UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA VEGETAL

Splicing kinetics revealed by nascent transcript sequencing

Kenny Patrick Figueira Rebelo

Mestrado em Bioinformática e Biologia Computacional
Especialização em Biologia Computacional

Dissertação orientada por:
Professora Doutora Maria Carmo-Fonseca
Professora Doutora Lisete Sousa

2017

" (...)

    She wraps man in darkness, and makes him for ever long for light. She creates him dependent upon the earth, dull and heavy; and yet is always shaking him until he attempts to soar above it.

<div align="right">(...)"</div>

<div align="right">**Johann Wolfgang von Goethe**</div>

Acknowledgements

This was the first time I was part of a scientific investigation team. A cog in the big knowledge producing machine, science. Which I hope to contribute, no matter how little, for a more aware future.

I want to thank my advisor Carmo-Fonseca for accepting me into her team at Instituto de Medicina Molecular de Lisboa. Since that first day she always was a brilliant professor to me with a contagious enthusiasm towards investigation. To Ana Rita Grosso for her endless support and understanding despite how easy and simple the task at hand was. And Lisete de Sousa for the help in co-advising this dissertation.

Tomás, for all the crystal clear guidance.

To Cláudio Vieira and Joana Tavares for all the time spent helping me deal with the bumps I found along the way. Lots of small things, all the time. And every time you guys went out of your way to lend a hand.

To all my friends that make my days whole. And of course to my caring parents for all the strength they provide me.

Thank you all from the bottom of my heart for your support.

Resumo

A transcrição é o primeiro passo da expressão génica, no qual um segmento particular de informação presente no DNA é copiado para RNA mensageiro (mRNA), que por sua vez irá ser traduzido tendo como resultado final o conjunto de proteínas que suportam a vida. A biogénese do RNA mensageiro (mRNA) é um processo constituído por múltiplas etapas. Imediatamente a seguir à transcrição o RNA é modificado na extremidade 5' pela adição de uma 7-methylguanosina. Segue-se o *splicing*, que resulta na remoção dos intrões. Finalmente o RNA é clivado na extremidade 3' e modificado pela adição de uma cauda poli(A). Uma vez que a maior parte destes processos ocorre enquanto a molécula de RNA nascente ainda está acoplada à RNA polimerase II (Pol II), considera-se que a biogénese do mRNA é maioritariamente um processo co-transcricional.

A associação entre Pol II e processamento de moléculas precursoras de mRNA (pré-mRNAs) é mediada pela maior subunidade da Pol II que se caracteriza por possuir, nos organismos eucariotas, um domínio carboxilo terminal (CTD) altamente longo e flexível. Este serve como plataforma para ancoragem de várias proteínas que regulam a transcrição e o processamento do pré-mRNA. O CTD da Pol II nas células humanas inclui 52 repetições da sequência proteica consenso YSPTSPS que é conservada entre leveduras e mamíferos. Para além de 21 repetições com a sequência consenso, as células humanas apresentam ainda 31 repetições de sequências mais degeneradas maioritariamente na última metade (desde a repetição 26 até à 52). Durante o ciclo da transcrição, composto por iniciação, elongação e terminação, o CTD é reversivelmente modificado por várias reações químicas, nomeadamente por fosforilações. Consoante o tipo de modificação, o CTD manifesta afinidade para diferentes proteínas. Assim, consoante a fase da transcrição, o CTD recruta proteínas responsáveis por etapas específicas do processamento do pré-mRNA. Antes da Pol II iniciar a transcrição o CTD encontra-se sobretudo desfosforilado. Durante a fase de iniciação da transcrição o CTD está maioritariamente fosforilado na serina 5 (S5P), enquanto que na fase de elongação o CTD vai gradualmente ficando mais fosforilado na serina 2 (S2P) perdendo ao mesmo ritmo o nível de fosforilação em S5P até à fase de terminação onde o CTD estará maioritariamente fosforilado na S2P. A transcrição termina quando a Pol II se dissocia da cadeia de DNA, ficando o CTD desfosforilado e pronto para começar um novo ciclo.

O *splicing* de pré-mRNAs consiste num processo químico constituído por duas fases em que é primeiro gerado um grupo hidroxilo 3' livre (3'-OH) no exão a montante e se cria uma ligação fosfodiéster 2'-5' na adenosina no ponto de ramificação ("*branch point*") responsável pela estrutura característica em forma de laço adotada pelo intrão. Na segunda fase, o grupo 3'-OH do exão a montante realiza um ataque nucleofílico no local de *splicing* (SS) a 3' de forma a ligar os exões e excisar a estrutura de laço formada pelo intrão. Estas reações ocorrem dentro do spliceossoma cataliticamente ativo, mas antes de chegar a este ponto é necessário que o SS a 3' esteja exposto, isto é, deixe de estar protegido pela estrutura da Pol II. De facto, as sequências de nucleótidos presentes nos locais local de *splicing* a 5' e a 3' precisam de estar acessíveis para serem reconhecidas por componentes do spliceossoma.

A maior parte do *splicing* ocorre co-transcricionalmente, isto é, enquanto a molécula de pré-mRNA ainda está acoplada à Pol II durante a transcrição da cadeia de DNA. Ao longo dos últimos anos foram realizados inúmeros estudos que revelam uma íntima relação entre o *splicing* e da transcrição, com importantes repercussões funcionais. Por exemplo, já foi descrito que a velocidade de elongação da Pol II influencia a capacidade de reconhecimento de locais de *splicing*, pelo que a existência de pausas da Pol II determina a escolha de padrões distintos de *splicing* alternativo. Esta associação entre a transcrição e a

biogénese do mRNA é determinante para a regulação da expressão génica. No entanto, sabe-se pouco sobre os mecanismos que ligam as duas maquinarias moleculares.

O objetivo deste trabalho foi realizar um estudo à escala genómica do acoplamento entre a transcrição e o *splicing* utilizando uma estratégia de sequenciação de nova geração denominada "*Native Elongation Transcript sequencing in mammalian cells*" (mNET-seq). Esta foi a abordagem escolhida porque era aquela que nos permite estudar perfis de transcrição nascente com resolução de nucleótido a nucleótido ao nível de todo o genoma, onde podemos então procurar padrões específicos de pausa da Pol II. O meu trabalho conducente a esta dissertação de mestrado consistiu na análise bioinformática de dados gerados, através da plataforma Illumina Hiseq 2500, no laboratório do Professor Proudfoot na Universidade de Oxford, Reino Unido, tendo sido o planeamento experimental conduzido em colaboração com a Professora Carmo-Fonseca no Instituto de Medicina Molecular da Faculdade de Medicina da Universidade de Lisboa.

Os resultados obtidos revelaram a presença de componentes do spliceossoma, nomeadamente os pequenos RNAs nucleares (snRNA) U1, U2, U4 e U5, num complexo associado especificamente à Pol II com o seu domínio carboxilo terminal contendo resíduos de serina 5 fosforilados (S5P Pol II). Encontramos também a nível global produtos intermediários da reação de *splicing* em associação com S5P Pol II. A capacidade da técnica de mNET-seq de mapear intermediários de *splicing* permitiu-nos uma análise global do *splicing* co-transcricional in vivo. O cálculo da percentagem de índice de *splicing* (PSI) mostrou que praticamente todos os exões que apresentam um intermediário de *splicing* associado a Pol II estão completamente incluídos no mRNA. Observamos níveis mais elevados de densidade da Pol II nos exões com *splicing* co-transcricional, em particular para o conjunto de dados de S5P CTD. Ademais identificamos moléculas de RNA já desprovidas de intrões associadas preferencialmente a S5P Pol II. Nestes casos observamos uma acumulação de transcritos já processados sugestivos de pausas da Pol II relacionadas com o processo de *splicing*. A nossa análise permitiu-nos ainda identificar a posição da Pol II no momento em que ocorreu *splicing*. Os resultados mostram uma tendência para o *splicing* ocorrer quando a Pol II transcreveu apenas cerca de 20 nucleótidos após o local de *splicing* a 3'. No entanto, ainda que raramente, foram detetados eventos de *splicing* a ocorrer quando a Pol II já se encontrava a transcrever o intrão a jusante. Em resumo, os resultados obtidos revelam que as reações de *splicing* podem ocorrer imediatamente após a transcrição pela Pol II e sugerem que o *splicing* impõe uma travagem à velocidade da Pol II.

Os resultados deste trabalho permitiram ainda identificar algumas limitações impostas pela técnica de mNET-seq, pelo que será necessário desenvolver aperfeiçoamentos e/ou técnicas adicionais para uma melhor caracterização global dos mecanismos de *splicing* co-transcricional no futuro. A principal limitação relaciona-se com o reduzido tamanho dos fragmentos de RNA que resistem à digestão pela *Micrococcal nuclease* (Mnase). Neste trabalho analisamos fragmentos de RNA com um tamanho entre 60 a 160 nucleótidos, o que não nos permite localizar a posição da Pol II no momento em que o *splicing* lento acontece nas células humanas. É previsível que o contínuo desenvolvimento de novas plataformas de sequenciação em larga escala vão também permitir uma análise mais precisa dos eventos de processamento do RNA que decorrem durante a transcrição. Desenvolvimentos futuros que permitam aumentar os níveis de resolução e sensibilidade das análises são absolutamente essenciais para uma correta interpretação da biologia subjacente. A par do desenvolvimento tecnológico, importa também investigar se as observações aqui reportadas em células humanas (células HeLa) são conservadas do ponto de vista evolutivo, nomeadamente noutros mamíferos e em organismos menos complexos. Em particular,

seria importante determinar em que estadio ao longo da evolução ocorreu a separação entre *splicing* rápido e lento, e qual das duas formas (rápida e lenta) terá surgido primeiro.

Palavras-chave: mNET-seq, Transcrição, Splicing, Spliceossoma, CTD

Abstract

The making of messenger RNA is a multi-step process that starts with transcription of a protein-coding DNA template by RNA polymerase II (Pol II) followed by capping the 5' end of the nascent transcript with a 7-methylguanosine, excision of introns, and formation of a 3' end by cleavage and addition of a poly(A) tail. Splicing of precursor mRNAs (pre-mRNAs) is a two-step chemical process that first generates a free 3' hydroxyl group on the 5' exon and creates a 2'-5' phosphodiester linkage at the adenosine branch site that causes the intron to adopt a characteristic lariat structure. In the second step, the 3'-OH of the 5' exon attacks the 3' splice site (SS) to form ligated exons and excised intron lariat. Most of these processes take place co-transcriptionally, i.e., while the nascent RNA molecule is still attached to Pol II on the DNA template. Co-transcriptionality of mRNA biogenesis has important regulatory implications for gene expression. Although several lines of evidence suggest that splicing of pre-mRNA is intimately coupled to Pol II transcription, mechanistic links between the transcription and splicing machineries are little-understood. The aim of this work was to study at genome-wide level the coupling between transcription and splicing using a recently developed next generation sequencing strategy termed Native Elongation Transcript sequencing in mammalian cells (mNET-seq) through bioinformatics analysis of datasets generated in the laboratory of Professor Proudfoot at the University of Oxford in collaboration with Professor Carmo-Fonseca at iMM Lisboa. We found components of the spliceosome, namely U1, U2, U4 and U5 snRNAs, specifically associated with serine 5-phosphorylated (S5P) CTD Pol II. Intermediates of many splicing events as well as a subset of spliced transcripts were also detected in S5P Pol II complexes. Finally, specific patterns of splicing-related polymerase pausing were identified. Our results reveal that splicing can be completed immediately after exposure of the 3'-splice site and suggest that splicing induces a barrier to transcription elongation.

Keywords: mNET-seq, Transcription, Splicing, Spliceosome, CTD

Table of Contents

Table of Figures

List of Abbreviations

Bp base pair(s)
mRNA messenger RNA
DNA desoxirribonucleic acid
Pol II RNA polymerase II
RNA ribonucleic acid
OH hydroxy group
SS splice site
mNET-seq *Native Elongation Transcript sequencing in mammalian cells*
snRNA small nuclear RNA
S5P serine 5 phosphorilated
S2P serine 2 phosphorilated
CTD carboxy terminal domain
snRNP small nuclear ribonucleo proteins
BS branch site
ChIP-seq chromatin Immunoprecipitation sequencing
TSS transcription start site
IM immediate *splicing*
MNase micrococcal nuclease
RPM reads per million
S7P serine 7 phosphorilated
T4P threonine 4 phosphorilated
Y1P tyrosine 1 phosphorilated
DMSO dimethyl sulfoxide
Pla-B pladienolide B
NpRNA nucleoplasmic mRNA
w with
wo without
**Ex** exon
chr chromosome

1. INTRODUCTION

The making of messenger RNA is a multi-step process that starts with transcription of a protein-coding DNA template by RNA polymerase II (Pol II) followed by capping the 5' end of the nascent transcript with a 7-methylguanosine, excision of introns, and formation of a 3' end by cleavage and addition of a poly(A) tail. Most of these processes take place co-transcriptionally, i.e., while the nascent RNA molecule is still attached to Pol II on the DNA template. Co-transcriptionality of mRNA biogenesis has important regulatory implications for the ultimate synthesis of proteins. For example, promoter elements can affect the decision to include an alternatively spliced exon (Cramer & Pesce 1997), and co-transcriptional processing can influence the rate of mRNA transcription, export from the nucleus and stability in the cytoplasm (reviewed in Moore & Proudfoot, 2009 ; Bentley, 2014).

Synthesis of mRNA relies specifically on Pol II, and RNAs experimentally transcribed by Pol I or Pol III fail to maturate properly (Sisodia & Sollner 1987 ; Smale & Tjian, 1985). A distinguishing feature of Pol II is the carboxy-terminus of its large subunit (Rpb1) known as the CTD, which consists of tandemly repeated blocks of seven amino acids (**Figure 1.1**). In the yeast *S. cerevisiae* the CTD is composed of 26 repeats with the consensus sequence tyrosine-serine-proline-threonine-serine-proline-serine, whereas human CTD is composed of 52 repeats with significant deviations from the consensus (Eick & Geyer, 2013). The CTD is thought to act by recruiting processing factors that become localized at the right place to act on the nascent transcript. During the transcription cycle of initiation, elongation and termination, the CTD is modified by reversible phosphorylation at multiple positions, and it has been suggested that these modifications change its binding properties (Eick & Geyer, 2013). Thus, different factors may interact with Pol II CTD during the transcription cycle, thereby facilitating the assembly of multisubunit complexes such as the capping and the 3' end processing machinery (Bentley, 2014). Whether Pol II CTD also participates in spliceosome assembly is unclear.

The spliceosome is a multimegadalton ribonucleoprotein complex that catalyzes splicing of precursors to mRNAs. Two distinct types of spliceosomes have been identified in human cells: the major or U2-dependent spliceosome, which catalyzes the removal of the vast majority of introns (U2-type introns), and the less abundant minor or U12-dependent spliceosome, which splices the rare U12-type class of introns (reviewed in Patel & Steitz, 2003). Both major and minor spliceosomes are comprised of five small nuclear RNPs (snRNPs) and over 170 proteins, of which approximately 45 are integral part of the spliceosomal snRNPs (reviewed in Will et al., 2012).

Splicing of precursor mRNAs (pre-mRNAs) is a two-step chemical process that first generates a free 3' hydroxyl group on the 5' exon and creates a 2'-5' phosphodiester linkage at the adenosine branch site that causes the intron to adopt a characteristic lariat structure. In the second step, the 3'-OH of the 5' exon attacks the 3' splice site (SS) to form ligated exons and excised intron lariat (Moore, Query, & Sharp, 1993).

Spliceosome assembly is a highly dynamic process that relies on intricate RNA-RNA interactions between spliceosomal snRNAs and short consensus sequences located at the intron boundaries. At the beginning, base-pairing between the extremely conserved 10 nucleotides at the 5' end of U1 and the 5' SS triggers the formation of a complex that commits the pre-mRNA to spliceosome assembly. Subsequently, a highly conserved sequence in U2 base-pairs with the branch site (BS), forming a short U2-BS duplex in which the branch adenosine is bulged out, specifying its 2' OH as the nucleophile for the first catalytic step of splicing. However, the U2 snRNP-BS interaction requires prior binding of U2AF to the 3' ss (Ruskin, Zamore, & Green, 1988).
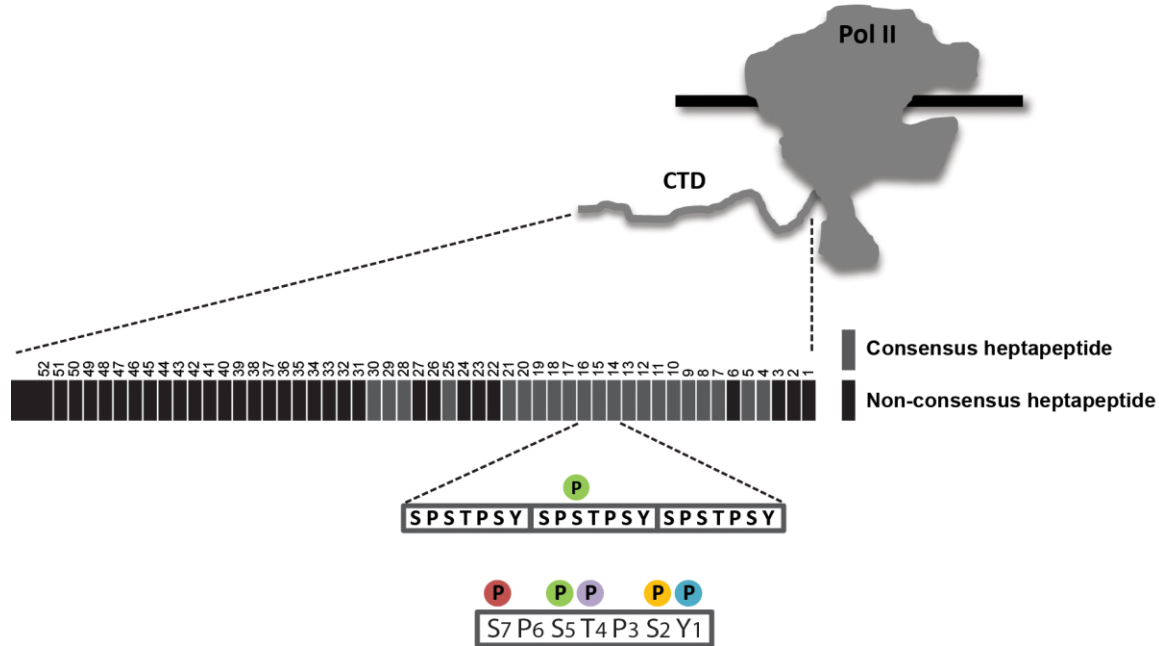
Figure 1.1 CTD structure and modification. The CTD is an inherently unstructured and highly flexible extension appended to Pol II. Human CTD is composed of 52 tandem heptapeptide repeats, of which 21 match the $Y_1S_2P_3T_4S_5P_6S_7$ consensus perfectly. The remaining 31 heptads have one or more amino acid substitutions. The amino acid residues of the heptapeptide can undergo extensive post-translational modifications, namely phosphorylation (P). Recent studies (Schuller et al, 2016 ; Suh et al., 2016) show that on average only one third of the heptads is phosphorylated. (Adapted from Custódio & Carmo-Fonseca 2016).

Thus, the 3' SS is necessary for formation of a pre-spliceosome complex comprising U1 and U2 snRNPs and, consequently, for the first step of splicing. This complex is then joined by the U4/U6.U5 tri-snRNP, giving rise to a fully assembled pre-catalytic spliceosome (or B complex). Within the tri-snRNP, the U6 and U4 snRNAs are extensively base paired with each other, but after association with the pre-spliceosome (or A complex), the U4/U6 interaction is disrupted, and the 5' end of U6 snRNA base pairs with the 5' ss, displacing the U1 snRNA in the process. Extensive base pairing is also formed between U6 and U2, which juxtaposes the 5' ss and BS for the first step of splicing. A central region of the U6 snRNA forms an intramolecular stem-loop structure that appears to play a crucial role in splicing catalysis. Additionally, the U5 snRNA interacts with exon nucleotides near the 5' ss. Following these rearrangements, the spliceosome is ready for the first transesterification step of splicing (C complex). In preparation for step 2, novel RNA-RNA interactions are established. Through base pairing with the two exons, U5 tethers the 5' exon to the spliceosome after step 1 and subsequently aligns both exons for the second transesterification reaction (reviewed by Turner, Norman, Churcher, & Newman, 2004).

What distinguishes U2- and U12-type introns are the sequence elements at the 5' and 3' ss and the BS. The major spliceosome is formed by the U1, U2, U4, U5, and U6 snRNPs, whereas the minor spliceosome is composed of U11, U12, U4atac, U5, and U6atac snRNPs. In U12 introns, the 5' ss is recognized by U11 and the BS base-pairs with U12. Overall, the major and minor spliceosomes share many common features,

with each snRNP in one spliceosome having a counterpart in the other: U1 and U11, U2 and U12, U4 and U4atac, U6 and U6atac (Tarn & Steitz Cell, 1996; Tarn and Steitz Science 1996).

Although several lines of evidence suggest that splicing of pre-mRNA is intimately coupled to Pol II transcription, mechanistic links between the transcription and splicing machineries are little-understood. Truncation of the CTD causes defects in splicing in vivo (McCracken, 1997; Fong, Bird, Vigneron, & Bentley, 2003; Rosonina & Blencowe, 2004) and purified pol II or the CTD itself activates splicing in vitro (Hirose, Tacke, & Manley, 1999 ; Berget, 2000). Remarkably, the CTD only stimulated the splicing of pre-mRNAs that contained complete exons with both 3 and 5' SS, leading to the speculation that the CTD may immobilize consecutive exons, thereby facilitating their juxtaposition (Berget, 2000). Further support to this model was provided by studying the splicing of pre-mRNAs containing introns engineered to be co-transcriptionally cleaved (Dye, Gromak, & Proudfoot, 2006). Disrupting the continuity of the nascent RNA had no effect on splicing of the exons that flanked the cleaved intron implying that newly transcribed consecutive exons are somehow tethered to the transcription complex until they are spliced together (Dye, Gromak, & Proudfoot, 2006). More recently, the laboratories of Professor Proudfoot at the University of Oxford and Professor Carmo-Fonseca at iMM Lisboa developed a strategy for Native Elongation Transcript sequencing using mammalian cells (mNET-seq) and found evidence indicating that splicing occurs within a stable complex formed between the spliceosome and Pol II (Nojima et al., 2015).

The aim of my work was to test the following predictions derived from the model that the splicing reaction is coupled to Pol II: First, spliceosome components, namely U snRNAs, should be detected in association with Pol II; Second, intermediates of the splicing reaction should be detected in association with Pol II; and third, newly spliced products should be detected bound to Pol II.

2. MATERIALS AND METHODS

2.1 mNET-seq and RNA-seq data processing

Data pre-processing was performed as described in Nojima et al., 2016. Read Coverage, gene elements comparison and manipulation were performed using bedtools (v2.25.0) (Quinlan & Hall, 2010). Meta-profiles, splice junction average profiles and individual profiles were generated as described (Nojima et al., 2015).

2.2 Nucleoplasmic RNA-seq data analysis

Strand-specific nucleoplasmic RNA-seq data was analysed with kallisto(v0.42.3) (Bray, Pimentel, Melsted, & Pachter, 2016) using Ensembl Homo sapiens GRCh37.74 release as template for quantification. TPM values for each transcript were converted to log2 and their distribution plotted. The threshold value chosen to identify expressed genes was the local minimum between the peak of high-expressed transcripts and low-expressed transcripts as described in (Hart, Komori, Lamere, Podshivalova, & Salomon, 2013).

2.3 Annotation

Annotations for TSS, introns and exons were retrieved from UCSC Table Browser for hg19 genome version (Karolchik et al., 2004; Rosenbloom et al., 2015).

2.4 Identification of immediately spliced events

Detection of immediately spliced events was based on spliced reads obtained in mNET-seq/S5P experiments. Reads that aligned uniquely and with no mismatch across two exons were classified as spliced reads. To reduce the possibility of fortuitous alignments, only reads that map to more than 3 nucleotides across any junction were considered. A splicing event was considered immediate (IM) if the 3' SS was covered by 10 or more reads, of which at least 90% were spliced.

2.5 Percent-spliced-in (PSI) calculation

To obtain a measure of splicing completeness for any given exon, the percent-spliced-in (PSI) metric (Barbosa-Morais et al., 2012) was applied to nucleoplasmic RNA-seq datasets. When calculating this index, exons that have alternative 3' or 5' splice sites were excluded from the analysis, as well as exons overlapping other exons on the same strand and exons with less than 100bp, or which the preceding intron had less than 100bp. Finally, only junctions that had a total of at least 5 reads in both regions were included in the analysis.

2.6 mNET-seq peak finding analysis

To identify spikes in the density of 3' ends of nascent transcripts we used an algorithm that finds nucleotides where the read density is at least three standard deviations above the mean in a local region

(Churchman & Weissman, 2011). Only positions with coverage of 4 or more reads were considered. To identify peaks at the 3' end of exons and introns indicative of splicing intermediates, the number of reads at these positions was compared to the mean read density across the corresponding exon or intron. To identify peaks corresponding to Pol II pause positions along exons, reads aligning to the last nucleotide position at the 3' end of exons and introns were removed in order to avoid contamination by splicing intermediates. The number of reads at each nucleotide position along the exon was then compared to the mean read density across the entire exon. Peaks were identified in the annotated exons and introns of expressed genes. Exons that intersected other isoform exons were discarded and the same principle was applied for introns. The genomic sequences surrounding pause sites were aligned by peak position and sequence consensus was estimated by WebLogo 3 (Crooks, Hon, Chandonia, & Brenner, 2004).

3. RESULTS

3.1 mNET-seq detects RNAs with free 3'OH in association with Pol II isoforms

To generate mNET-seq datasets, chromatin-bound Pol II and associated nascent transcripts are first purified by immunoprecipitation with phospho-CTD specific antibodies. Next, RNAs derived from within the Pol II complexes are sequenced (Nojima et al., 2015 ; Nojima et al., 2016). In brief, the mNET-seq protocol starts with the isolation of a native chromatin fraction from purified nuclei. This chromatin fraction is then extensively digested with Micrococcal nuclease (MNase). This is a very efficient nuclease that digests all exposed DNA and RNA, effectively releasing the Pol II elongation complex from chromatin into a soluble fraction. Only RNA protected by protein complexes such as Pol II and the spliceosome will be resistant to MNase digestion. Nascent RNA fragments of different sizes are detected in Pol II complexes. Since MNase RNA digestion products have a 5'OH this is first converted back to a 5'P by kinase treatment. Consequently the isolated RNA fraction will have 5' phosphate and 3' hydroxyl (OH). Specific RNA ligation of a linker to the 3'OH allows directional sequencing of the RNA 3' ends present at the Pol II active site (**Figure 3.1**).
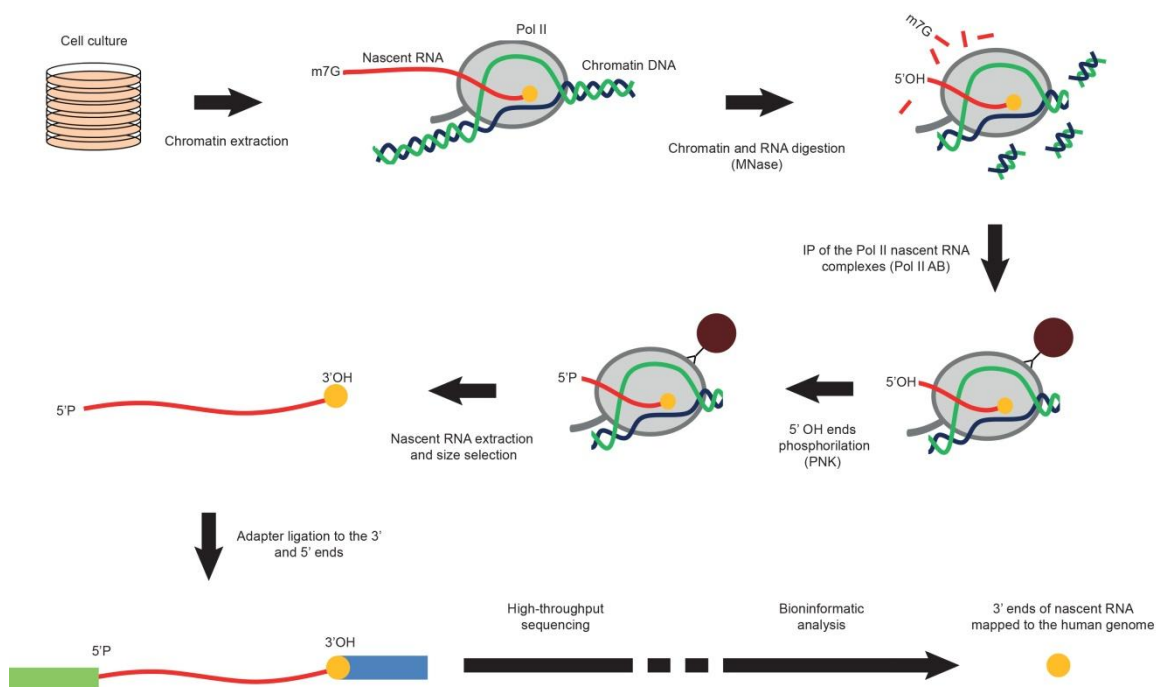

Figure 3.1 Overview of the mNET-seq methodology.

3.2 Spliceosomal U snRNAs associate with Pol II

Assuming that the spliceosome associates with Pol II for co-transcriptional splicing, we reasoned that mNET-seq should detect the free 3' OH ends of mature spliceosomal snRNA associated with Pol II (**Figure 3.2.1**). To analyze spliceosome snRNAs, a list of ENSEMBL genes was filtered to select all genes containing U1, U2, U4, U5, or U6 in the respective gene symbol, and genes were grouped by each snRNA category. This list was then manually curated to remove incorrectly classified and overlapping genes. To determine which snRNA genes are expressed in HeLa cells, we analyzed nucleoplasmic RNA-seq datasets. Genes with a log2(RPM) equal or greater than 1 were selected as expressed. For this set of snRNA genes, reads obtained in mNET-seq experiments were counted and normalized to RPM, adding 5 bases to the annotated coordinates to account for possible annotation uncertainties. The results show an accumulation of mNET-seq signal mapping to the 3' ends of U1, U2, U4 and U5 snRNAs (**Figure 3.2.2**). In contrast, no mNET-seq peak was found at the end of U6 snRNA (**Figure 3.2.2**). This result is in agreement with the previously reported observation that the predominant form of human U6 snRNA terminates with a 2',3' cyclic phosphate group (Lund & Dahlberg, 1991). The strongest U5 snRNA 3' end signal suggests that relative to the other spliceosomal components the U5 snRNP is more efficiently protected from MNase digestion. When cells were treated with the splicing inhibitor pladienolide B (Pla-B) (Kotake et al., 2007) prior to mNET-seq analysis, the 3' end peaks were effectively eradicated over the U5 snRNA gene end (**Figure 3.2.3**) indicating that the association is splicing-dependent. Antibodies to CTD S2P precipitated significantly less reads mapping to U5 snRNA 3' end, and antibodies to S7P, T4P and Y1P yielded residual 3' end signals (**Figure 3.2.4**). Altogether this data strongly supports the view that S5P Pol II forms a stable entity with the catalytically active spliceosome.
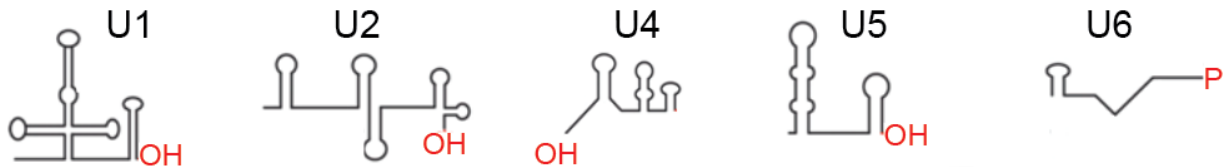
Figure 3.2.1 Spliceosomal snRNA structure overview.
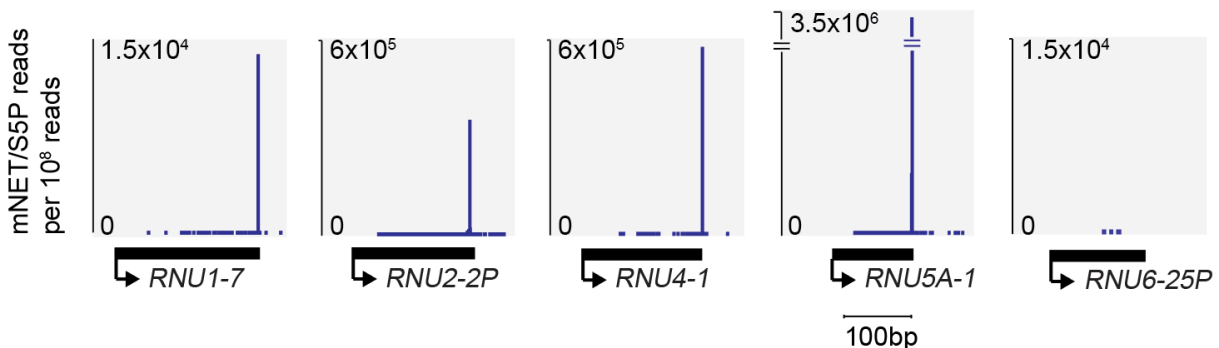
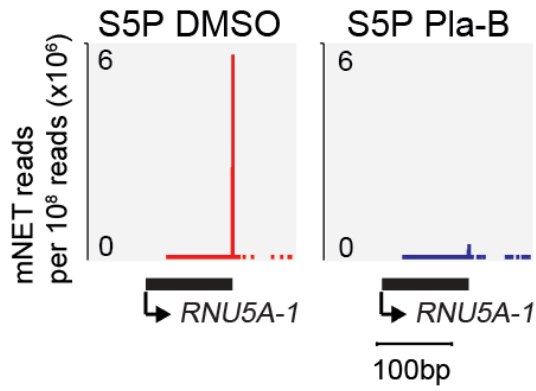**Figure 3.2.2 Spliceosomal snRNA genes mNET-seq S5P profiles.**

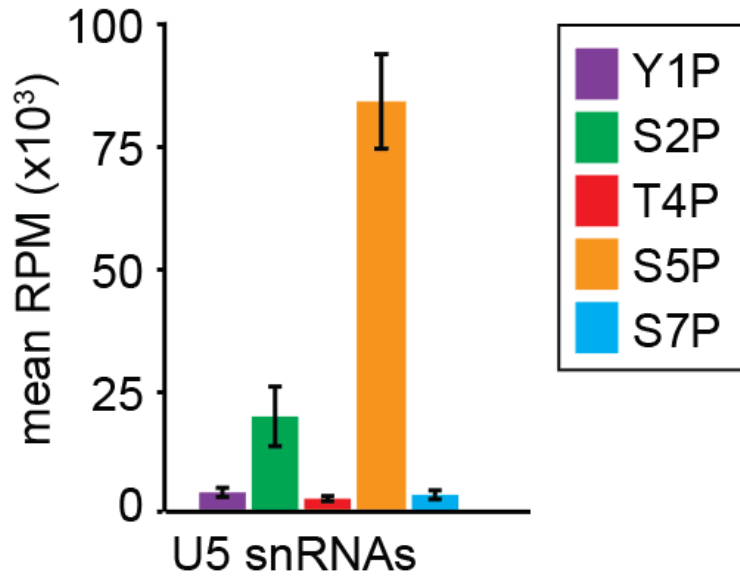**Figur**e **3.2.3 RNU5A-1 gene mNET-seq DMSO/PlaB profiles.**

Figure 3.2.4 Color coded mean RPM counts of U5 spliceosomal snRNA genes for mNET-seq Y1P/S2P/T4P/S5P/S7P datasets.

**3.3 mNET-seq reveals widespread co-transcriptional splicing**

A feature of the mNET-seq method is that only short RNA fragments are protected from MNase digestion, being the nascent RNA directly synthesized by elongating Pol II or RNA cleaved during co-transcriptional events such as pre-mRNA splicing and pre-miRNA processing (Nojima et al., 2015; Nojima et al., 2016). In our previous mNET-seq analysis we selected immunoprecipitated RNA fragments containing 35 to 100 nucleotides to generate libraries for paired-end Illumina high-throughput sequencing with a read length of 50 bp (Nojima et al., 2015). Here we collected and sequenced longer RNA fragments (60 to 160 nucleotides) using a read length of 150 bp. Sequencing of libraries derived from long RNA fragments revealed mNET-seq peaks at the very end of exons indicative of co-transcriptional cleavage of 5' splice sites. These derive from ligation of sequencing adaptors to 3'-OH ends of splicing intermediates where exons have been cleaved at the 5' splice site but not yet ligated to neighboring exons results in mNET-seq peaks at the very end of exons (Nojima et al., 2015). Analysis of nucleoplasmic RNA-seq data reveals that exons with a mNET-seq peak at the end are included in the mature mRNA as shown for exons 5 and 6 of *DDX3X* gene (**Figure 3.3.1**).
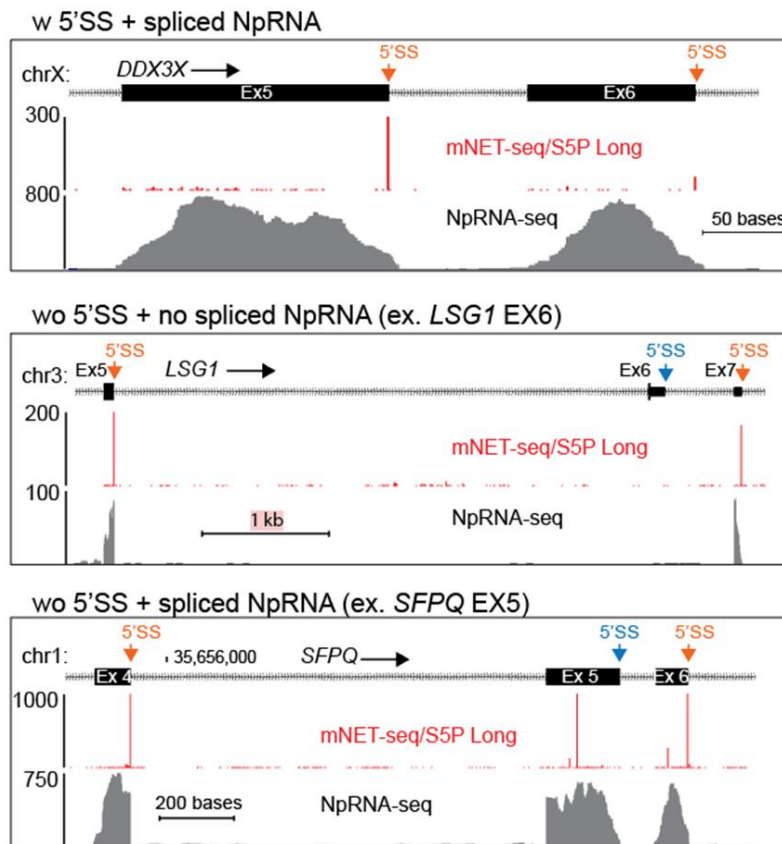
Figure 3.3.1 mNET-seq S5P and Nucleoplasm RNA-seq profiles. Sections of the following genes: *DDX3X, LSG1, SFPQ*.

The ability of mNET-seq to map splicing intermediates enables in-depth investigation of the extent of co-transcriptional splicing in vivo. To identify exons with mNET-seq peaks at the end we adapted a previously described algorithm that finds nucleotides where the read density is at least three standard deviations above the mean in a defined window (Churchman & Weissman, 2011). Upon sequencing three independent libraries we found approximately 25 thousand exons that consistently had a peak at the end (**Figure 3.3.2**).

3.3.2
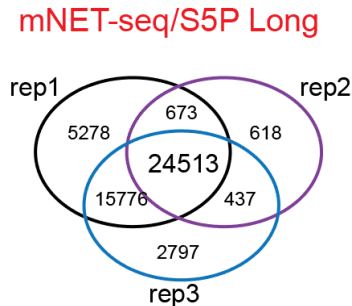
Figure 3.3.2 Venn diagram of all exons with with a peak at the end for 3 independent mNET-seq S5P replicates.

3.3.3
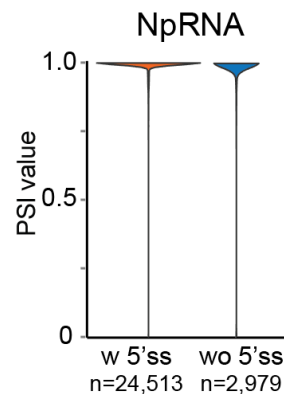
Figure 3.3.3 Distribution of PSI values for the exons with/without a peak at the 5' SS.

A global analysis of the percent splicing index (PSI) that provides the inclusion level of each exon (Barbosa-Morais et al., 2012) confirms that the vast majority of exons with an mNET-seq peak at the end are fully included in mature mRNAs (PSI > 0.9; **Figure 3.3.3**). As expected, the majority of exons lacking a 5'ss peak were absent in the processed mRNA, as shown for exon 6 of *LSG1* gene (**Figure 3.3.1**). However, we further identified approximately 3 thousand exons that show little or no accumulation mNET-seq signal at the 5'ss and yet are spliced in the mature mRNA (**Figure 3.3.3 and 3.3.4**), as shown for exon 5 of the *SFPQ* gene (**Figure 3.3.1**).
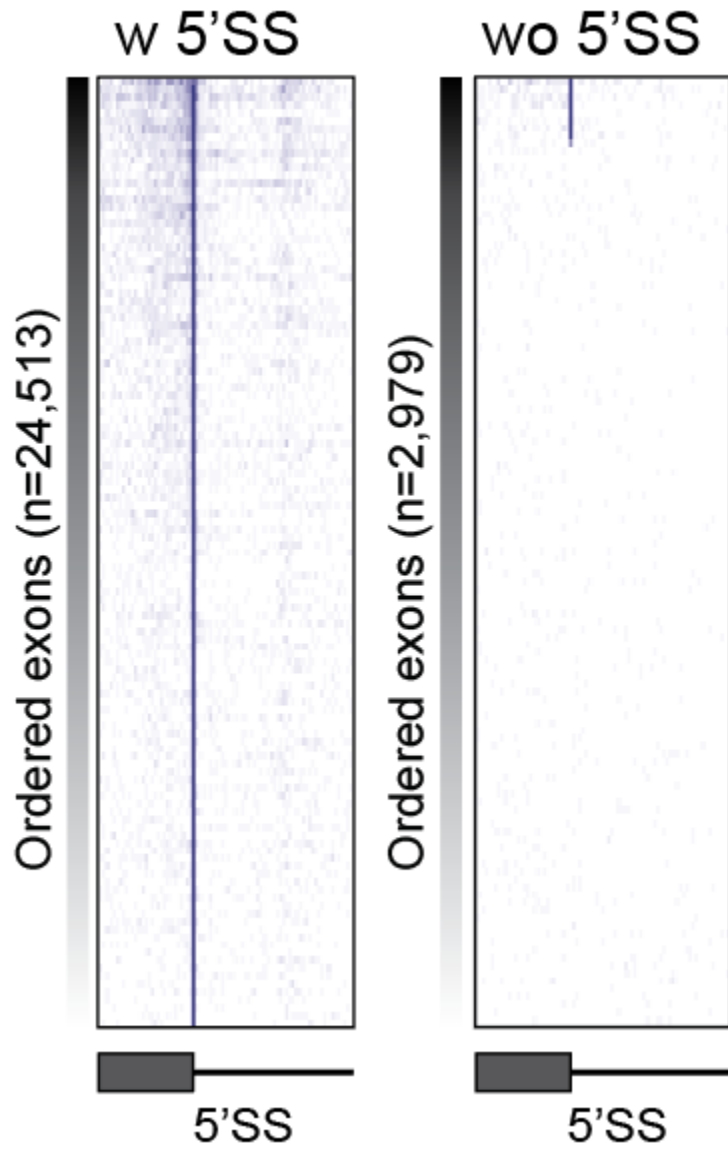
Figure 3.3.4 Heatmaps of mNET-seq/S5P profiles ordered by signal intensity at 5' SS. Profiles are aligned to 5' SS.
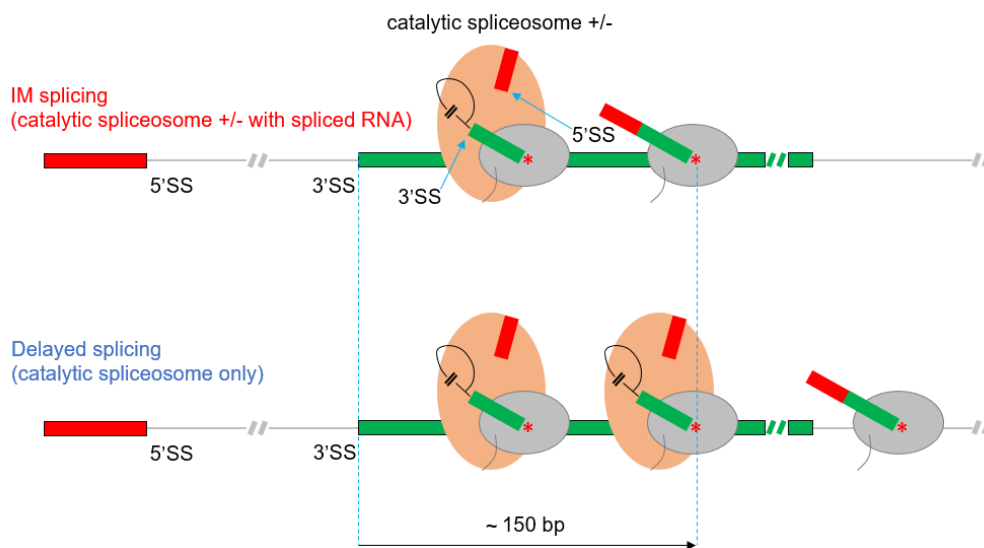
3.4 Immediate and delayed co-transcriptional splicing

Depending on how fast splicing occurs following complete intron extrusion from the Pol II exit channel, then mNET-seq spliced reads reflecting exon joining might be detectable. We do indeed detect significant numbers of mNET-seq reads that aligned across two exons. To reduce the possibility of fortuitous alignments, only reads that map to more than 3 nucleotides across any junction were considered (**Figure 3.4.1**).
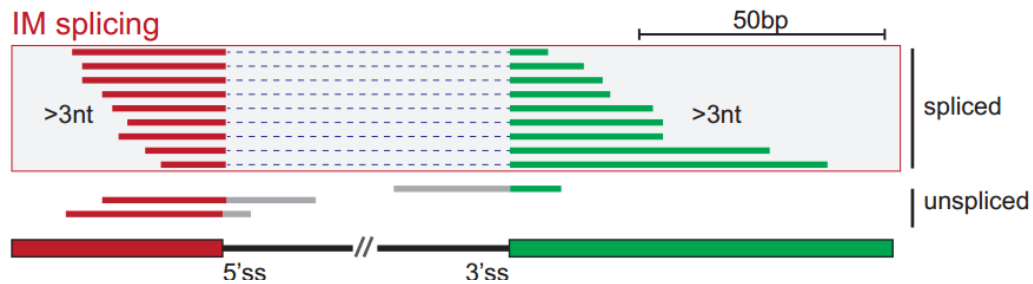

Figure 3.4.1 Schematic illustrating spliced and unspliced mNET-seq reads. Only reads that map to more than 3 nucleotides across any junction were considered.

From 3 independent experiments we identified 1262 splice junctions covered by at least 90% of mNET-seq/S5P spliced reads, corresponding to roughly 3.2% of all included exons (i.e., all exons with PSI>0.9). Because mNET-seq analysis is limited to short RNA fragments, it can only detect splicing events that occur shortly after transcription. Thus, we refer to splicing events detected by mNET-seq as immediate splicing (**Figure 3.4.2**). All other events characterized by mNET-seq 5'ss peaks indicative of splicing intermediates but devoid of spliced reads are referred to as delayed splicing (**Figure 3.4.2**). For these, we cannot map the position of Pol II when spliced occurred.


**Figure 3.4.2 Overview of immediate and delayed splicing.**

To better define the timing of immediate splicing relative to transcription, we calculated the percentage of spliced reads mapping to each bp position after the 3' splice site, determined as the ratio [spliced reads/ (spliced + unspliced reads)]. The first bp position with more than 90% of aligned spliced reads was considered as the position occupied by Pol II when splicing occurs. Analysis of Pol II position at the moment of splicing shows that in the vast majority of cases splicing takes place while the polymerase is still transcribing the exon (**Figure 3.4.3**). In rare cases we detected splicing taking place when the polymerase was already transcribing the downstream intron (**Figure 3.4.4**).
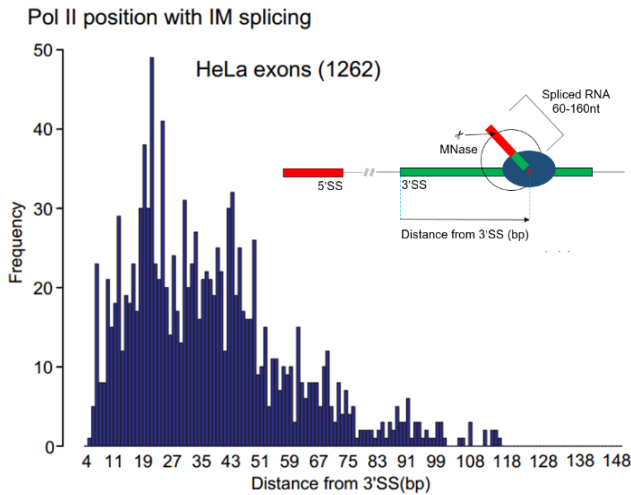
Figure 3.4.3 Distribution of the Pol II position at the moment of splicing in the exon.
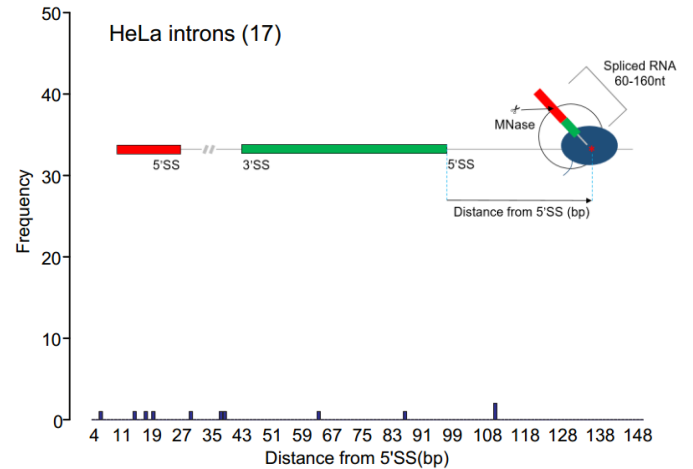
Figure 3.4.4 Distribution of the Pol II position at the moment of splicing in the intron.
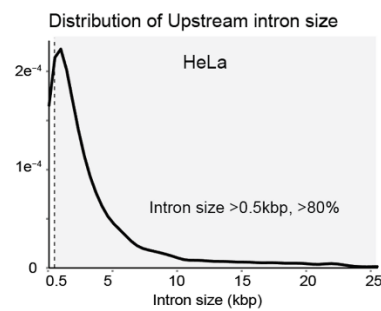

Figure 3.4.5 Density plot depicting upstream intron size distribution. 82% of the introns are above 0.5 kbp.

The remarkable fact that splicing can be completed when Pol II is still transcribing the exon is incompatible with the exon definition model. This model posits that spliceosome assembly in mammals involves interactions with the U1 snRNP bound to the 5' ss downstream of the exon (S. Berget, 1995 ; Sterner, Carlo, & Berget, 1996). As *in vitro* splicing assays indicate that splice sites are recognized across the exon when the intron size is longer than 200-250 bp (Fox-walsh et al., 2005), we determined the size of introns associated with these immediate splicing events (**Figure 3.4.5**). Notably the majority (82%) of

introns in this category are over 500 bp in length. Apparently splice site recognition mechanisms operate differently *in vivo* than they do *in vitro*.

Previous studies indicate that alternative and U12-associated splicing tends to occur later than splicing of constitutive exons and U2-type introns (Khodor et al., 2012; Turunen et al., 2013). In agreement, we did not detect immediately spliced introns flanking alternatively spliced exons. We also did not find minor
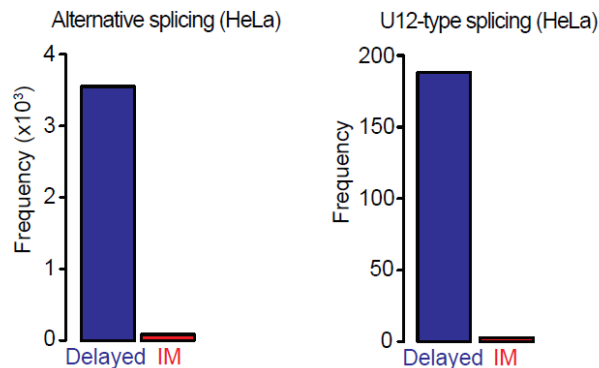
U12-type introns in the fast splicing category (**Figure 3.4.6**).

Figure 3.4.6 Immediate splicing excludes alternative splicing and U12 introns. Only very few alternative and minor U12-type introns are included in the fast splicing category; most probably these result from mis-annotation.

3.5 Pol II pauses for immediate splicing

It is well established that transcription elongation rates influence splicing outcomes (Fong et al., 2014 ; Dujardin et al., 2014) and higher Pol II density at human exons versus introns was reported using ChIP (Brodsky et al., 2005 ; Schwartz, Meshorer, & Ast, 2009) and NET-seq (Mayer et al., 2015). High Pol II density across exons with particular enrichment towards the beginning of exons was further observed in Drosophila using PRO-seq (Kwak et al., 2013).

Analysis of mNET-seq/S2P and mNET-seq/S5P data revealed significantly higher Pol II density at co-transcriptionally spliced exons relative to flanking introns, and the difference is particularly striking for S5P modified Pol II (**Figure 3.5.1**). This suggests that co-transcriptional splicing imposes a slower transcription elongation rate at exons and that these exons are predominantly transcribed by S5P Pol II.

Moreover, analysis of S5P Pol II mNET-seq profiles across exons associated with immediate splicing revealed a prominent spike in the density of 3'-ends of nascent transcripts indicative of Pol II pausing. This peak shown for exon 5 of *OTUD3* results from accumulation of nascent spliced transcripts with the same 3' end suggesting that the polymerase is paused at the position when splicing occurs. The peak is substantially reduced following treatment with the splicing inhibitor Pla-B indicating that pausing is splicing-dependent (**Fig. 3.5.2**).
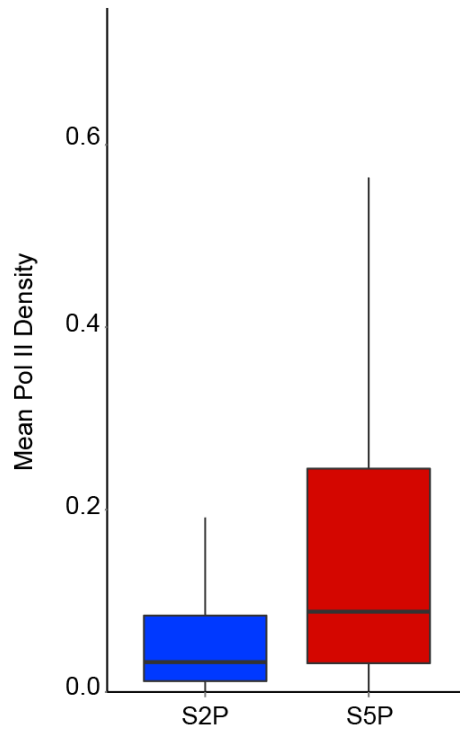
Figure 3.5.1 Box-plots representing the mean Pol II density difference for Co-Transcriptional exons and the respective upstream introns. mNET-seq S2P/S5P datasets used. The * is indicative of p-value < 0,001 resulting from the application of Wilcoxon test for paired samples.
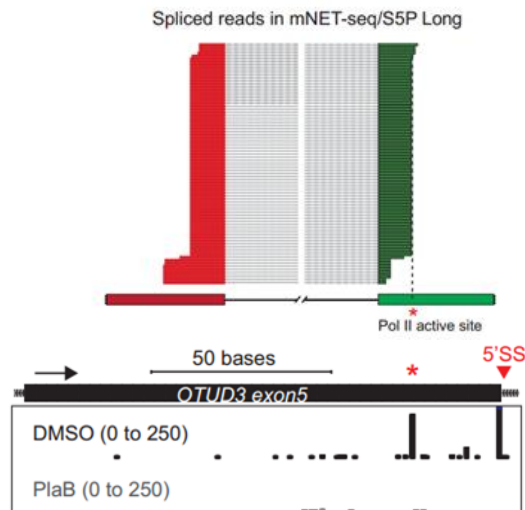
Figure 3.5.2 Schematic depicting the accumulation of spliced mNET-seq reads interpreted as Pol II pausing. mNET-seq DMSO/PlaB profiles for exon number 5 of *OTUD3* gene. The red * highlights the Pol II pause position.

To characterize in more detail the Pol II density profile across all immediate exons, we generated a heatmap of mNET-seq signal within a 10 bp sliding window showing exons sorted by the distance between the 3'ss and the position when splicing occurs (**Fig. 3.5.3**). In each exon, a 10 bp region stands out for having the highest mNET-seq/S5P density value along that exon and flanking introns. For all exons analyzed, the position when splicing occurs is located within that region. No such strong density spike is detected in mNET-seq/S2P datasets (**Fig. 3.5.3**), suggesting that the paused Pol II is modified on S5P, not S2P.

We further used the previously described peak calling algorithm to identify positions where the read density is highest along the entire length of the exon. In agreement with the density heatmap, only one major peak was generally detected per exon. These peaks are preferentially located towards the beginning of immediately spliced exons (**Figure 3.5.4**).
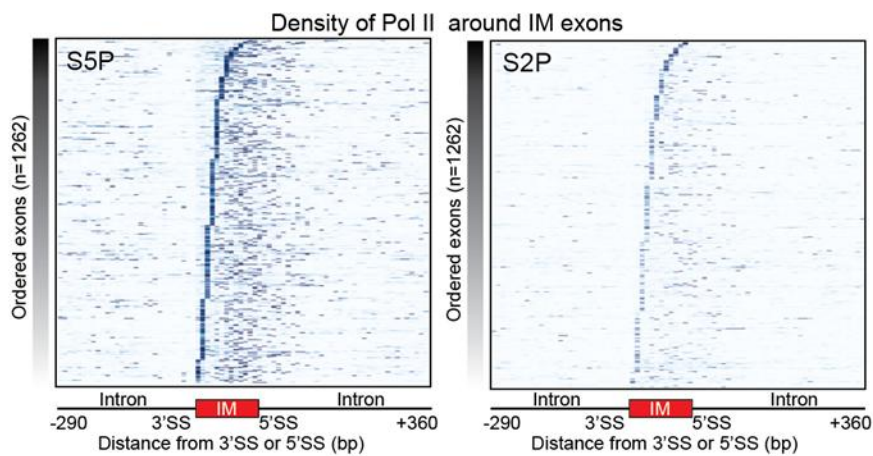
Figure 3.5.3 Heatmaps for the Pol II density of immediate exons. Exons are ordered by increasing distance of the position when splicing occurs relative to the 3' SS.
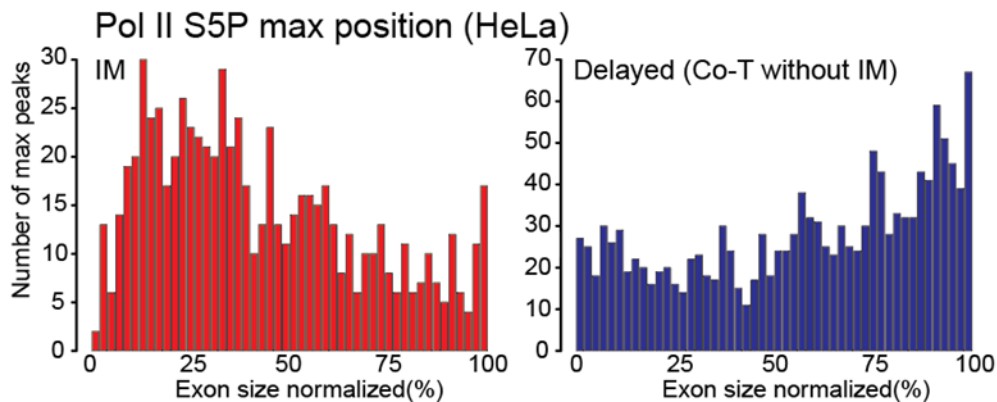
Figure 3.5.4 Pol II pause (location of peak) distribution in immediate exons and co-transcriptional exons that are not immediate.

Notably, even though the splicing-coupled pause sites that we identify have no nucleotide preference at the 3' end of the transcript, the majority of pause sites are immediately followed by a specific TGGA
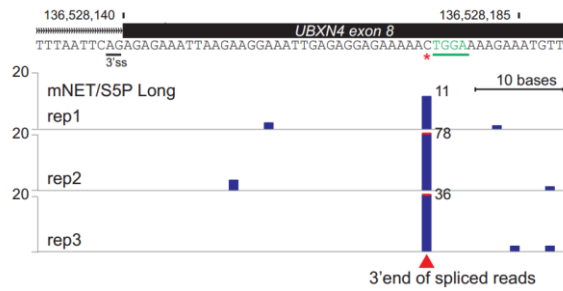
sequence (**Figure 3.5.5 and 3.5.6**).

Figure 3.5.5 mNET-seq S5P profiles for 3 independent replicates. The red * highlights the Pol II pause position. **B**
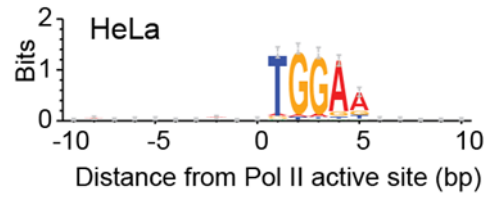
Figure 3.5.6 DNA sequence Weblogo aligned by Pol II pause position.

4. CONCLUSIONS

Analysis of co-transcriptional splicing detected by mNET-seq has revealed a number of hitherto unanticipated features of mammalian splicing. As previously described (Nojima et al., 2015), we observed that actively spliced exons correlate with S5P CTD modified Pol II. Furthermore, a splicing intermediate generated by the first catalytic step in splicing was detected. We hypothesize that this cleaved upstream exon is retained on the elongating Pol II complex within a co-immunoprecipitated spliceosome. Our observation that U1, U2, U4 and U5 snRNAs are present in Pol II complexes provides additional evidence that the catalytically active spliceosome is directly associated with S5P CTD modified Pol II during co-transcriptional splicing. A remarkable outcome of this study is the realization that most splicing in HeLa cells occurs co-transcriptionally, but with radically different kinetics. By analyzing the complete set of splicing events detectable in HeLa cell nucleoplasmic mRNA, we show that spliced exons generally give a significant 5'SS mNET-seq/S5P peak which is diagnostic of co-transcriptional cleavage at the intron 5' ss. This is generated by the first catalytic step in splicing. We further show that a subset of spliced exons complete the whole splicing reaction of joining adjacent exons together as soon as the 3' end of the intron including its 3' branch site, polyprimidine tract and 3'ss emerges from the Pol II RNA exit channel. Thus spliced reads are associated with these immediately spliced introns even though only 60 to 160 nucleotides RNA sequences can be determined in mNET-seq due to the limited length of RNA protected by the Pol II complex from micrococcal nuclease digestion. To give time for this rapid reaction to occur, the spliceosome acts to pause Pol II right at the beginning of the downstream exon. Notably a clear sequence consensus TGGA is positioned a few nucleotides downstream of this pause site. Interestingly this same TGGA motif has been previously associated with a small class of constitutive but not cryptic exons in a mouse transcriptomic study (Zavolan et al., 2003). This may point to evolutionary conservation of the fast splicing mechanism in mammals. Again this exon pausing effect is specific to the S5P CTD form of Pol II. Unexpectedly this Pol II pausing is itself induced by splicing, as it is substantially reduced by pre-treatment with the splicing inhibitor Pla-B. This argues that the association of the spliceosome on the upstream intron, recruited by interactions with Pol II also acts to restrain elongation of the polymerase. This will result in immediate pausing of Pol II in the downstream exon. Our findings on splicing induced exon pausing is in agreement with previous studies in budding yeast (Alexander, Innocente, Barrass, & Beggs, 2010).

5. REFERENCES

Alexander, R. D., Innocente, S. A., Barrass, J. D., & Beggs, J. D. (2010). Splicing-Dependent RNA Polymerase Pausing in Yeast. *Molecular Cell*, *40*(4), 582–593. doi:10.1016/j.molcel.2010.11.005

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., & Gueroussov, S. (2012). The Evolutionary Landscape of Alternative Splicing in Vertebrate Species, *1587*. doi:10.1126/science.1230612

Bentley, D. L. (2014). Coupling mRNA processing with transcription in time and space. *Nature Publishing Group*, (February). doi:10.1038/nrg3662

Berget, S. (1995). Exon Recognition in Vertebrate Splicing.

Berget, S. M. (2000). Participation of the C-Terminal Domain of RNA Polymerase II in Exon Definition during Pre-mRNA Splicing, *20*(21), 8290–8301.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification, (October 2015), 4–7. doi:10.1038/nbt.3519

Brodsky, A. S., Meyer, C. A., Swinburne, I. A., Hall, G., Keenan, B. J., Liu, X. S., … Silver, P. A. (2005). Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells, *6*(8), 1–9. doi:10.1186/gb-2005-6-8-r64

Churchman, L. S., & Weissman, J. S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, *469*(7330), 368–373. doi:10.1038/nature09652

Cramer, P., & Pesce, C. (1997). Functional association between promoter structure and transcript, *94*(October), 11456–11460.

Crooks, G. E., Hon, G., Chandonia, J., & Brenner, S. E. (2004). WebLogo : A Sequence Logo Generator, 1188–1190. doi:10.1101/gr.849004.1

Custódio, N., & Carmo-fonseca, M. (2016). Co-transcriptional splicing and the CTD code, *9238*(September). doi:10.1080/10409238.2016.1230086

Dujardin, G., Lafaille, C., Mata, M. De, Marasco, L. E., Mun, M. J., Jossic-corcos, C. Le, … Kornblihtt, A. R. (2014). How Slow RNA Polymerase II Elongation Favors Alternative Exon Skipping, 1–8. doi:10.1016/j.molcel.2014.03.044

Dye, M. J., Gromak, N., & Proudfoot, N. J. (2006). Exon Tethering in Transcription by RNA Polymerase II, 849–859. doi:10.1016/j.molcel.2006.01.032

Eick, D., & Geyer, M. (2013). The RNA Polymerase II Carboxy-Terminal Domain (CTD) Code.

Fong, N., Bird, G., Vigneron, M., & Bentley, D. L. (2003). A 10 residue motif at the C-terminus of the RNA pol II CTD is required for transcription , splicing and 3 ¢ end processing, *22*(16).

Fong, N., Kim, H., Zhou, Y., Ji, X., Qiu, J., Saldi, T., … Bentley, D. L. (2014). Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate, 2663–2676. doi:10.1101/gad.252106.114.nascent

Fox-walsh, K. L., Dou, Y., Lam, B. J., Hung, S., Baldi, P. F., & Hertel, K. J. (2005). The architecture of pre-mRNAs affects mechanisms of splice-site pairing, *2005*.

Hart, T., Komori, H. K., Lamere, S., Podshivalova, K., & Salomon, D. R. (2013). Finding the active genes in deep RNA-seq gene expression studies.

Hirose, Y., Tacke, R., & Manley, J. L. (1999). polymerase II stimulates pre-mRNA splicing, (212), 1234–1239.

Kotake, Y., Sagane, K., Owa, T., Mimori-kiyosue, Y., Shimizu, H., Uesugi, M., … Mizui, Y. (2007). Splicing factor SF3b as a target of the antitumor natural product pladienolide, *3*(9), 570–575. doi:10.1038/nchembio.2007.16

Kwak, S., Hong, S., Bajracharya, R., Yang, S., Lee, K., & Yu, K. (2013). Drosophila Adiponectin Receptor in Insulin Producing Cells Regulates Glucose and Lipid Metabolism by Controlling Insulin Secretion, *8*(7), 2–11. doi:10.1371/journal.pone.0068641

Lund, E., & Dahlberg, J. (1991). Cyclic 2 ', 3 ' -Phosphates and Nontemplated Nucleotides.

Mayer, A., Iulio, J., Stamatoyannopoulos, J. A., Churchman, L. S., Mayer, A., Iulio, J., … Churchman, L. S. (2015). Human Transcriptional Activity at Nucleotide Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution. *Cell*, *161*(3), 541–554. doi:10.1016/j.cell.2015.03.010

McCracken, S. (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription.pdf.

Moore, M. J., & Proudfoot, N. J. (2009). Review Pre-mRNA Processing Reaches Back to Transcription and Ahead to Translation. *Cell*, *136*(4), 688–700. doi:10.1016/j.cell.2009.02.001

Moore, M. J., Query, C. C., & Sharp, P. A. (1993). *Splicing of Precursors to mRNA by the Spliceosome*.

Nojima, T., Carmo-fonseca, M., Proudfoot, N. J., Rita, A., Grosso, F., Kimura, H., … Dhir, S. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing Article Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing, 526–540. doi:10.1016/j.cell.2015.03.027

Nojima, T., Gomes, T., Carmo-fonseca, M., & Proudfoot, N. J. (2016). Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide, *11*(3), 413–428. doi:10.1038/nprot.2016.012

Patel, A. A., & Steitz, J. A. (2003). SPLICING DOUBLE : INSIGHTS FROM THE SECOND SPLICEOSOME, *4*(December). doi:10.1038/nrm1259

Quinlan, A. R., & Hall, I. M. (2010). BEDTools : a flexible suite of utilities for comparing genomic features, *26*(6), 841–842. doi:10.1093/bioinformatics/btq033

Rosonina, E., & Blencowe, B. J. (2004). Analysis of the requirement for RNA polymerase II CTD heptapeptide repeats in pre-mRNA splicing and 3 ' -end cleavage, 581–589. doi:10.1261/rna.5207204.mechanisms

Ruskin, B., Zamore, P. D., & Green, M. F. (1988). A Factor , U2AF , Is Required for U2 snRNP Binding and Splicing Complex Assembly, *52*, 207–219.

Schuller, R., Schu, R., Forne, I., Straub, T., Cramer, P., Imhof, A., … Decker, T. (2016). Resource Heptad-Specific Phosphorylation of RNA, 305–314. doi:10.1016/j.molcel.2015.12.003

Schwartz, S., Meshorer, E., & Ast, G. (2009). Chromatin organization marks exon-intron structure. *Nature Structural & Molecular Biology*, *16*(9), 990–995. doi:10.1038/nsmb.1659

Sisodia, S. (1987). Specificity of RNA Maturation Pathways : RNAs Transcribed by RNA Polymerase III Are Not Substrates for Splicing or Polyadenylation, *7*(10), 3602–3612.

Smale, S. T., & Tjian, R. (1985). Transcription of Herpes Simplex Virus tk Sequences Under the Control of Wild-Type and Mutant Human RNA Polymerase I Promoters, *5*(2), 352–362.

Sterner, D., Carlo, T., & Berget, S. (1996). Architectural limits on split genes, *93*(December), 15081–15085.

Suh, H., Ficarro, S. B., Kang, U., Chun, Y., Marto, J. A., & Buratowski, S. (2016). Direct Analysis of Phosphorylation Sites on the Rpb1 Article Direct Analysis of Phosphorylation Sites on the Rpb1 C-Terminal Domain of RNA Polymerase II. *Molecular Cell*, *61*(2), 297–304. doi:10.1016/j.molcel.2015.12.021

Tarn, W., & Steitz, J. A. (1996). A Novel Spliceosome Containing U11 , U12 , and U5 snRNPs Excises a Minor Class (AT − AC) Intron In Vitro, *84*, 801–811.

Tarn, W., & Steitzt, J. A. (1996). Highly Diverged U4 and U6 Small Nuclear RNAs Required for Splicing Rare AT-AC lntrons, *273*(September).

Turner, I. A., Norman, C. M., Churcher, M. J., & Newman, A. J. (2004). Roles of the U5 snRNP in spliceosome dynamics and catalysis, *32*, 928–931.

Will, C. L., Lührmann, R., Noller, H. F., Gottesman, S., Storz, G., Volpe, T., … Lu, R. (2012). Spliceosome Structure and Function. doi:10.1101/cshperspect.a003707

Zavolan, M., Kondo, S., Scho, C., Adachi, J., Hume, D. A., Group, R. G. E. R., … Gaasterland, T. (2003). Impact of Alternative Initiation , Splicing , and Termination on the Diversity of the mRNA Transcripts Encoded by the Mouse Transcriptome, 1290–1300. doi:10.1101/gr.1017303.7