of the duplicated 5′ss motif (I, Figure 1A). Outside the mammalian lineage the level of conservation decreases, particularly in the region between exon 4i and the duplicated 5′ss motif. Between human and fishes the conservation is limited to the splice signals, the 4i exon, and the duplicated 5′ss motif, while between humans and insect species only the PPT, 3′ss and duplicated 5′ss motif are recognizable. Together, the sequence conservation suggests that in addition to the duplicated 5′ss motif, the upstream splice sites constitute the main functional entities of this conserved intronic element. Surprisingly, the duplicated 5′ss motif together with an upstream 3′ss was also detected in plant 48K genes, but in the 3′UTR instead of an intronic location (Figure 1C in I). The presence of this motif in the same gene from humans to plants suggests a critical cellular or organismal function.

Importantly, EST data provided evidence for two alternatively spliced mRNA isoforms containing exon 4i. Both show an activation of an alternative 3′ss upstream of the duplicated U12-type 5′ss-like motif and inclusion of exon 4i sequences, either alone or together with intronic sequences downstream of exon 4i (see Figure 5). EST data was confirmed by RT-PCR detection of both exon 4i-containing isoforms from endogenous
genes and reporter constructs (Figure 4A in I; Turunen et al., 2013b). Similarly, plant isoforms were identified by RT-PCR from *A. thaliana* and *Populus trichocarpa* (Figure S1 in I).

![Figure 5. 48K and 65K splice isoforms. Black boxes represent coding exons, black lines represent introns, and grey boxes represent untranslated regions.](image)

Both 4i-containing isoforms are predicted to elicit NMD as they lead to appearance of premature stop codons in the mRNA. Exon 4i inclusion leads to a frameshift in the coding sequence and the appearance of a PTC close by downstream in exon 5. With the other isoform, which contains exon 4i and downstream intronic sequences, the intron retention similarly leads to the insertion of a downstream PTC. These observations led to a working model in which the 5’ss-like motifs could function in feedback regulation in a manner similar to AS-NMD described for SR proteins (Lareau et al., 2007; Ni et al., 2007). In this hypothetical scenario, the binding of U11 snRNP or U11/U12 di-snRNP would lead to activation of an upstream 3’ss, PTC inclusion and a decrease in 48K mRNA levels via NMD. Given the central role of the 48K protein in 5’ss recognition, this would establish a negative feedback loop to regulate the 48K protein levels and cellular homeostasis of U12-type intron recognition. The putative splicing regulatory element containing the two U12-type 5’ss-like motifs was named USSE, for U11 snRNP-binding splicing enhancer.

Bioinformatic searches for USSE-like elements elsewhere in the genome resulted in only one other hit that was evolutionarily highly conserved, namely RNPC3. This gene codes for the 65K protein, which, similarly as the U11-48K protein, is a component of the U11/U12 di-snRNP. Also in this case the USSE element is embedded within a highly conserved region, located in the 3’UTR, thus resembling the plant 48K USSE (Figure 1B in I; see also above). It is also associated with an upstream PPT and 3’ss, along with isoforms displaying alternative splicing in the 3’UTR as detected by RT-PCR analysis.
The 65K USSE element is conserved within vertebrates but was not detected in other species.

In addition to sequence conservation, the distance between upstream 3’ss and USSE, and also the distance between the two 5’sss-like elements within the USSE sequence, both show a striking evolutionarily conservation (Figure 1 in II). In terrestrial vertebrates (Tetrapoda), an overwhelming majority displayed exactly 63 nt distance between 48K USSE and the upstream 3’ss, with only a few species that had a 1-2 nt variation in either direction. In the same group of species, 65K 3’ss-USSE distance was exactly 44 nt with only 1 nt variation. In fishes, the 3’ss-USSE distance was slightly more variable; nevertheless, fish 48K 3’ss-USSE distances clustered closer to other vertebrates’ 48K 3’ss-USSE distance and fish 65K 3’ss-USSE with other vertebrates 65K 3’ss-USSE distance. Insect 48K 3’ss-USSE distance showed the largest variation, with a range between 14 and 72 nt with little obvious clustering. In plants, the 3’ss-USSE distance is clustered between 34 and 41 nt and thus was similar to that in 65K gene. This may reflect the fact that in both cases the USSE is located within the 3’UTR. However, unlike in the mammalian 65K, there is more variation in element spacing in plants, for example, a few plant species show very long 3’ss-USSE distances (up to 121 nt).

Similar conservation was also observed with the length distribution of the spacer separating the two 5’sss-like elements within the USSE (Figure 1B in II). With tetrapod species 48K USSE seems to favor a 9 nt spacer length with only a few exceptions. Interestingly, 65K USSE in the same phylogenetic group typically has a 6 nt spacer, again with very little variation. Plant USSE spacers are most often 3 nt in length with only a few exceptions. Fish and insect spacer lengths are more variable and do not seem to correlate with location of the USSE (48K vs. 65K).

Together, the phylogenetic data suggests that the upstream 3’ss (including PPT) and the USSE sequence itself are the central elements needed for alternative splicing regulation. In addition to sequence conservation, both the distance between the 3’ss and USSE and the length of the spacer between the 5’sss-like motifs are maintained within strict limits. The strong evolutionarily conservation detected in the phylogenetic analysis suggests that these parameters are important at cellular or organismal level and subject to purifying selection. It also suggests that deviations from these parameters would have downstream consequences at the level of U12-type intron recognition or splicing.

4.1.2 Both USSE motifs bind U11/U12 di-snRNP (I, II)

To confirm the proposed model of the feedback regulation above, the binding of the U11/U12 di-snRNP to the USSE was investigated initially using in vitro splicing
conditions. *In vitro* transcribed RNA substrates containing 48K USSE sequences were incubated in HeLa nuclear extracts and the resulting complexes were resolved on nondenaturing gels (Figure 2 in I). Two complexes migrating similar to a spliceosomal A complex were observed with wt RNA, and were lost with an RNA substrate that contained the A+3G mutations that inhibit binding of the 48K protein to the canonical U12-type 5’ss (Figure 2B in I; Turunen *et al*., 2008). The binding properties of the USSE element were studied further with pulldown of biotinylated RNA substrate with either 48K or 65K USSE element, followed by the identification of copurifying small RNAs on northern blots (Figures 2D-F in I). In these experiments it was observed that both U11 and U12 snRNAs associate with USSE sequences, and the association is dependent on the functionality of the 5’ss-like motifs. Mutations that disrupt the binding of either 48K (A+3G as above) or U11 snRNA (CC+5/6GG) lead to reduction of U11 and U12 snRNA association. To achieve a complete loss of U11/U12 binding, both 5’ss-like motifs need to be mutated.

The above results were confirmed by psoralen crosslinking (Figure 3 in I). Both U11 and U1 snRNAs, but not U12, formed crosslinks with 48K substrate. U11-specific crosslinks were dependent on the functional U12-type 5’ss sequence suggesting that U11 snRNA base-pairs with the USSE sequence, while the lack of crosslinks to U12 snRNA suggests it does not form base-pairing interactions with the substrate but is associated to USSE in the pulldown experiments (Figures 2D-F in I) due to the recruitment of the entire di-snRNP. In contrast, U1 associates with the U2-type 5’ss that is present in the crosslinking substrates (Figure 3B in I). The indispensable role of U11 snRNA on USSE recognition was unequivocally shown with a rescue experiment in which USSE mutations that prevent base pairing between U11 snRNA can be rescued by coexpression of a U11 snRNA containing compensatory mutations *in vivo* (Figure 4F in I).

One of the key questions was the function of the 5’ss duplication within the USSE element: are two U11 (or U11/U12) snRNPs binding to the two 5’ss sequences, or is the function of the duplication to increase the probability that a single U11 snRNP binds? In this respect the *in vitro* experiments provide ambiguous results: with the pull-down experiment (Figure 1 in I), mutations to either 5’ss element decrease the amount of U11 snRNA in the pellet roughly by 50 %, which suggest either that two U11 snRNPs are binding simultaneously, or that high-affinity binding of a single U11 requires a duplicated site. In contrast, in crosslinking experiments (Figure 3 in I) mutations of individual 5’ss elements did not affect the crosslinking efficiency of the other site. This suggests that a single 5’ss site is sufficient for U11 binding.

In contrast, both the *in vivo* and phylogenetic data suggest that both U11 binding sites are needed for the activation of the upstream 3’ss. If either one is targeted by
mutations that prevent the binding of U11-48K protein (A+3G) or U11 snRNA (CC+5/6GG or CT+6/7AG) the result is a complete loss of alternative splicing with either 48K or 65K reporter construct (Figures 4A,B,F in I). Strong evidence supporting the binding of two U11 snRNPs to a single USSE sequence comes from the rescue experiment mentioned above (Figure 4F in I). This experiment showed that two different U11 snRNAs (wt and one containing compensatory mutations) are needed for the 3’ss activation when one of the 5’ss motifs has been mutated to prevent the binding of wt U11 snRNA. The ultimate support to this model comes from experiments aiming to identify the minimal 5’ss -5’ss spacer length within the USSE element able to support the activation of the upstream 3’ss splice site (Figure 1D in II). There, reduction of the spacer length below 3 nt resulted in a strong inhibition of activation of upstream 3’ss. Given that even with 0 nt spacers both 5’ss motifs are expected to be fully functional for binding of a single U11 snRNP, the interpretation of these results is that 5’ss elements that are too close lead to a steric clash of the two U11/U12 di-snRNP particles, reducing or disabling the activity of the USSE.

Interestingly, lengthening the 65K spacer beyond the wild type 6 nt resulted in only slight reductions in splicing efficiency up to 30 nt (longest spacer tested; Figure 1E in II). A 9-nt spacer, which is the length of the majority of vertebrate 48K USSE spacers, supported normal levels of the long isoform. This result supports the above notion that both 5’ss sequences within the USSE element are occupied by a U11 snRNP, but also indicate that there is a stabilizing interaction between the complexes assembled on the individual 5’ss sequences. In contrast to the situation where spacer shortening can lead to a steric clash, the minimal effect of long spacers can be explained by flexibility of the single-stranded RNA spacer, which allows the spacer to loop out to support interactions between distantly spaced U11 snRNPs.

Together, both the in vitro and in vivo data indicate that USSE elements are recognized by U11 snRNP or U11/U12 di-snRNP. U12 snRNP most likely does not have any direct role in the USSE recognition, but rather is a passive component participating through binding to U11 snRNP. In vivo evidence provides a strong support that both 5’ss elements are recognized simultaneously. It is possible that the two U11 or U11/U12 snRNPs bound to USSE are stabilizing each other’s binding through currently undiscovered interactions. The role of such interactions would be to stabilize U11:5’ss recognition in the absence of U12:BPS interaction that normally occurs during the splicing of U12-type introns (Frilander and Steitz, 1999).

4.1.3 Feedback mechanism relies on alternative splicing and exon definition interactions (I, II)

To confirm the biochemical evidence on USSE-dependent activation of the upstream 3’ss we carried out in vivo investigations targeting both endogenous 48K and 65K genes
and also used splicing reporters containing the USSE elements. The 48K USSE reporter contained a previously described SmE minigene reporter (Pessa et al., 2006), with an insertion of the 48K 4i exon and surrounding sequences. Transfected to HEK293 cells, the wild-type reporter mRNA is processed to yield two isoforms, the majority of which excluded exon 4i and a minority with exon 4i inclusion (Figure 4A in I). Mutation of 48K (A+3G mutation) or U11 binding sites (CC+5/6GG) abolished 4i exon inclusion. Similar responses were observed for a reporter containing a firefly luciferase gene in association with the 65K 3’UTR (Figure 4B in I). This reporter produced a short 3’UTR when no USSE activation takes place and a long 3’UTR when USSE is activated, as supported by mutation and oligo blocking experiments. Analogously to the 48K USSE reporter, mutations of 48K (A+3G as above) or U11 (CC+5/6GG as above or CT+6/7AG) binding sites abolished the long isoform.

Next, the endogenous alternative splicing pathway was investigated using overexpression of 48K cDNA from plasmid ("r48K") and antisense oligonucleotides binding to USSE to block U11 snRNA base-pairing ("block" oligo). A mock oligo binding to a nonconserved downstream site was used as a control. We found that the overexpression of 48K dramatically increases the inclusion of exon 4i, and this splicing activation was attenuated with the blocking oligo (Figure 4C in I). The endogenous 65K alternative splicing also responded to the oligo block (Figure 4D in I). These results support the previous result showing that the USSE sequence is necessary for the activation of the upstream 3’ss.

We then tested whether reduction of minor spliceosome protein levels would affect the feedback loop. We inactivated 48K, 35K, or 20K using RNAi, and found that the knockdown of 35K produced a 3-4 fold upregulation of both 48K and 65K mRNAs (Figure 5C in I). This knockdown also caused a shift in the 65K 3’UTR isoform ratio, leading to a marked increase in the short isoform levels and an almost complete loss of the long isoform (Figure 5E in I). Knockdown of 48K produced a similar upregulation and isoform switch in 65K mRNA, whereas knockdown of 20K, another protein of the di-snRNP but with no known function in 5’ss recognition, had very little effect (Figure 5D, E in I). The strong effect of 35K knockdown suggests that 35K is involved in interactions that enhance recognition of the upstream U2-type splice sites. This is supported by the observation that the 35K protein structure is similar to that of U1-70K, which in the major spliceosome is involved in exon definition interactions (Will et al., 2004).

Since 35K knockdown had such a dramatic effect on the 48K and 65K mRNA levels, we investigated the USSE-mediated activation of alternative splicing using reporters that replace USSE with four λN hairpin loops (Gehring et al., 2008) within the context of the 65K 3’UTR (Figure 2A in II). 35K was tagged with the λN peptide that binds to the
Results and discussion

After introducing the hairpin, and with a V5 peptide for immunodetection. The hairpin reporter on its own was inactive with regard to splicing to the long isoform 3’ss, but tethering 35K to the reporter restored alternative splicing, albeit with lower efficiency than seen with the wt USSE (Figure 3C in II). Mutating the 35K RRM to prevent association with the U11 snRNA did not eliminate reporter splicing; in fact, there was a slight increase in long isoform levels. This suggests that the RS domain of 35K is the most likely candidate for interacting with upstream factors involved in 3’ss recognition. To confirm this interpretation we inactivated the RS domain by changing all its arginine residues to alanines. This led to a loss of 3’ss activity and is consistent with the well-known role of the RS domain in splicing activation and localization (Cazalla et al., 2002). Consistently, the paralog of U11-35K in the major spliceosome, U1-70K, and its RS domain alone, was able to substitute for 35K in 3’ss activation when tethered to reporter construct, leading to 3’ss activation at a level similar to the wt USSE. The most likely explanation for strong activation with U1-70K constructs is that it contains two stretches of DR/ER/SR dipeptide repeats, whereas 35K has only one stretch, which is equivalent in length to one 70K RS repeat. Earlier studies with other splicing factors have shown that the number of SR dipeptides is proportional to the potency of activation (Philipps et al., 2003).

Given the phylogenetic conservation of the 3’ss - USSE distance (see 4.1.1), we asked whether there is a minimal distance in the human 65K USSE system needed to support exon definition interactions between the USSE and the upstream 3’ss. We found a progressive loss of 3’ss activation upon shortening of the 3’ss - USSE distance, with virtually no 3’ss activity when the distance was 0 nt (Figure 1D in II). This observation is consistent with the hypothesis that the 3’ss - USSE distance has been selected to accommodate the binding and interaction of U11 snRNP and the 3’ss-binding factors. In exon definition of constitutively spliced exons, it has been observed that shortening of an exon from 50 to 33 nt causes it to be skipped (Dominski and Kole, 1991), and similar results have been obtained with alternative splice sites regulated by U1 snRNP (Hwang and Cohen, 1997). This distance may represent an approximate lower limit that accommodates 3’ss - 5’ss bridging interactions; however, very short exons are sometimes efficiently spliced (Berget, 1995).

4.1.4 USSE activation regulates stability and localization of 48K and 65K mRNAs (I)

Activation of the upstream 3’ss by the USSE in 48K mRNA is predicted to lead to decay by the NMD pathway, as the inclusion of the 8-nt exon leads to a frameshift and eventual inclusion of a PTC with both USSE-induced mRNA isoforms. In the case of the exon 4i-containing transcript this was experimentally verified using the translation inhibitor cycloheximide (CHX; see Figure 5A in I). As NMD is activated during the
pioneer round of translation (reviewed by Maquat et al., 2010), inhibition of protein synthesis is expected to stabilize any NMD target. Consistently, both CHX treatment and expression of a dominant negative form of Upf1 (Figure 5B in I; Lykke-Andersen et al., 2000), an essential NMD factor, led to a substantial stabilization of exon 4i containing mRNAs, indicating that the 48K mRNA is indeed targeted by the NMD pathway.

In contrast, with the 65K mRNA, CHX treatment did not lead to stabilization of the long isoform (Figure 5F in I), consistent with the location of the alternatively spliced 3’ss within ca. 50 nt of the actual translation termination codon. Instead, the 65K short 3’UTR, when cloned to a luciferase reporter and assayed for luciferase activity, produced 2-3 fold higher luciferase signals than the long 3’UTR (Figure 5G in I). Furthermore, an equimolar cotransfection of the long and short 3’UTR reporters led to the expression of the short isoform in both nucleus and cytoplasm, as assayed by fractionation (Figure 5G in I), whereas the long 3’UTR reporter was retained mainly in the nuclear fraction. The endogenous 65K mRNAs localized in the same way (Figure 5H in I). Overall, the luciferase activity and both the reporter and endogenous mRNA localization support two nonconflicting models, whereby the 65K long 3’UTR mRNA would either be retained in the nucleus, and/or the cytoplasmic long isoform would be destabilized. Interestingly, in a related study, it was found that the longer 4i-containing 48K transcript was not stabilized with CHX treatment (Turunen et al., 2013b). Rather, it was speculated that the presence of the USSE element in the mRNA (as opposed to removal of the USSE sequence via splicing, as in the exon 4i-only isoform) may have led to nuclear retention of this transcript.

What are the implications and the extent of the USSE-mediated regulation of the 48K and 65K mRNAs? Remarkably, the qPCR quantification above indicated (4.1.3) that up to 75% of the transcripts were directed to the decay pathway. This suggests that the level of both the transcripts and also their protein products are actively downregulated, at least in the cell lines used in this study. It is not known whether this autoregulatory feedback loop can be regulated differentially in other conditions or cell types. The hnRNP H/F proteins identified as critical negative factors regulating the activation signals by the USSE element may represent an example of an additional regulatory layer (Honoré et al., 2004; Turunen et al., 2013b).

4.1.5 The function of the USSE and evolutionary implications

Splice site sequences, including U12-type splice site-like motifs, have previously been observed by others to function as splicing regulatory elements (see section 1.4.6; Hwang and Cohen, 1997; McNally and McNally, 1999; Lewandowska et al., 2004; Wang et al., 2004; Yeo et al., 2007; Dhir et al., 2010). In contrast to the present work they have been mostly shown to act as negative regulators, typically suppressing the
recognition of pseudo splice sites in in vitro splicing reactions, or they have been identified in SRE screens as inhibitory elements. The mode of action in such cases has been proposed to be a direct blocking of productive interactions by steric hindrance, or by destabilization of the binding of other splicing factors. Some of these cases differ from the USSE system in having binding sites very close to the elements that are being inhibited (Hwang and Cohen, 1997; Lewandowska et al., 2004; Dhir et al., 2010). In contrast, with the USSE, it is likely that the distance between the element and the target site can accommodate both USSE-binding di-snRNPs and proteins and snRNPs recognizing the actual splice sites.

Given that a large group of spliceosomal proteins, and also proteins functioning more generally in RNA processing, are regulated by a feedback mechanism (Saltzman et al., 2008), it is likely that similar negative feedback loops involving spliceosomal complexes will be identified. Indeed, a very recent work described the feedback regulation of a major spliceosome component (the U1-70K protein) that operates by very similar principles as the USSE system (Rösel-Hillgärtner et al., 2013). In that case, activation of an alternative 3′ss within the U1-70K mRNA is regulated by U1 snRNP binding to unproductive U2-type 5′ splice site sequences located downstream. High levels of functional U1 snRNP lead to U1 binding to the unproductive 5′ splice sites and to the activation of the upstream 3′ss, presumably using exon definition interactions similar to the USSE-mediated alternative splicing. Alternative 3′ss activation introduces a PTC in the U1-70K mRNA and leads to decay via NMD. The recognition of the unproductive 5′ splice sites is regulated through the cellular levels of U1-C protein (a functional analog of the U11-48K protein). Thus, like the USSE, the system controls the cellular U1-70K/U1-C homeostasis needed for correct U1 snRNP assembly. The identification of analogous feedback systems in both spliceosomes, that regulate the cellular homeostatic of the central snRNP protein components, suggests that such stringent control is necessary for the correct functioning of both spliceosomes.

With the U12-dependent spliceosome the need for correct 48K protein levels is obvious, as it has been shown that a significant reduction in the cellular levels of 48K leads not only to failure to splice U12-type introns and intron retention, but also to the activation of nearby cryptic U2-type splice sites (Turunen et al., 2008). This can severely compromise the expression of genes containing U12-type introns by introducing insertions or deletions that can compromise the function of the proteins or lead to NMD-mediated decay of the incorrectly spliced mRNAs. Similarly, mutations of the human 65K that impair its normal function in bridging the U11 and U12 snRNPs to form the U11/U12 intron recognition complex lead to a developmental disorder, presumably as a consequence of activation of cryptic U2-type splice sites and other missplicing events (Argente et al., 2014). Therefore, as the correct level of these two
proteins is necessary for the recognition of U12-type introns, it is conceivable that the levels of both proteins are being regulated through similar feedback systems.

One way to explain the existence of the intricate feedback regulation of U12-dependent spliceosome components is the need to maintain the observed slow processing of U12-type introns, which in turn provides the rate-limiting regulation for the genes with U12-type introns (III; Patel et al., 2002; Patel and Steitz, 2003). However, it is unclear whether intron recognition would be the appropriate step to regulate splicing rate, because, as discussed above, failure to recognize a U12-type intron can activate nearby cryptic U2-type splice sites. Therefore, if the slow splicing of U12-type introns is under regulatory control, then it is conceivable that the rate-limiting regulatory step should be after the intron recognition step (see 4.2.3). Interestingly, in vitro splicing data indicate that while the intron recognition (A-complex assembly) is relatively fast with the U12-type introns, the subsequent activation step seem to be rate-limiting (Frilander and Steitz, 1999). Consistently, recent work has suggested that U6atac snRNA levels can be regulated via signaling pathways to influence the efficiency of U12-type intron removal (Younis et al., 2013), suggesting that tri-snRNP levels regulated through U6atac availability could provide the means to regulate the slow splicing of U12-type introns. Furthermore, as the 65K protein has been shown to bind not only the U12 snRNA (Benecke et al., 2005), but also U6atac snRNA (Benecke, 2004), it is also possible that the feedback regulation of the 65K levels are linked to the efficiency of tri-snRNP recruitment to the spliceosome, therefore providing the means to regulate the speed of U12-type intron removal.

While the USSE sequence itself can be found from humans to plants, suggesting an ancient evolutionary origin, its presence in variable locations (either intronic or 3’UTR) suggests that it may have arisen on several occasions during eukaryotic evolution. 65K USSE appears to be a more recent addition as it is present in all vertebrates but not in any other species of the present study (Figure 7 in I). Since it brings also 65K under the influence of the feedback loop, the function of USSE in the 65K may be to provide an even more stringent level of regulation, suggesting that for vertebrates, the regulation of minor spliceosome activity may be of crucial importance. This is also supported by the strict conservation of the USSE distance parameters in vertebrates (II). On the other hand, the absence of the USSE in either 48K or 65K in some lineages suggests that it has been lost from either gene on several occasions, possibly concomitant with the reduction in the number of U12 introns (e.g. in Diptera) or in combination with the loss of rate-limiting regulatory function of U12-type introns.
4.2 U12-DEPENDENT TRANSCRIPTS ARE SPLICED LESS EFFICIENTLY AND TARGETED TO NUCLEAR DEGRADATION (III)

4.2.1 Exosome inactivation leads to stabilization of intron-retaining transcripts

It has been long known that U12-type introns are spliced less efficiently than the major type introns (Tarn and Steitz, 1996b; Frilander and Steitz, 1999; Patel et al., 2002). However, it was not known whether the slow splicing of U12-type introns is a global phenomenon. Similarly, the subsequent fate of the transcripts containing unspliced U12-type introns has not been investigated. We hypothesized that U12-type intron retention would lead to increased dwell times in the nucleus, which would expose the unspliced transcripts to nuclear surveillance pathways that target nuclear RNAs based on their processing efficiency (a process called kinetic proofreading; see section 1.6.5).

To test the potential interconnections between the minor spliceosome and nuclear quality control pathways, we disabled various RNA degradation pathways by knocking down different decay factors that are known to function in the nuclear quality control, including RNA exosome components and the 5′ to 3′ decay factors. The expectation was that relevant pathways would show a stabilization of unspliced U12-type introns signals upon the knockdown. Initially, we assayed intron retention in three genes, VPS16, MAPK12, and RCD8, which each contain either 1 or 2 U12-type introns (III, Figure 1A). The U12-type introns in these genes are relatively short, which permits simultaneous amplification of unspliced pre-mRNAs and spliced exon-exon junctions using primers in the flanking exons. We found that knockdown of the exosome core subunit RRP41 produced the most robust and reproducible stabilization of unspliced U12-type transcripts, whereas the knockdown of exonuclease subunits DIS3 and RRP6 produced milder effects (Figures 1 and S1 in III). Other exosome subunits tested (data not shown) or 5′ to 3′ decay factors XRN2 and DCP2 had no effect on transcript stability (Figures 1 and S1 in III). The stabilization of U12-type introns caused by RRP41 knockdown was partially rescued by the overexpression of RRP41 with silent point mutations in the siRNA target site (Figure 1D,E in III).

4.2.2 U12-type intron retention on global scale

Since the inefficient splicing of U12-type introns has previously been reported only in a handful of genes, we decided to investigate this on a transcriptome-wide scale. Control and knockdown cells were fractionated and the resulting RNA pools were sequenced. First we compared the retention levels of U12-type introns to their neighboring U2-type introns in the control knockdown, and found that they were on average two times higher (Figures 2C and 3G in III). This reproduces, on a transcriptome-wide scale, the previously reported small-scale findings of increased U12-type intron retention (Patel
et al., 2002; Pessa et al., 2006). After knockdown of exosome subunits, 119 introns were differentially retained with statistical significance in either knockdown. 80% of these were identified in the RRP41 knockdown, and a majority of these were stabilized, as expected (Figure 3A,B in III).

The fold change values were larger in the RRP41 knockdown than for DIS3; in fact, the fold change distribution of the whole set of U12-type intron-containing genes shifted to stabilization in RRP41 knockdown (Figure 3E in III). DIS3 knockdown, in addition to intron stabilization, showed considerably more intron destabilization than RRP41 knockdown (Figure 3C,F in III). Finally, the overlap between stabilized introns in the two knockdowns was larger and only marginally significant (Figure 3C in III). The reasons for the partial discrepancy between the DIS3 and RRP41 knockdowns are not known, but it is possible that DIS3 may be compensated by other exoRNases that associate with the exosome core, even to such an extent that the expression of alternative exoRNases is stimulated (Tomecki et al., 2010). This could lead to the observed destabilization of the introns not shared with the knockdown of RRP41.

We hypothesized that if the inefficiency of splicing makes U12-type introns susceptible to nuclear degradation, then the more slowly spliced the intron is, the more likely it is degraded. Thus, we attempted to identify a signature of such inefficiently spliced introns by analyzing intron characteristics such as splice site score, subtype (AT-AC vs. GT-AG), but found no correlation between these characteristics. Also, we reasoned that maybe larger introns and downstream introns are more likely to be retained and therefore more likely to be targeted, but again there was no correlation. Neither could we find enrichment or absence of sequence motifs that would suggest regulation by splicing factors.

To control for any effects that the exosome knockdown might directly have on the efficiency of U12-type splicing, we performed several controls. First, we probed on northern blots that the knockdowns did not affect the levels and sizes of minor spliceosomal snRNAs (Figure S2A in III). Second, we confirmed from the RNAseq data that the expression of minor spliceosome-specific proteins is not changed in the cytoplasmic fraction (III, Figure S2D). Third, we counted U12-type exon-exon junction reads in the nuclear fraction, which represent the efficiency of splicing for nascent RNAs (III, Figure S2B). In fact, we found that with both of the subunit knockdowns, the number of junction reads increased. This is consistent with the kinetic proofreading hypothesis, as the knockdown of exoRNase activity allows the splicing of more U12-type intron-containing transcripts to reach completion. However, there were very few genes containing U12-type introns whose cytoplasmic expression changed after the exosome knockdown.
Finally, we did not find evidence of cryptic splicing of U12-type introns after exosome knockdown. It has been shown that when the U12-type recognition is impaired, nearby cryptic U2-type splice sites are often activated (Turunen et al., 2008; Argente et al., 2014). As we did not find any evidence of increased levels of cryptic splicing near U12-type introns after the exosome knockdown, we conclude that U12-type introns have been correctly recognized by the U11/U12 di-snRNP in the transcripts that are targeted for decay by the nuclear surveillance pathways. Instead, a subsequent slow step in spliceosome assembly or catalysis, as discussed earlier (section 4.1.5), leads to a slow splicing and nuclear decay.

### 4.2.3 U12- and U2-type introns show differential decay rates

To confirm the slower splicing kinetics and the effect of the exosome knockdowns we investigated the decay rates of a subset of pre-mRNAs in control and exosome knockdown cells. The cells were treated with DRB (5,6-dichloro-1-6-D-ribofuranosylbenzimidazole), an inhibitor of CDK7 phosphorylation and of Pol II progression from initiation to elongation, and the decay of three U12-type introns was followed by RT-qPCR (III, Figure 4C). The genes, STX10, MORC4, and CHD1L, were chosen from the group of stabilized introns so that the total length of pre-mRNA would not be very large (i.e. they would complete transcription within 15 min following the introduction of DRB) and the locations of the U12-type introns would vary (i.e. located at the 5′, middle and 3′ part of the transcript, respectively). One U2-type intron from each gene was analyzed as a control, and was chosen to match the length of the U12-type intron as well as possible.

We found that the half-lives of the U2-type introns in the DRB experiment were surprisingly long, 30-60 min, and showed little change upon exosome knockdown. In contrast, the half-lives of U12-type introns, approximately 90-150 min in control cells, were further increased by 1.5 to 2-fold upon exosome knockdown. Additionally, the RT-qPCR data also showed the increased levels of transcripts containing U12-type introns for each gene tested. The interpretation of these experiments is that while the decay of U2-type pre-mRNA signal results predominantly from the cotranscriptional splicing as the exosome knockdown did not have a strong effect on the decay rate, a significant fraction of transcripts containing U12-type introns are targeted by the nuclear exosome. Considering that U2-type introns are removed cotranscriptionally in the 5′ to 3′ order they emerge from Pol II (Pandya-Jones and Black, 2009), our data suggest that U12-type introns may deviate from this processing order and are removed last.

Previous work on the kinetics of pre-mRNA splicing has shown that U12-type introns can be spliced cotranscriptionally, as the initial RT-qPCR signals for successful splicing of U12-type introns after the start of transcription appear before Pol II has finished
transcription (Singh and Padgett, 2009). This is consistent with the detection of the snRNA components of the minor spliceosome in the chromatin fraction in cell fractionation experiments (Tilgner et al., 2012). On the other hand, our kinetic data suggest that at least with a subset of genes a significant fraction of the U12-type splicing occurs post-transcriptionally. Interestingly, after exosome knockdown, the U12-type intron decay kinetics show a relatively long lag period following the addition of DRB, during which no decay takes place. This observation suggests that splicing does not occur at these introns for a considerable time period after the splice sites emerge from Pol II.

Taken together, the transcriptome-wide study confirms that the majority of U12-type introns, over 70% (III, Figure 3G), are retained to a larger extent than their U2-type neighbors. While the highly conserved regulatory loop may not explain the inefficiency of U12-dependent splicing, pre-mRNAs with retained U12-type introns are stabilized following exosome inactivation. The strong upregulation of the pre-mRNA signals together with considerably delayed decay kinetics suggest that, under steady-state conditions, U12-type splicing is indeed rate-limiting. The slow removal causes a substantial fraction of the nascent U12-type intron containing pre-mRNAs to be degraded before reaching cytoplasm, and therefore, is highly important for maintaining the correct expression level of genes with U12-type introns.
5 CONCLUDING REMARKS

Since the discovery of the U12-dependent spliceosome, the maintenance of two parallel spliceosomes in many eukaryotic lineages has remained an enigma. The minor spliceosome is clearly essential for the development of the organisms that retain it, yet U12-type introns are easily lost via mutagenesis and conversion to U2-type introns, and some lineages have lost the U12-dependent spliceosomal components altogether. This work has shown that not only are the introns and spliceosomal components conserved in plant and animal species, but also the regulatory elements that control the abundance and activity of critical subunits show remarkably strict conservation. The function of the USSE autoregulatory loop is to maintain the intron recognition components of the minor splicing pathway at optimal levels. Deviation from this homeostasis would lead to intron retention or activation of U2-type cryptic splice sites in the vicinity of the unrecognized U12-type introns, and potentially harmful missplicing events.

The hypothesis of rate-limiting U12-type introns in gene expression is well established, but it has not been established whether all U12-type introns are inefficiently removed. The transcriptome-wide results in this work show that the majority of U12-type introns indeed are retained when compared to U2-type introns. The data presented here on the kinetics of U12-type intron decay indicate that there is a considerable delay in the processing of these introns. The lack of cryptic splice site usage in the RNAseq data suggests, however, that the processing is stalled after splice site recognition. Therefore, the rate-limiting step of U12-type splicing pathway may not be intron recognition but instead further downstream. Indeed, multiple arguments point to the availability of U6atac contributing to the rate limiting step of the pathway.

Finally, the rate-limiting hypothesis implies that regulating the activity of U12-dependent spliceosome is important for maintaining correct removal rates, and it can be further speculated that its activity may be different under some physiological conditions. Until recently there has been no evidence on how the U12-dependent spliceosome activity could be regulated. The current work on the feedback loop regulation, together with other recent publications, has opened the field for the study of minor spliceosome regulation.
6 ACKNOWLEDGEMENTS

This study was carried out mainly at the Institute of Biotechnology at University of Helsinki, while parts of the work were performed at the facilities of the Radboud Institute for Molecular Life Sciences, Radboud University Nijmegen. I would like to thank both of the Institutes for providing excellent research facilities and infrastructure. The Helsinki Graduate Program in Biotechnology and Molecular Biology is gratefully acknowledged for supporting my doctoral studies financially. My work has also been supported by the Orion-Farmos foundation and the Magnus Ehrnrooth foundation, whose generosity has allowed me to travel abroad for collaboration and meetings.

I am grateful to my supervisor Mikko Frilander for guidance and support during my years in the lab. You have created a stimulating working atmosphere with a constant influx of new ideas while at the same time fostering independence and thinking for oneself. I am especially grateful for your openmindedness, flexibility and help with setting up the exosome collaboration. I also thank my colleagues in the RNA splicing lab, Janne, Heli, Jens, Jouni, Bhupendra, Ali, Visa, Prosanta, Antto, Lilli and Annukka for friendship and discussions, both science-related and otherwise. Janne, you have set the benchmark in the lab very high with your vast knowledge, beautiful experiments and attention to detail. Lilli deserves special thanks for excellent technical assistance, for knowing just how I want my cloning, for managing the lab smoothly, and for creating a warm working atmosphere. You really are the heart and soul of the lab.

I am delighted to have professor Michelle Hastings as my opponent and look forward to our discussions. I thank professor Kari Keinänen for serving as my custos. Dr. Maria Vartiainen and Dr. Matjaz Barboric provided valuable discussions and feedback while reviewing my thesis and are gratefully acknowledged. Dr. Vartiainen also served as my thesis group member along with Dr. Osamu Shimmi, to whom both I am thankful for following my work during the years. Your enthusiasm and encouragement was highly appreciated.

I am indebted to professor Ger Pruijn who collaborated with me on the exosome project and allowed me to spend time in his lab to perform critical experiments. The members of the Biomolecular Chemistry lab are acknowledged for welcoming me so warmly into your lively lunch and coffee breaks, borrels, brain storming dinners, pizzas, movie nights etc. Sandy, thanks for your support, friendship, and humor, and for being my RNA roomie on many conference trips. Raymond, your knowledge in all exosome-related things was invaluable. Geurt and Sander, thanks for your insightful advice on my project and your friendship. Tamara, you are absolutely awesome, and my days are a lot duller without you. Manuela, thanks for your friendly company in the lab and on
many excursions as far as to Utrecht and Heidelberg. Carla S., thank you for all the help in the lab, I learned all my protein purification and Western tricks from you. Wilma, thanks for your wise words and practical knowledge in all exosome-related experiments. Lettie, Wilbert, Siebe, Sanne, Lonneke, Marina, Helma, Joyce, Annemarie, Judith, Carla O., Chantal, and Remon, you created an excellent and fun working environment. Thank you Els for taking care of all practical issues involved with my stay. Marieke, many thanks for your friendship and sharing the ups and downs of motherhood in the Netherlands. My days with baby Topi would have been a lot more monotonous without the company of you and Cas and our numerous outings to Blonde Pater and elsewhere. Minja, I am so glad that in the end I found the only other Finn in the building. Your sparkling wit and Nordic frame of mind made excellent lunch and party company that alleviated any homesickness that might have lingered.

I have been lucky to collaborate with brilliant scientists during my work. Cindy Will and professor Reinhard Lührmann from the Max Planck Institute for Biophysical Chemistry in Göttingen as well as Janne Ravantti from the Department of Biosciences at University of Helsinki are acknowledged for the collaboration for my first thesis publication. Dario Greco, currently at the Finnish Institute of Occupational Health, is acknowledged for help on the bioinformatical aspects of the transcriptome-wide data.

I thank the Helsinki Graduate Program in Biotechnology and Molecular Biology (later Doctoral School for Integrative Life Sciences) for organizing interesting courses and events and especially coordinator Erkki Raulo for support and flexibility with my somewhat unique arrangement. I am grateful to Anita Tienhaara and Eeva Sievi from the graduate schools, to Jonna Katajisto from the Faculty Office and Satu Sankkila from BI for help with practical matters concerning my thesis and graduation.

The past and present lab neighbours in the fifth floor are acknowledged for a supportive and friendly atmosphere. Harri Savilahti, Tapio Heino, Oscar Puig, Osamu Shimmi, and Tomi Mäkelä and their labs are warmly remembered. Special thanks to the lunch crew including Bianca, Emma, Heini, Kaisa, Niina, Saana, Saara, Ying and others, for guaranteed laughs every working day and the sharing of reagents and equipment. I really look forward to working with all of you in the future. Juha Laurén, at one time of Neuroscience Center, you have been an important teacher and a role model. Among many things, you taught me how to culture cells, which is the foundation of the career of any molecular biologist, as well as set the example on how to run away abroad with graduate school money.

My friends outside the world of biosciences are warmly thanked for diverting my thoughts regularly. My dear friend Tove, thank you for being there for me as long as I can remember. During our busy lives we may have repeatedly lived far apart but we
have always picked up from right where we left. Your positive outlook on life and your perseverance cheered me on many occasions. The Sunday lunchers, Mirkka, Jaakko, Elina, Mikko and the kids are thanked for culinary experiences and lively company. The book clubbers, including Jens, Heidi, Fergal, Tiina, Natalia, Michele, Janneke, Sylvie, Suvi, Jonathon and Elina P. are thanked for fun and brilliant discussions and refreshing ideas.

I thank my in-laws, Arja and Risto, for welcoming me into your family so warmly. Mum and Dad, you have brought me up with an ever inquisitive mind, independence and a world wide open with possibilities. Your love and trust have supported my career and you have never even blinked at any of my choices in life.

Mauri, thank you for your endless love and support. I am fortunate to have a fellow scientist at home who understands the occasional pipetting and writing marathons and conference travel as well as challenges my thoughts and encourages me on. Malla and Topi, thank you for pulling me away from the lab at the end of regular working hours and reminding of what is important in life. You mean the world to me.

Helsinki, August 2014

Elina
7 REFERENCES


pre-mRNA and nucleates the assembly of 40S heterogeneous nuclear ribonucleoprotein particles. Mol Cell Biol 14, 518-533.


McCracken, S., Fong, N., Rosonina, E., Yankulov, K., Brothers, G., Siderovski, D., Hessel, A., Foster, S., Shuman, S., and Bentley, D.L. (1997a). 5'-Capping enzymes are
targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. Genes Dev 11, 3306-3318.


References


References


