UNIVERSITY OF CALIFORNIA SAN DIEGO

Splice Site Recognition During Early Spliceosome Assembly

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Chemistry

by

Charles Leonard Mallari Lumba

Committee in charge:

Professor Gourisankar Ghosh, Chair
Professor Judy Kim
Professor Ulrich Muller

2019

The Thesis of Charles Leonard Mallari Lumba is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California San Diego

2019

DEDICATION

I dedicate this to my girlfriend, Yeonjoo, and my family, Charlie, Marilyn, and Charlene, for their constant support throughout my journey.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

BME	β-mercaptoethanol
DTT	dithiothreitol
FL	full-length
WT	wild type
SDS	sodium dodecyl sulfate
PAGE	polyacrylamide gel electrophoresis
PIC	protease inhibitor cocktail
PMSF	phenylmethylsulfonyl fluoride
snRNP	small nuclear ribonucleic protein
snRNA	small nuclear RNA
SR	serine arginine rich
U2AF	U2 auxiliary factor
SF	splicing factor
SS	splice site
BS	branch site
PPT	polypyrimidine tract
MBP	maltose binding protein
Bg	beta-globin

AdML	adenovirus major late
RBD	RNA binding domain
RRM	RNA recognition motif
RS	arginine serine rich domain
RE	arginine glutamate (mutated RS domain)
RBD	RNA binding domain

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

First and foremost, I would like to thank my principle investigator, Professor Gourisankar Ghosh. It is through meeting him that I learned how fortunate one is to be able to conduct research, and how valuable such a gift is. With his passion and charisma, he has taught me the joys of science, and that through an earnest approach coupled with diligent and careful work, special things can come about.

I extend my thanks to my lab members and the friends I have made in my program. Being new to biochemistry it was an arduous experience learning the ropes but being surrounded by such lovely people has made this ordeal so much easier. I give special thanks to Dr. Kaushik Saha for his unending patience with me, guiding me not only in science but also in life. Thank you to Mike Fernandez, demonstrating what independent thinking and drive backed by insightful thought can accomplish. Shandy Shahabi the older brother I never had, Myung Soo Ko, Kyle Shumate, Sonjiala Hotchkiss, Lei Yu, Dr. Maria Mulero-Roig, Dr. Tapan Biswas, Samantha Cohen, Dr. Yidan Li, Dr. Anup Mazumder, Dr. Vivien Wang, Qian Du, Boqing Gu, and Yunshu Song, I am indebted to you all. A special thanks to Dr. Ulrich Muller and Dr. Judy Kim for being part of my committee. Dr. Muller, your in-depth analyses and how you lead discussions within your seminar course set the bar for the breadth of critical thinking expected for my graduate studies. Thank you to Dr. Kim as well for demonstrating true thoughtfulness and kindness when working with students, the opportunities I had teaching with you proved valuable in my own teaching approach.

A much deserved thanks to my family and my awesome group of friends, without you I would not be where I am today. I appreciate your unwavering support in whatever I choose to do, for giving me a place to always return to. Fostering such a loving environment, whether in the comfort of my own home or in the company of my closest friends, the stress from everyday life is

lessened. Thank you for pushing me to strive for more and to be the best version of myself. To my parents, you are outstanding role models on how life should be lived, sincerely and with an open heart.

Finally, I acknowledge my girlfriend Yeonjoo Choi. Your belief in me and constant cheering throughout this process has given me so much strength, time and time again. You have added so much color to my world. With you it has been nothing but the highest of highs and I can't imagine life without you. Thank you for loving me.

ABSTRACT OF THE THESIS

Splice Site Recognition During Early Spliceosome Assembly

by

Charles Leonard Mallari Lumba

Master of Science in Chemistry

University of California San Diego, 2019

Professor Gourisankar Ghosh, Chair

Protein coding sequences in most eukaryotic pre-messenger RNAs (pre-mRNA) are interrupted by intervening sequences called introns, while the protein coding sequences are termed exons. The process by which introns are removed, and exons are joined, generating a processed mRNA is called splicing, catalyzed by the spliceosome enzyme. Within pre-mRNA the spliceosome distinguishes exon and intron boundaries, also referred to as splice sites (SS). In

addition to SS, there are two essential, but highly degenerate sequence elements within an intron, a branch site (BS) and a poly-pyrimidine tract (PPT) sequence, both near the 3'SS. Lacking such highly conserved features in pre-mRNAs, how the splicing machinery recognizes authentic SS amongst several similar sequences remains unclear. In the first step of spliceosome assembly U1 snRNP recognizes the 5'SS, while the BS, PPT and 3'SS are recognized by their respective partners, SF1, U2AF65 and U2AF35. However, each factor interacts with its partner RNA sequence *in vitro* with low affinity, suggesting that early pre-mRNA substrate recognition must be more complex. The serine-arginine (SR) protein family is thought to facilitate early spliceosome assembly, but no systematic analysis on SR protein mediated early spliceosome assembly has been done. Using *in vitro* binding experiments, I found that the interaction between U1 snRNP and pre-mRNA through the 5'SS is assisted by the presence of the PPT, suggesting both the 5'SS and 3'SS are jointly recognized very early in assembly. I further show that an SR protein can support U1 snRNP binding in certain splicing substrates but not all.

I. Introduction

A. RNA Splicing and the Spliceosome

In the human genome only 2% is devoted to protein coding sequences, equating to roughly 25,000 protein coding genes (Ast, 2004). Additionally, in over 98% of these coding sequences, genes are interrupted with intervening sequences. Therefore, the transcript of a protein coding gene, referred to as the pre-messenger RNA (pre-mRNA), contains both the coding and noncoding intervening sequences. The segments within the pre-mRNA that code for proteins are known as exons, while the non-coding sequences flanked by two exons are known as introns. Pre-mRNA splicing is an RNA processing event by which the intervening intronic sequences are removed, and exons are then joined together forming a mature mRNA. Among these splicing events, there are two major types- constitutive and alternative splicing. In constitutive splicing an exon is always included in the mRNA, diversely in alternative splicing, there can be exon skipping or inclusion events (Figure 1.1). Consequently, alternative splicing then gives rise to tens of splice variants from a single pre-mRNA comprised of tens of exons, while each of the splice variants can perform distinct cellular functions (Modrek et al., 2001). Therefore, from a limited gene pool, a significantly greater number of functionally distinct splice variants, and hence proteins, can be generated (Graveley 2001, Smith & Valcárcel, 2000).

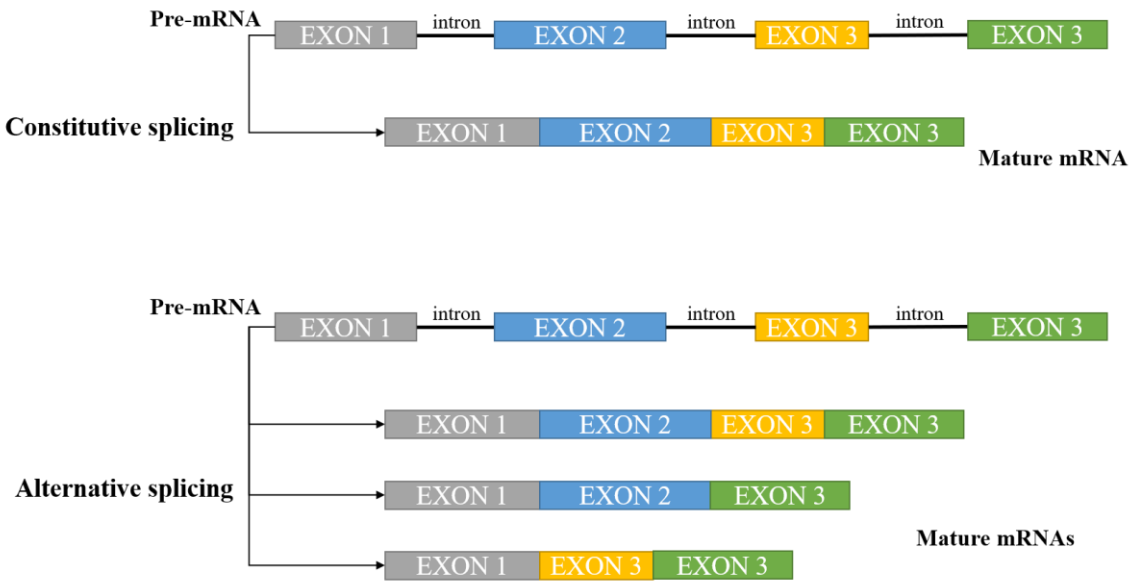

Figure 1.1 Constitutive vs alternative splicing
Schematic illustrating constitutive splicing, where all exons are included, and alternative splicing, where exons may or may not be included. The result is different mRNAs, thus encoding for different proteins, derived from a single gene.

The splicing reaction is catalyzed by an intricate RNA-protein complex machine known as the spliceosome. The human spliceosome is composed of five uridine-rich ribonucleoproteins (snRNPs) complexes and numerous other protein factors, at least 11 RNA helicases and several SR proteins, all of which are essential (Will & Luhrmann, 2011). The snRNPs and other factors assemble to the pre-mRNA in multiple steps (Figure 1.2). The assembly can be divided into major steps; first, the recognition of authentic exon-intron-exon boundaries and second, formation of the catalytic enzyme, in which the substrate becomes part of the enzyme. The reaction is characterized by two sequential transesterification steps. In the first step, the exon-intron boundary at the 5'SS junction is cleaved and in the second step the two exons are joined removing the intron as a lariat structure. Thus, the spliceosome does not function as a classical

enzyme where a single enzyme performs multiple turnover of substrate into product. Instead, the spliceosome carries out a single turnover of the substrate into product.

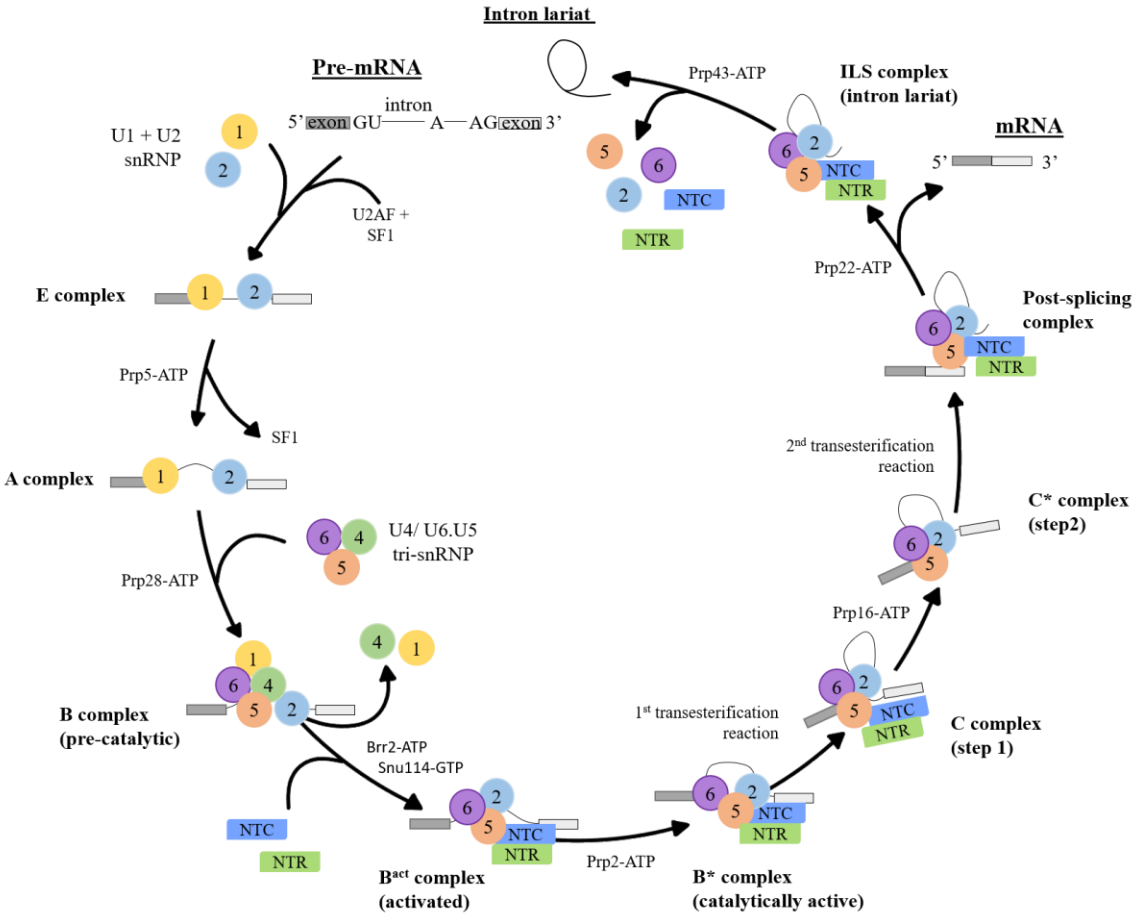

Figure 1.2 Pre-mRNA splicing by the spliceosome
Schematic representation of a single pre-mRNA which the spliceosome assembles on, a stepwise process involving the five snRNPs along with numerous other protein factors, resulting in a mature mRNA.

B. SR Protein Family

In the case of alternative splicing, a single gene can code for multiple proteins as exons may or may not be skipped. Many factors can regulate splicing, in mammals specifically, serine-

arginine-rich (SR) proteins are thought to play a role in both constitutive and alternative splicing (Zhou & Fu, 2013). This family of RNA binding proteins contain an N-terminal RNA binding domain (RBD), one or two RNA recognition motifs (RRM), and an Arg-Ser-rich (RS) domain at the C-terminus. Examining the splice signals within the pre-mRNA, if these are weak, in terms of how closely it aligns with the consensus sequence (Figure 1.3), this can prevent incorporation of the exon. It is thought that the SR proteins facilitate inclusion of alternative exons by binding to exonic splicing enhancer sequences, or ESE (Zhu et al, 2001). Typically, ESEs are short single stranded RNA and each SR protein binds to a specific set of ESEs (Wang et al., 2005). However, ESEs are highly degenerate, which questions sequence specificity in the interaction between SR proteins and ESEs. Recent genome-wide experiments have also revealed that SR proteins bind to entire length of exon in multiple copies (Pandit et al., 2013).

Additionally, our laboratory has shown that the RS domain of SRSF1, one of the SR proteins, must undergo phosphorylation to participate in splicing (Cho et al., 2011). SR protein phosphorylation is performed by two kinases, SR protein kinase (SRPK) and Clk (cdc2-like kinases), as shown in Figure 1.4 (Gui et al., 1994, Colwill et al., 1996). Phosphomimetic mutants of SRSF1, where serines in the SR domain are mutated to Glu or Asp, can also be involved in the splicing reaction (Cho et al., 2011). The SR proteins are known to recruit snRNPs to the spliceosome, but the mechanism has remained unclear (Chandler et al., 1997). Although there is activity co-transcriptionally of SR proteins to RNA polymerase II, little is known on how regulation of pre-mRNA binding by SR protein occurs, as well as the consequences of said binding (Sapra et al., 2009). With the role of SR proteins unclear, our focus is on the activity of these proteins and how they are involved in assembly of the spliceosome.

5’ splice site	Branch point site	Polypyrimidine tract	3’ splice site
GURAGU	YNCURAC	Rich in pyrimidines	YAG

Figure 1.3 Schematic of metazoan consensus splice signals
List of the sequences comprising the consensus of each splice signal, in the case of the polypyrimidine tract however, there is loose conservation. (Y= pyrimidine, N= any nucleotide, and R= purine).

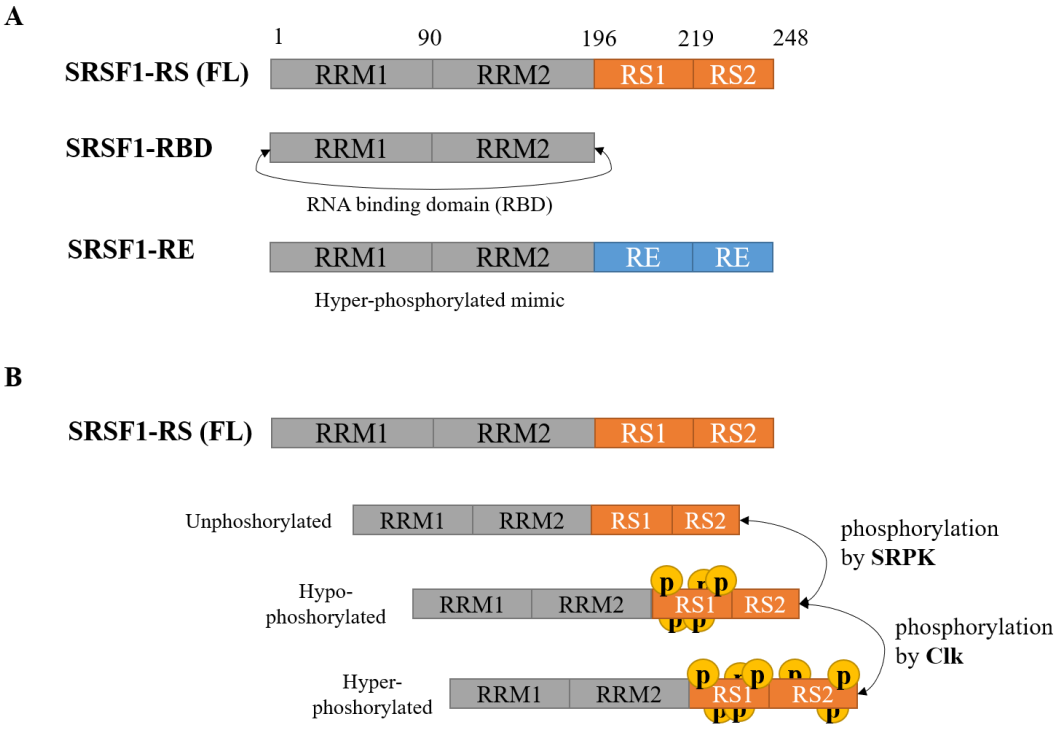

Figure 1.4 Schematic representation of SRSF1 variants and SRSF1 phosphorylation events
Three SRSF1 variants (A) were utilized: full length (FL) SRSF1 consisting of two RRMs followed by the RS domains at the C-terminus, the RNA binding domain (RBD) composed of the two RRMs, and a hyper-phosphorylated mimic, SRSF1-RE (serine to glutamate mutations in RS domain). B) SRSF1 undergoes two phosphorylation events, the 12 serines in RS1 are phosphorylated by SRPK (serine arginine rich protein kinase), and the remaining 6 serines in SRSF1 are phosphorylated by kinase Clk.

C. Early Spliceosome Recognition Complex

Within metazoan pre-mRNA substrates there are four major splice signals (Figure 1.5), the splice sites (SS) at the 5' and 3' ends of the intron, 20-40 nucleotides upstream the 3'SS is the branch-point site (BS), and a poly-pyrimidine tract (PPT) of various lengths between the BS and 3'SS (Will & Luhrmann, 2011). These splice signals are respectively recognized by factors U1 snRNP, U2AF35, SF1, and U2AF65, which are all found in the early spliceosomal complex, the E-complex. Within E-complex a stretch of nucleotides in U1 snRNA of U1 snRNP is recognized by the 5′SS (Kondo et al., 2015). However, the low sequence conservation in the 5′SS can limit complementarity with U1 snRNA, suggesting the presence of additional contacts (Mount et al., 1983, Zillmann et al., 1987). Therefore, how authentic splice sites are faithfully determined is unknown, and attempting to identify the basal components proved difficult considering the amount of proteins involved (Makarov et al., 2012). The 3' components, U2AF35, U2AF65, and SF1 create an intricate network to interact with the 3' end of intron, but the sequence of assembly also remains unclear. It has been suggested that protein binding is dependent on RNA structure, providing a context for said proteins to bind (Taliaferro et al., 2016, Lambert et al., 2014). Recent reports have also indicated secondary and tertiary structures of transcripts *in vivo* (Ding et al., 2014, Spitale et al., 2015, Corley et al., 2017). Considering this information, it provides insight to just how complex these spliceosomal component interactions might be.

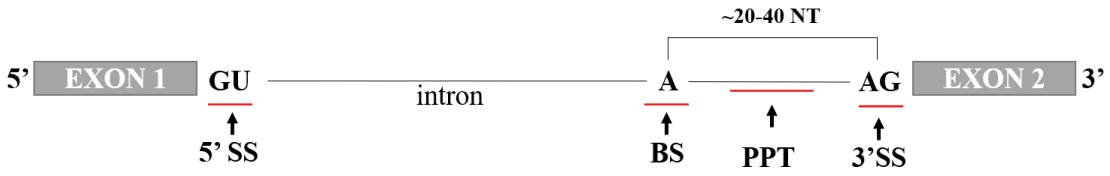

Figure 1.5 Schematic of pre-mRNA substrate indicating the four major splice signals
The schematic presents the four splice signals, 5'SS, BS, PPT, and 3'SS, as well as their relative positions in the pre-mRNA.

It has been shown SR proteins SRSF1 and SRSF2 influence recognition of the splice sites in both constitutively and alternatively spliced pre-mRNAs (Zhou and Fu, 2013). These SR proteins are also known to promote splicing in certain pre-mRNA substrates (Fu, 1993, Zhu and Krainer 2000, Hertel and Maniatis (1999). In alternative splicing, SR protein binding to an alternative exon resulted in exon inclusion, while binding of SR proteins to a flanking exon led to exclusion of the inner alternative exon, therefore demonstrating the effect SR proteins can have in splice site recognition. We focus on identifying a pre-spliceosomal complex, before E-complex, in which all splice signals are recognized by their respective protein factors, as shown in Figure 1.6 (Chen et al., 2016). Our lab has shown that pre-mRNA is bound by multiple copies of SR protein, and to the RNA-SR protein complex, recruitment of U1 snRNP to the 5'SS allows for the recognition of the remaining major splice sites, BS, PPT, and 3'SS. In this early complex all splice signals are recognized, defining the intron, and thus allowing for assembly of the spliceosome on the RNA substrate.

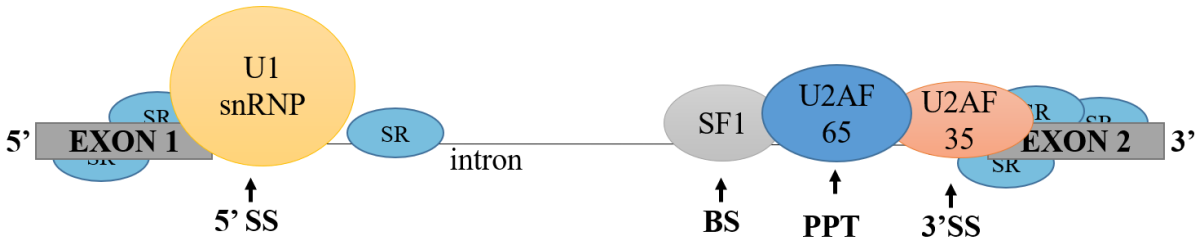

Figure 1.6 Schematic representation of early recognition complex
Within an early spliceosomal complex, leading up to E-complex, the SR protein binds to the pre-mRNA, and the major splice signals are all recognized by protein factors U1 snRNP, SF1, U2AF65, and U2AF35. At this stage the intron is defined and primed for the splicing reaction.

D. Focus of Study

The focus of our lab is elucidating how assembly occurs within earlier stages of spliceosome assembly. To examine the interactions of early splicing factors to the pre-mRNA,

and to each other, I will create splicing substrates through molecular biology, as well as purify proteins known to bind the major splice signals. My goal is to ascertain the relationship between the 5' and 3' splice signals, as well as interrogate the relationship of SR proteins in spliceosome assembly. To study the assembly mechanism of proteins on the pre-mRNA, pull-down assays are required. RNA substrates are tagged with MS2 binding loops, taking advantage of their strong interaction with MS2 viral protein, proteins assembled on the substrate can be pulled down. To determine the influence of the 3' splice signals, BS, PPT, and 3'SS, on U1 snRNP bound to the pre-mRNA, and likewise, if binding of the 3' factors is influenced by U1 snRNP already at the 5'SS, I plan to make mutations about the pre-mRNA substrates. Mutations will be made to alter the splice signals, for example, the PPT sequence, rich in pyrimidines, will be mutated to a purine rich sequence, and the 5'SS will be altered to create a mismatch with the consensus sequence. Through testing wild type and mutant substrates, I plan to uncover possible communication between the 5' and 3' splice signals through examining any changes in protein interactions to the mutant substrates, relative to the wild type. I will also employ pull down experiments of MS2 tagged pre-mRNA to examine efficiencies in U1 snRNP binding and discern differences in recruitment with the addition of SR protein. Urea gels will be used to study bands correlating to both the RNA substrates and U1 snRNA, while SDS-PAGE gels will be used to indicate the presence of protein components in my pull-down experiments. By looking at both RNA and proteins via urea and SDS-PAGE gels, respectively, it will determine the presence and possible interactions between U1 snRNP and other protein factors SF1, U2AF65, U2AF35, and SR protein, to the pre-mRNA substrate.

II. Materials and Methods

A. Protein Purification Protocols

1. His tagged U2AF65

Protein was expressed in *E. Coli*, Rosseta (DE3) cells, grown to A_{600} of 0.8 and induced overnight with isopropyl β-D-1-thiogalactopyranoside (IPTG) at room temperature. The harvested cells were resuspended in 20 mM Hepes pH 7.5, 250 mM ammonium sulfate, 1 M urea, 30 mM imidazole pH 7.5, and 5 mM b-mercaptoethanol (BME), and lysed, with the addition of 1 mM phenylmethanesulfonyl fluoride (PMSF) and 1:1000 protease inhibitor cocktail (PIC). Lysate was clarified through centrifugation at 15,000 rpm for 1 hr. The supernatant was loaded onto a nickel-NTA (GE Healthcare Lifesciences) resin column equilibrated with buffer and later washed with 20 column volumes of the same buffer. The protein was eluted with elution buffer, resuspension buffer with 500 mM imidazole. Elution fractions were checked qualitatively for protein using a quick Bradford assay, mixing one part eluent and one part Bradford dye (Bio-Rad). Through SDS-PAGE elution samples were visualized by Coomassie staining and were pooled. Pooled samples were loaded onto a Superdex 75 (GE Healthcare) size exclusion column, equilibrated with 20 mM Hepes pH 7.5, 250 mM ammonium sulfate, 1M urea, 40 mM imidazole pH 7.5, and 5 mM BME. Peak fractions were concentrated up to 25 μM with PEG (polyethylene gycol) and stored at -20° C with 50% glycerol.

2. His tagged D_3B

Protein, was co-expressed in *E. Coli*, BL21(DE3) pLysS cells, grown to A_{600} of 0.6-0.8 and induced overnight with 0.5 mM IPTG at 16°C. The cells were harvested and resuspended in 20 mM Hepes pH 7.5, 0.5 M NaCl, 0.5 M urea, 50 mM imidazole pH 7.5, 5 mM BME, and lysed with buffer, with the addition of 1 mM PMSF and 1:1000 PIC. Lysate was clarified through

centrifugation at 15,000 rpm for 1 hr. The supernatant was loaded onto a nickel-NTA column pre-equilibrated with buffer and washed with 20 column volumes of the same buffer. The protein was eluted over a linear gradient, 0-100% elution buffer, resuspension buffer with 500 mM imidazole, over 20 column volumes. Through SDS-PAGE elution samples were visualized by Coomassie staining and were pooled. Pooled samples were digested with TEV protease, 1:25 dilution. After digestion, protein was passed through the Ni-NTA column. The sample was then loaded onto Superdex 75 size exclusion column, equilibrated with 20 mM Hepes pH 7.5, 250 mM NaCl, and 5 mM BME. Peak fractions were stored in -80° C, concentrated up to 150 µM.

3. His tagged EFG

EFG, was co-expressed in *E. Coli*, BL21(DE3) cells, overnight without IPTG induction (leaky expression). The cells were harvested and resuspended in 20 mM Hepes pH 7.5, 0.5 M NaCl, 0.5 M urea, 30 mM imidazole pH 7.5, 5 mM BME, and lysed with buffer, with the addition of 1 mM PMSF and 1:1000 PIC. Lysate was clarified through centrifugation at 15,000 rpm for 1 hr. The supernatant was loaded onto a nickel-NTA column equilibrated with buffer and later washed with 20 column volumes of the same buffer. The protein was eluted with elution buffer, same as resuspension buffer but with 500 mM imidazole. Through SDS-PAGE elution samples were visualized by Coomassie staining and were pooled. Pooled samples were digested with TEV protease, 1:10, against a buffer consisting of 20 mM Hepes 7.5, 250 mM NaCl, 50 mM imidazole, 5 mM BME overnight. After digestion, protein was passed through the Ni column equilibrated in the same buffer. The sample was loaded onto a Superdex 75 size exclusion column, equilibrated with 20 mM Hepes 7.5, 250 mM NaCl, and 5 mM BME. Peak fractions were concentrated up to 800 µM and frozen in -80° C.

4. His tagged TEV Protease

Protein was expressed in *E. Coli*, Bl21 (DE3) cells, grown to A_{600} of 0.5 and induced for 6 hours with IPTG at 22°C. The harvested cells were resuspended in 20 mM Tris pH 7.5, 500 mM NaCl, 25 mM imidazole pH 7.5, and 5 mM BME, and lysed, with the addition of 1 mM PMSF and 1 mM benzamidine. Lysate was clarified through centrifugation at 20,000 rpm for 30 min. The supernatant was loaded onto a nickel-NTA column equilibrated with buffer and later washed with 20 column volumes of the same buffer. The protein was eluted over a linear gradient, 0-100% elution buffer, resuspension buffer with 500 mM imidazole, over 15 column volumes. Through SDS-PAGE elution samples were visualized by Coomassie staining and were pooled. Pooled samples were diluted 1:5, with 20 mM MES pH 6.0, 10% glycerol, and 10 mM BME. Diluted protein was passed through SP ion exchange column, pre-equilibrated with 20 mM MES pH 6.0, 10% glycerol, 0.1 M NaCl, and 10 mM BME. SP column is washed with 20 column volumes of buffer, and eluted with high salt, 2 M NaCl, over a 15 column volume gradient (0-100%). Peak fractions were concentrated up to 250 µM and stored at -20°C with 50% glycerol.

B. Baculovirus- U2AF35 (His) Expression and Purification

1. Sf9 Cell Culturing

Sf9 cells were maintained in culture at 1-2 million cells/mL in SF9 III medium (Gibco) supplemented with 1% PSG at 170 rpm and 27° C, at 1/3 flask volume. Cells were passaged every 24 - 48 hours.

2. Protein Expression

Virus previously made by fellow lab members was used for infection. Sf9 cells were infected

at an optimal dilution of the virus, determined previously. The cells were at 1 million cells/mL when infected and left for 72 hours at 27° C, rotating at 170 rpm. The cells were collected and spun down at 3000 rpm for 5 minutes. The supernatant was discarded, and the cell pellet was frozen and kept at -80° C for later nickel affinity purification.

3. Lysis of Sf9 Cells

Cells were lysed in lysis buffer (25mM Tris pH 8.0, 200 mM NaCl, 10 mM Imidazole, 10% Glycerol, 5mM BME, PMSF, PIC). The cells were lysed by sonication at 5 and 50% duty cycle for two to three 1-minute bursts followed by 5 minutes of rest. The lysate was clarified by centrifugation at 4000 rpm at 4° C for 45 minutes.

4. Purification

The pellet is resuspended in lysis buffer, 20mM Tris-HCl pH 7.5, 150mM KCl, 1.5mM MgCl2, 0.2mM EDTA, 1% NP40, 1mM DTT, 0.5 mM PMSF, and PIC 1:100. The cells are sonicated and centrifuged at 8000 rpm for 30 min at 4° C. Supernatant is discarded and an extraction buffer of 20M HEPES pH 7.9, 250mM ammonium sulfate, 8M urea, 5mM BME, and 40mM Imidazole is added. Resuspended pellet is incubated in said buffer with rocking for 30 min. Spun at 15,000 rpm for 1hr, the supernatant is passed through a Ni-NTA column and washed with 20 column volumes of buffer. The protein is eluted with elution buffer, extraction buffer with 600 mM imidazole. Protein is checked through SDS-PAGE and pooled. Samples are then subjected to a refolding buffer, 20mM HEPES pH 7.9, 250mM Ammonium Sulfate, 1M Urea, 5mM BME, and 40mM imidazole overnight. Finally, the protein is concentrated to 30-50 μM, via PEG concentration, and stored at -80°C.

C. Plasmid Cloning

1. PCR Amplification

Reactions consisted of 1x Pfu DNA polymerase buffer (200 mM Tris, 100 mM KCl, 100 mM ammonium sulfate, 20 mM magnesium sulfate, 1% Triton X-100, 1mg/mL BSA), 200 μM dNTP mix, 200 μM of the forward primer, 200 μM of the reverse primer, 25 ng of template, and Pfu DNA polymerase. Thermocycler conditions are as follows: denaturation at 95° C for 5 min (first cycle) or 30 s (repeating cycles), annealing at 2°C below primer Tm values for 30 s, extension at 72° C (1 kilobase/min of expected product), for a total of 34 cycles. Thermo cycling conditions for mutagenesis are as follows: denaturation at 95° C for 30 s, followed by cycles consisting of 95° C for 30 s, 55° C for 1 minute, 68° C (1 minute/kilobase of plasmid length), repeated for 12 cycles. Following mutagenesis and PCR amplification (Bioneer), 1 μl of Dpn I restriction enzyme (10 U/μl, NEB) is added directly to each PCR reaction, incubated at 37° C for 1 hour. Following enzyme incubation, samples are transformed (*E*. coli DH5α strains) and sequenced.

2. Restriction Digestion and Ligation

Digestion reactions are made to 50 uL, consisting of the DNA template (300 ng/μL), 1x CutSmart buffer (New England BioLabs), and appropriate restriction enzymes. The reaction is incubated at 37° C overnight. After digestion, DNA is purified with Bioneer AccuPrep PCR purification kit, following manufacturers protocol and quantified by nanodrop. A 10 μL ligation reaction is prepared with 1 μL of T4 ligase buffer 10x (NEB), 100 ng of vector DNA, insert DNA (added in a 3:1 molar ratio of insert to vector), 1 μL of T4 DNA ligase (NEB), and water added to volume. The reaction is incubated at room temperature for 1 hr. Ligated DNA was transformed into *E*. coli DH5α on LB agarose plates with the appropriate antibiotic. 5 mL of bacterial culture

was grown overnight from a single colony. Plasmid DNA was miniprepped and checked for successful ligations by PCR amplification or test digestion.

D. Refolding pre-mRNA

The pre-mRNA is diluted to 800 mM in concentration, followed by the addition of 50 mM NaCl and 10 mM EDTA. The mix is incubated at 90° C for 2 min, then moved to ice for 5 min, lastly, 10 mM of $MgCl_2$ is added to the RNA.

E. MS2 RNA Pulldowns

1. Preparing RNA and MBP-MS2 Mixtures

MS2 tagged RNA, containing 3 MS2 binding loops, was added to a final concentration of 500 nM with 1,500 nM of MBP-MS2, incubated at 30° C for 15 min.

2. Amylose Resin

250 µL pull-down assays required 7.5 µL of amylose resin (New England Biolabs). 200 µL of water was added to an aliquot of resin, resuspended by tapping and centrifuged at 1000 g for 3 min. Supernatant was removed, adding 200 µL of buffer (20 mM Hepes, 250 mM NaCl, 14mM $MgCl_2$, 5% glycerol, 2.5 mM DTT), resuspended and centrifuged once more at 1000 g for 3 minutes. This washing process was repeated for a total of four times.

3. Pull-down Assay

RNA:MBP-MS2 is added to the reaction buffer, 20 mM Hepes, 250 mM NaCl, 14mM $MgCl_2$, 5% glycerol, 2.5 mM DTT, 500 mM urea and 0.3% PVA (polyvinyl alcohol**)**. U1 snRNP is added,

first incubated with RNA:MBP-MS2 for 5 min at 30° C, and let sit at room temperature for 5 min. Following is the addition of the other protein factors, SRSF1, U2AF35, and U2AF65/SF1 heterodimer. The mixture is incubated for 5 min at 30° C. 7.5 µL of equilibrated amylose resin is added to the mixture and rotated at room temperature for 20 min, immediately followed by 5 min of rotating at 4° C. The resin is centrifuged at 1000 g for 3 min, collecting the supernatant. To the resin, it is washed with 200 µL of buffer, and is repeated two additional times. The samples are then either analyzed by SDS-PAGE or by extracting the RNA and ran on a 6% urea gel.

4. RNA Extraction

Resin is collected with the removal of the supernatant and 200 µL of proteinase K buffer, 10 mM Tris-HCl, 1 mM EDTA, add 0.5% SDS, is added. Proteinase K (NEB) is incubated within each reaction tube for 10 min, rotating at 37°C. The RNA is then phenol extracted.

5. Phenol Extraction and Ethanol Precipitation

A mix of phenol, chloroform, isoamyl alcohol (25:24:1) is added to each reaction in a 1:1 ratio, vigorously mixed by hand and centrifuged at 13,000 rpm for 10 min. To the aqueous layer equal volume of chloroform is added, then mixed and centrifuged at 13,000 rpm for 10 min. To the aqueous layer, a final concentration of 300 mM sodium acetate is added, followed by the addition of 2.5x sample volume of 100% ethanol. The mixture is inverted several times and RNA is precipitated overnight.

F. Data Quantitation

Urea gels were visualized with toluidine blue and scanned. Using Image Lab software (Bio-

Rad) scanned gels were annotated and the bands quantified. In each respective lane within the urea RNA gels, the volume of the band corresponding to U1 snRNA is divided by the volume from the band corresponding to the substrate RNA (AdML or beta-globin), obtaining ratios that can be compared to determine the effect of U1 snRNP binding, in the presence or absence of SRSF1, to the substrates. The values were then normalized for each replicate, normalizing to U1 snRNA ratios to substrate RNA without the presence of SRSF1.

III. Results

A. Strategy to Study Substrate Selection by the Early Spliceosomal Factors

To study the process by which early spliceosomal components select authentic splicing substrates through the recognition of splice sites, our laboratory generated several substrates through molecular biology (Table 3.1) and expressed and purified proteins involved in these early spliceosome assembly stages (Table 3.1B, Figure 3.1). Two model substrates were used for my experiments, a viral substrate AdML (Adenovirus major late) and an endogenous beta-globin (Bg), which are both constitutively spliced. In addition to the full-length substrates, several mutations were made, altering individual splice signals, being reflected in both Bg and AdML. To examine the assembly mechanism, I primarily performed pulldown experiments. A pre-mRNA substrate was mixed with protein or protein complexes, followed by separation of the bound components from the unbound via an affinity pull down of the RNA. To accomplish this, a 3xMS2 binding RNA tag was incorporated at either the 5'- or the 3'-end of the substrates, as shown in the Figure 3.2. MS2 viral coat protein is known to strongly bind a specific RNA stem-loop sequence. We generated MS2 as a fusion protein with maltose binding protein (MBP), which binds to amylose affinity matrix. With the proteins assembled on the RNA substrate, MS2-MBP bound to the RNA then interacts with amylose resin, allowing complexes to be pulled down. These results are then analyzed by either the proteins components via a SDS-PAGE gel, or by RNA via a urea gel.

Table 3.1 Generated substrates and purified proteins
Substrates made through molecular biology (A) and proteins purified by column chromatography (B) are presented.

A

Constructs	
U2AF35, pFastbac	Bg FL WT, MS2 pcDNA 3.1
U2AF35 RRM, pET 24d	Bg FL Δ5'SS, MS2 pcDNA 3.1
U2AF35 UHM, pET 24d	Bg FL Δ PPT, MS2 pcDNA 3.1
SRSF3, pgv67	Bg FL Δ3'SS, MS2 pcDNA 3.1
SRSF2, pET 24d	Bg90 (exon 2), pcDNA 3.1
MBP, pET 30a	Bg90 WT, pcDNA 3.1
AdML, pUC18 with MS2 tag	Bg90 Δ5'SS, MS2 pcDNA 3.1
AdML ΔBS, pUC18	Bg90 Δ PPT, MS2 pcDNA 3.1
AdML Δ3'SS, pUC18	Bg90 ΔBS, MS2 pcDNA 3.1
AdML EH1+2, MS2 pUC18	Bg90 Δ3'SS, MS2 pcDNA 3.1
AdML shortest, pUC18	Bg136, pcDNA 3.1
	Bg Trans, pcDNA 3.1
	Bg shortest, pcDNA 3.1

B

Purified Proteins
U1 snRNP: D3/B + EFG (U1 snRNP assembled by Kaushik Saha)
U2AF35 FL
U2AF65
U2AF65/ SF1 heterodimer
TEV protease

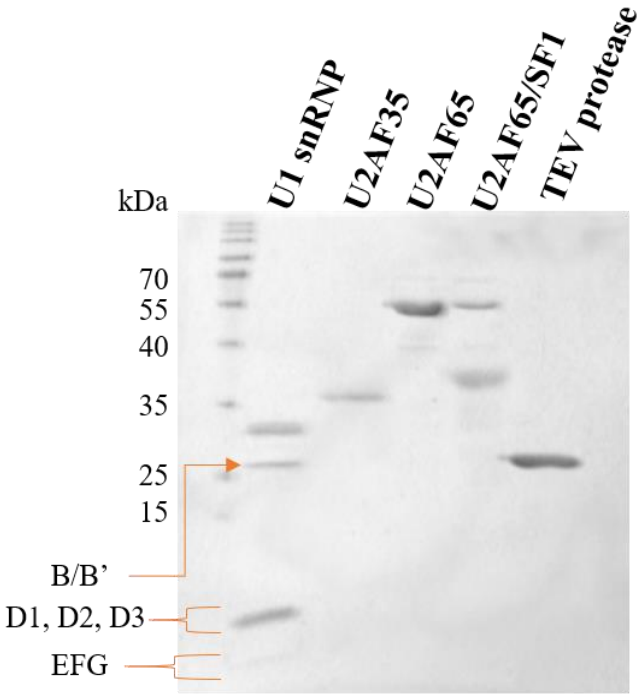

Figure 3.1 Purified proteins on SDS-PAGE gel
SDS-PAGE gel corresponding to Table 3.1B. Proteins were expressed in *E. Coli*, except for U2AF35, expression was done through insect cells.

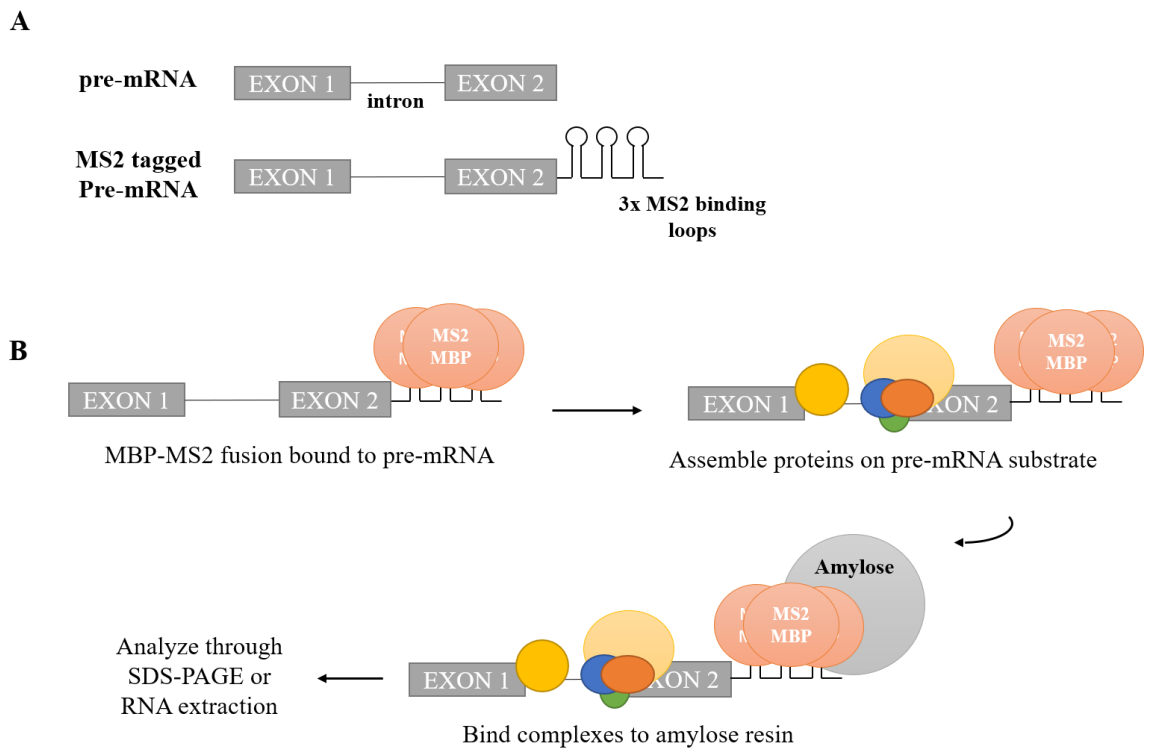

Figure 3.2 Schematic of MS2 RNA tag and MS2 pull-down experiments
Pre-mRNA was tagged with 3x MS2 coat protein binding stem loops at the 5' or 3' end (A). MS2-MBP fusion protein binds to the tagged pre-mRNA, factors bound to the substrate can then be pulled down by MBP's affinity to the amylose resin, allowing analysis of assembled complexes (B).

B. Two Constitutive Splicing Substrates Recruit U1 snRNP to the 5'SS Differently

Our goal was to investigate if SR protein SRSF1 is required for the recruitment of U1 snRNP to the substrate. Both AdML and Bg have been extensively used by the splicing community since the discovery of splicing and are thought to be constitutive splicing substrates based on the strength of the known four splice site signals, as shown in Figure 3.3. Since the 3'-exon of the native Bg substrate is much longer (220 nt) compared to the 3'-exon of AdML (90 nt), a shortened Bg substrate was made. This shortened Bg construct, called Bg90, matched the length of AdML 3'-exon by removing 130 nucleotides from the 3'end, with all four splice signals still intact. *In vivo*

splicing of this substrate has been tested previously in the lab, results showed both Bg90 and WT Bg substrates spliced with similar efficiencies. A phospho-mimetic version of SRSF1 (SRSF1-RE) was used, where 20 serines within the SR domain were converted into glutamates, since it was established that SRSF1 must be phosphorylated for it to be involved in splicing. The substrate RNAs were prepared in two ways, either through unfolding and refolding (1) or those without any treatment (2).

500 nM RNA, of AdML or Bg90, were incubated with 1,500 nM MBP-MS2. To the mixture of MS2 tagged RNA and MBP-MS2, 600 nM of U1 snRNP was added in the presence or absence of 100 nM SRSF1-RE. AdML proved to be efficient at recruiting U1 snRNP to the 5′SS, however the same cannot be said for Bg90. The binding was analyzed in two ways, through a urea gel to visualize both the substrate RNA and U1 snRNA (Figure 3.4 A, B) and by SDS-PAGE to visualize proteins (Figure 3.4 C, D). Visualization of both the U1 snRNP proteins and U1 snRNA will be indicative of U1 snRNP binding to the substrate. SRSF1-RE bound to both substrates as revealed by SDS-PAGE. AdML substrate bound to U1 snRNP to similar extents in the absence or presence of SRSF1-RE. Relative to AdML, however, Bg90 showed to be inefficient in binding free U1 snRNP. In the presence of SRSF1 however, binding improved (Figure 3.4 E, F). U1 snRNP binding was similar irrespective of how the substrates were prepared, suggesting that the pre-mRNA adopts a native structure during or following transcription. These data highlight the differential efficiencies in U1 snRNP recruitment, suggesting a difference in 5′SS recognition by our model substrates, as well as a possible stabilizing role of SRSF1 in U1 snRNP interaction with the RNA in the case of beta-globin.

	5’ splice site	Branch site	Poly-pyrimidine tract	3’ splice site
Consensus	GURAGU	YNCURAC	Rich in pyrimidines	YAG
AdML substrate	GUGAGU	UGCUGAC	UGUCCCUUUUUUUU	CAG
βg substrate	GUUGGU	CACUGAC	CUAUUUUCCCACCC	UAG

Figure 3.3 Schematic aligning substrates AdML and Bg with consensus sequences
The consensus sequences for the four major splice signals are aligned with those of model substrates AdML and Bg. Due to the strong correlation to the consensus, it is believed these two substrates are constitutively spliced. (Y= pyrimidine, N= any nucleotide, and R= purine).

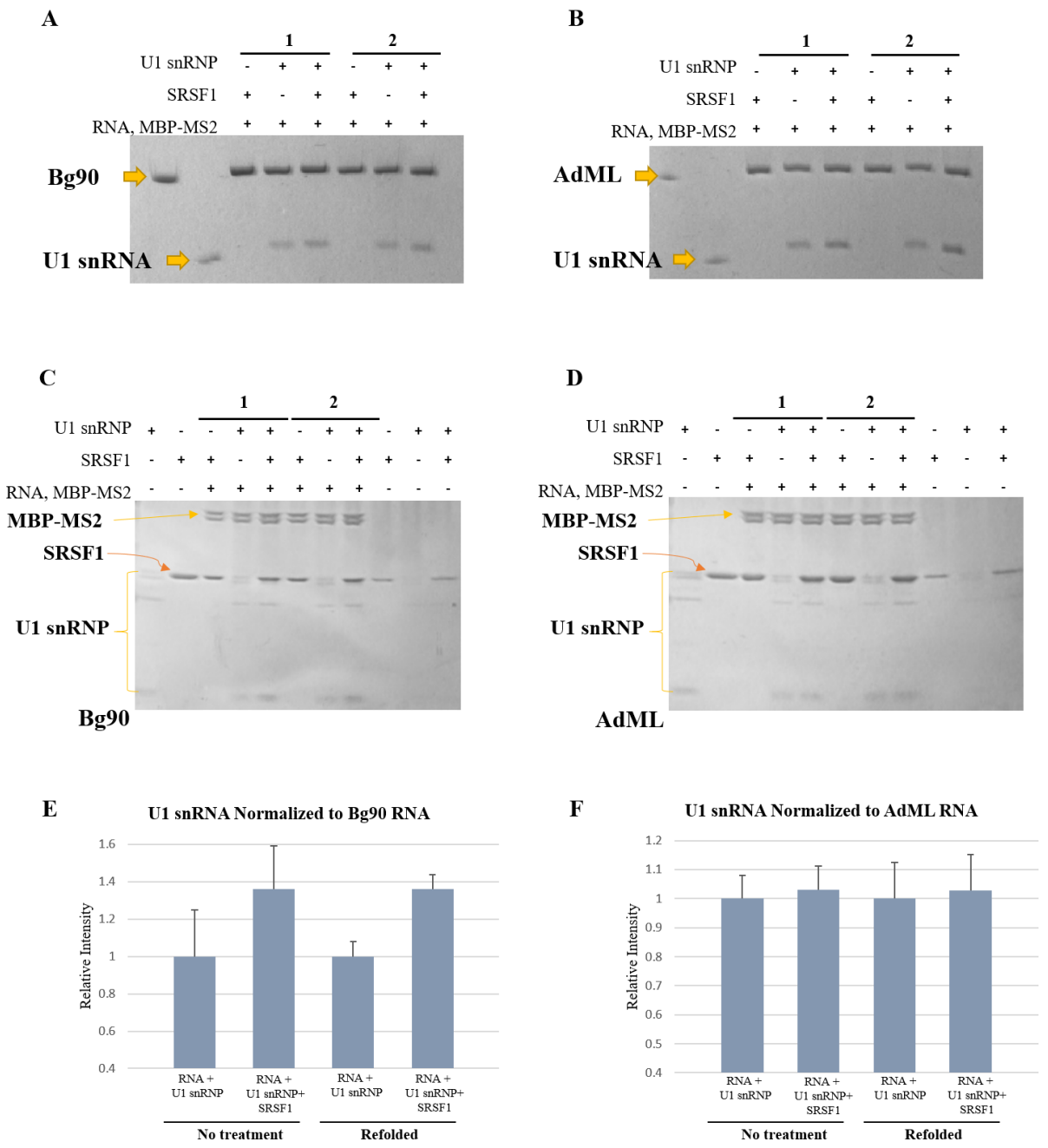

Figure 3.4 Differential U1 snRNP recruitment by AdML and Bg90
To refolded RNA (1) and the RNA without treatment (2), U1 snRNP is added with or without SRSF1-RE, demonstrating the difference in recruitment of the snRNP. AdML (A) proved efficient in binding U1 snRNP, irrespective of SR protein addition, while Bg (B) showed an enhanced effect when SR protein was present. C and D show the corresponding protein gels for A and B, respectively, while E and F quantitate data from A and B, respectively.

C. BS - PPT - 3'SS (3' splice signals) Influences U1 snRNP Recruitment at the 5'SS

In addition to the wild type (WT) substrates, mutants were made to better understand the importance of the 3' splice signals in the early recruitment of the spliceosome components. Specifically, I wanted to observe whether U1 snRNP binding to the 5'SS is influenced by the 3' splice signals, and conversely, if the 3' factors need U1 snRNP already at the 5'SS to bind. Two separate mutations were made, several conserved residues at the 5'SS (Figure 3.5A) were mutated, which we called Δ5'SS. Similarly, pyrimidine residues at the PPT were mutated to purine residues to generate ΔPPT (Figure 3.5B). I repeated the pull-down assay of AdML WT along with the two AdML mutants, Δ5'SS and ΔPPT (Figure 3.6). As expected with the Δ5'SS mutant, it was deficient in bringing U1 snRNP. With the ΔPPT mutant however, an unexpected result was recorded; U1 snRNP recruitment was also impaired. This suggests that even within the early steps of spliceosome assembly, U1 snRNP was communicating with the other splice signals, not just the 5'SS. It is not clear at this stage if this communication is a consequence of direct contact with the PPT sequence or if indirectly, the PPT provides an essential structural feature for U1 snRNP binding to occur. The enhancement effect of SRSF1-RE in U1 snRNP recruitment found in Bg90 is visible here as well in both AdML mutants, as the interaction is slightly strengthened due to the addition of the SR protein to the mutant substrates, suggesting that SRSF1 can partially compensate for the observed defect in U1 snRNP recruitment, hinting at an indirect role of the PPT in U1 snRNP recruitment at the 5'SS.

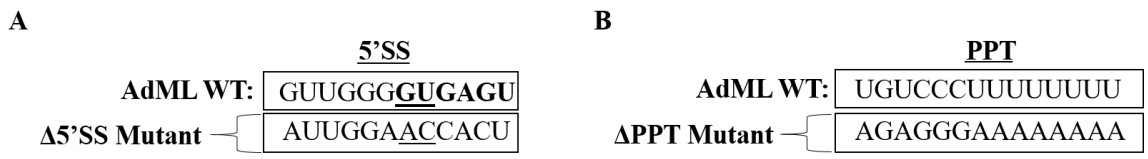

Figure 3.5 Schematic of AdML wild type sequences relative to 5'SS and PPT mutants
In addition to the wild type AdML substrate, two mutants were created. The 5'SS mutant (A) alters the 5'SS, bolded with GU dinucleotide underlined, and other surrounding residues. The PPT mutant (B) changes the stretch of polypyrimidines to purines.

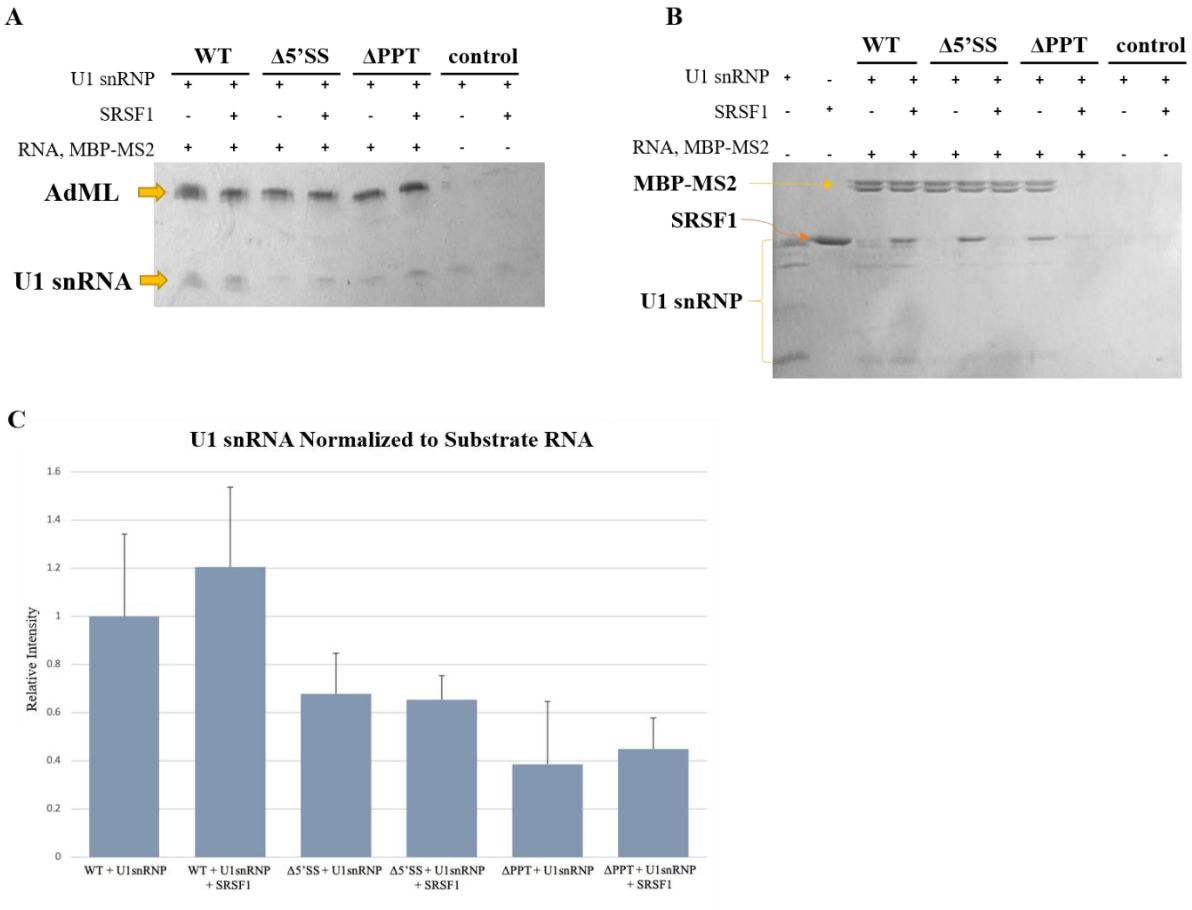

Figure 3.6 AdML WT vs 5'SS and PPT mutants in U1 snRNP recruitment
Pre-mRNA was incubated with U1 snRNP, or U1 snRNP plus SRSF1-RE. Urea gel (A) of AdML wild type, Δ5'SS, ΔPPT, and controls, examine differences in recruitment of U1 snRNP, shown with the corresponding protein gel (B). The data from A was quantified in C.

D. Split RNA Substrate to Study the Recruitment of SF1, U2AF65 and U2AF35 to the 3' Splice Signals

To study the recruitment of the 3' proteins (3' p), SF1, U2AF65, and U2AF35, in more depth, we decided to create a split substrate. By splitting the substrate into discernable halves, we can better understand the cross talk between the 5' and 3' signals. Two RNAs were made, a 5' RNA (Bg 136) containing the 5'SS and the flanking exon and intron, and a corresponding 3' RNA (trans Bg) that houses the 3' splice signals and part of the 3' exon (Figure 3.7). If 3'p bound to trans Bg interact with U1 snRNP and/or SRSF1 bound to Bg136, trans Bg will be pulled down along with Bg136. I added MS2 binding stem loops to the 5'-end of Bg136 for the pulldown experiments.

50 nM of the MS2 tagged Bg136 with 1,500 nM MBP-MS2 was mixed with 50 nM trans Bg in a series of conditions, as shown in Figure 3.8. We also used tRNA instead of trans Bg to test binding specificity. First, I tested if the two pairs of RNA interact with each other in the absence of any additional other factor, (lanes 1 and 2, Figure 3.8). As expected Bg136 failed to pull down 3' RNAs, trans Bg and the control tRNA. I next tested interaction between the RNA in the presence of SRSF1 (lanes 3 and 4, Figure 3.8). Surprisingly, I found Bg136 interacts with trans Bg in the presence SRSF1-RE, and considering Bg136 does not pull down tRNA, this suggests SRSF1 mediated interaction between Bg136 and trans Bg is specific. Through these experiments I hoped to see only in the presence of trans Bg, and not tRNA, more efficient recruitment of the 3' proteins. In the RNA gel however, although tRNA is not present in any of the pulldowns, showing specificity for the 3' Bg RNA, the amount of U1 snRNP is noticeably less when combined with SRSF1 and the 3'p, illustrated by lanes 7 and 9 in Figure 3.8. These results indicate some sort of competition, as less U1 snRNA is pulled down with the addition of 3' p.

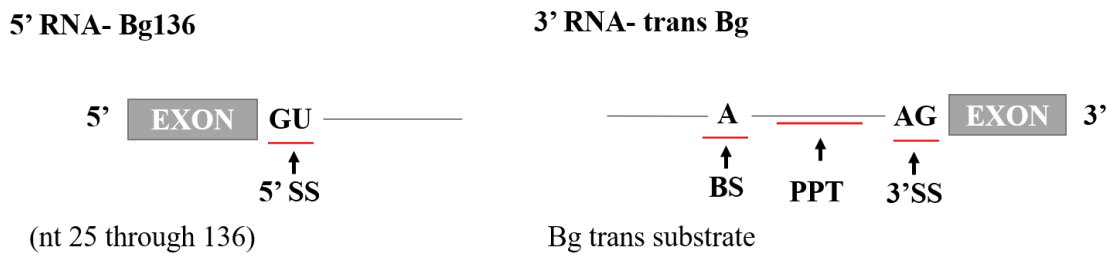

Figure 3.7 Schematic of a split pre-mRNA substrate
Two RNAs were made to mimic a full substrate, consisting of a 5' segment with the 5'SS and parts of the flanking exon and intron (Bg136), and a 3' RNA containing the 3' splice signals, BS, PPT, and 3'SS, as well as part of the 3' exon (trans Bg).

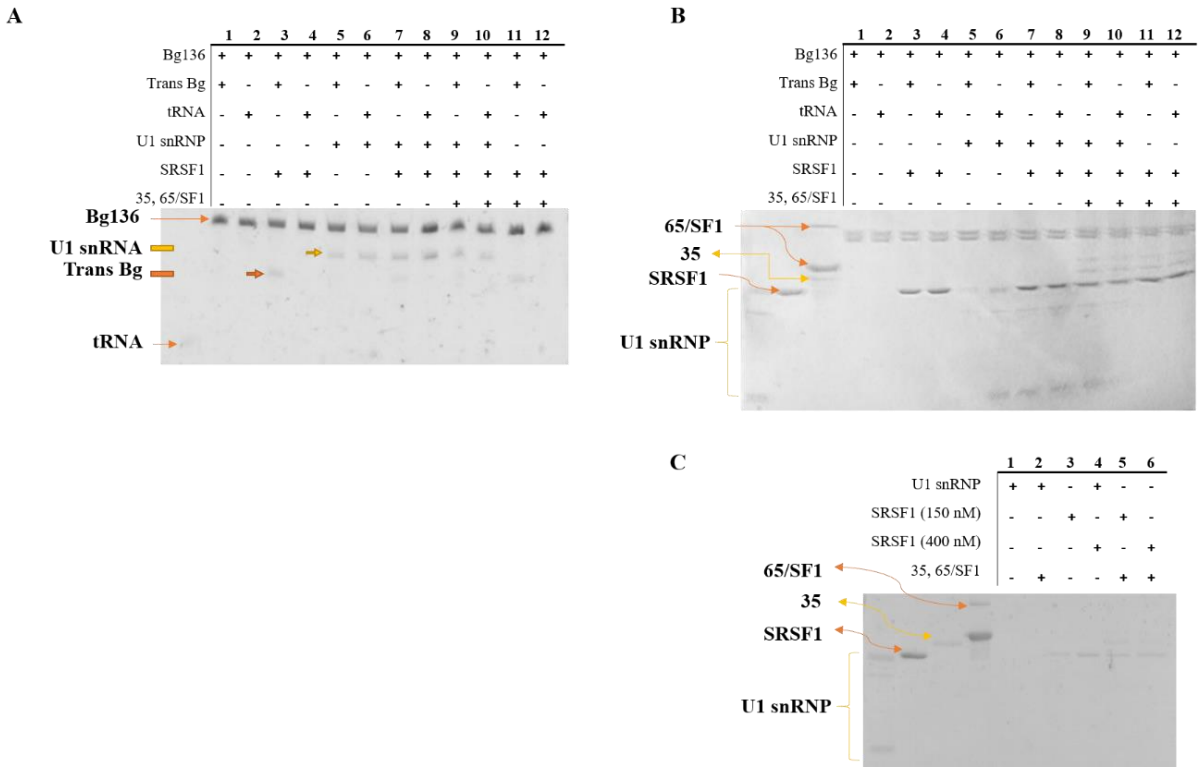

Figure 3.8 Split substrate in the presence of early splicing factors
Bg136 was added with trans Bg or a nonspecific tRNA and examined the RNA via a urea gel (A) and SDS PAGE gel (B), with accompanying controls (C). Different combinations of U1 snRNP, SRSF1-RE, and U2AF35 with U2AF65/SF1 heterodimer were added to 5' and 3' RNAs, showing bridging of only trans Bg and not tRNA, in certain conditions.

There are three variants of SRSF1 used in our experiments. To further test the bridging capabilities of SRSF1, as well as U1 snRNP recruitment efficiencies, these SRSF1 variants were employed. SRSF1-RBD is comprised of the RNA binding domain and proven inefficient in splicing. SRSF1-RS in addition to the RBD, possesses a C-terminal RS domain (rich in arginine and serine) which does not induce splicing, and SRSF1-RE with arginine glutamate repeats serving as a hyperphosphorylated mimic, with previous studies showing its importance in splicing. I tested each SRSF1 protein with both Bg136 and either trans Bg or control tRNA to see if the SR protein alone could bridge two RNAs. Results showed that without SR protein, only Bg136 is pulled down, however varying levels of trans Bg can be seen when SR protein is added (lanes 1 and 5, Figure 3.9). With SRSF1-RE and RS, trans Bg shows a strong interaction with the 5' RNA, while SRSF1-RBD proved inefficient to connect the two RNA. From these data, it shows that SRSF1 alone can bridge these two pieces of RNA, while showing SRSF1's specificity for the 3' signal containing trans Bg and not tRNA. Interestingly, despite SRSF1-RS' inability to induce splicing, more of the trans Bg is pulled down relative to SRSF1-RE, which is known to be involved in splicing.

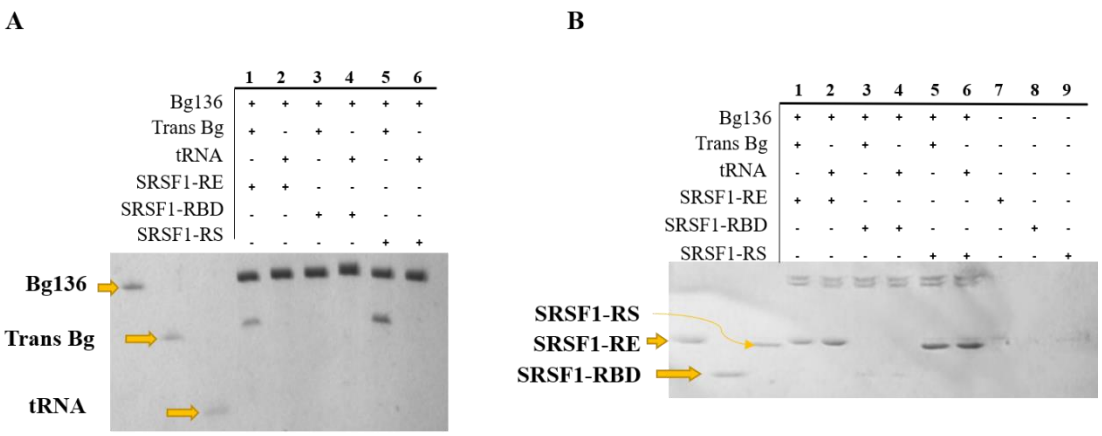

Figure 3.9 SRSF1 bridging the 5' and 3' RNAs
The bridging capability of the different SRSF1 variants were tested, using Bg136 with either trans Bg or tRNA. It is observed that only trans Bg is being connected to the Bg136 (A), however SRSF1-RS brings the 3' RNA more efficiently than SRSF1-RE. B) Shows.

IV. Discussion

My goal was to study the metazoan cis-acting splice signals responsible for defining an intron during the assembly of the spliceosome. Due to its prevalence in diseases, as about a third of genetic disorders are linked to splicing, I was interested in elucidating how the early spliceosomal components select legitimate splicing substrates through splice site recognition, despite degenerate nucleotide sequences (Singh & Cooper, 2012, Fredericks et al., 2015). To that end, a strategy was devised. First, substrates were made through molecular biology and the protein factors known to bind said splice signals were purified. With the necessary components assembled, in order to examine the assembly mechanism of proteins on the RNA, pull down experiments were performed. Through this approach, protein assemblies on constitutive pre-mRNA substrates AdML and Bg, can be analyzed either through SDS-PAGE, examining the protein components, or by urea gels to visualize RNA.

Focus was placed on the SR family of proteins, namely SRSF1 for being known to be involved in both constitutive and alternative splicing, and their influence on splice site recognition. With the role of SR proteins unclear, our focus is on the activity of these proteins and how they are involved in assembly. To examine if SRSF1 is needed to recruit U1 snRNP to the substrate, pulldowns of MS2 tagged AdML and Bg90 (shortened to closely match the length of AdML's 3'-exon) substrates were done in parallel. Both substrates were mixed with U1 snRNP, with or without phospho-mimetic SRSF1-RE, proven important in splicing. Notably, there were differences between the two RNAs in terms of U1 snRNP interaction (Figure 3.4). AdML demonstrated efficient binding of U1 snRNP to the 5′SS, although Bg90 did not exhibit this same feature. AdML substrate bound to U1 snRNP to similar extents in the absence or presence of SRSF1-RE. Unlike AdML however, Bg90 proved inefficient in that regard, while in the presence of SRSF1, binding of U1 snRNP improved. These results suggest differential recognition of the 5′SS by the model

substrates, in addition to possible stabilization of the U1 snRNP:RNA complex mediated by SRSF1, but only in the case of beta-globin.

I then wanted to observe the relationship between the 5'SS and 3' splice signals. With studies indicating the 5'SS and 3'SS being in close proximity was necessary for E-complex, I wanted to see if U1 snRNP bound to the substrate is influenced by the 3' splice signals, and similarly, if the 3' factors are influenced by U1 snRNP already at the 5'SS (Michaud & Reed, 1993, Kent & MacMillan, 2002). Thus, additional substrates were made, consisting of mutations to alter the splice signals, known to be a close match to the consensus sequences. Along with AdML WT, mutants Δ5'SS and ΔPTT were made to better discern the importance of the 3' splice signals in early recruitment of the spliceosome components. Unexpectedly, Δ5'ss was found deficient in U1 snRNP recruitment, the ΔPPT mutant substrate however showed interesting results, as it also displayed impairment in U1 snRNP interaction (Figure 3.6). This suggests a line of communication between the splice signals, besides just 5'SS, and U1 snRNP. The nature of this communication remains elusive however, it us unknown whether direct contact to the PPT sequence is required or whether it acts via an indirect method, providing a sort of essential context for U1 snRNP recruitment, thereby requiring further study in elucidating the role of PPT in U1 snRNP recruitment to the 5'SS.

To increase the breadth of our study in regards to SF1, U2AF65, and U2AF35 recruitment to the 3' splice signals, along with the role of U1 snRNP bound to the 5'SS in this process, a split substrate was created. By having two separate RNAs, a 5' RNA segment (Bg136) containing the 5'SS and parts of the flanking intron and exon, and a 3' portion (trans Bg) that contains the BS, PPT, and 3'SS signals, this allowed for closer inspection of any connection between the 5' and 3' signals (Figure 3.8). Bg136 was shown to specifically interact with 3′ RNA trans Bg and not the

control tRNA, but only in the presence of SRSF1-RE. However, despite this specific SRSF1 mediated interaction between Bg136 and trans Bg, although tRNA is not present in any of the pulldowns, the amount of U1 snRNP is noticeably less when SRSF1 is combined with the 3' proteins (3'p), relative to the amount of U1 snRNP and SRSF1 without the 3'p.These results indicate some sort of competition, as less U1 snRNA is pulled down with the addition of 3' p, prompting further experiments to better understand this observation. Experiments with the split substrate will be repeated but with a 3' trans Bg ΔPPT mutant, this will examine the effect of U1 snRNP recruitment at the 5' RNA segment paired with a 3' RNA lacking a PPT sequence, in hopes of better understanding the role of the PPT in the context of its relationship with 5'SS and U1 snRNP, as well as help reinforce earlier observations in deficient U1 snRNP recruitment of AdML ΔPTT.

I tested each SRSF1 protein variant (SRSF1-RE, SRSF1-RBD, and SRSF1-RS) with both Bg136 with either trans Bg or tRNA to see if the SR protein alone could bridge two RNAs. With SRSF1-RE and RS, trans Bg showed better interaction with the 5' RNA, relative to SRSF1-RBD which proved incapable in connecting the two RNA (Figure 3.9). These results suggest SRSF1 is sufficient in connecting both 5' and 3' RNAs, while showing specificity for the 3' signal containing trans Bg and not tRNA. Following up on these results, experiments will be done once more with all three SRSF1 proteins but now testing and comparing U1 snRNP interaction efficiencies among the SRSF1 variants. We would expect SRSF1-RE, known for its involvement in splicing, to be more efficient in recruiting U1 snRNP despite SRS1-RS showing slightly enhanced levels of bridging the opposite ends of the split RNA substrate.

REFERENCES

Alsafadi, S., Houy, A., Battistella, A., Popova, T., Wassef, M., Henry, E., & Stern, M. Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. *Nature Communications* **7** (2016). doi:10.1038/ncomms10615

Ast, G. How did alternative splicing evolve? *Nature Reviews Genetics* **5**, 773–782 (2004). doi:10.1038/nrg1451

Chandler, S.D., Mayeda, A., Yeakley, J.M., Krainer, A.R., & Fu, X.D. RNA splicing specificity determined bythe coordinated action of RNA recognition motifs in SR proteins. *Proceedings of the National Academy of Sciences* **94**, 3596–3601 (1997). doi:10.1073/pnas.94.8.3596

Chen, L., Weinmeister, R., Kralovicova, J., Eperon, L.P., Vorechovsky, I., Hudson, A.J., & Eperon, I.C. Stoichiometries of U2AF35, U2AF65 and U2 snRNP reveal new early spliceosome assembly pathways. *Nucleic Acids Research* **94**, 3596–3601 (1997). doi:10.1073/pnas.94.8.3596

Cho, S., Hoang, A., Sinha, R., Zhong, X.Y., Fu, X.D., Krainer, A.R., & Ghosh, G. Interaction between the RNA Binding Domains of Ser-Arg Splicing Factor 1 and U1-70K SnRNP Protein Determines Early Spliceosome Assembly. *Proceedings of the National Academy of Sciences* **108**, 8233–8238 (2011). doi:10.1073/pnas.1017700108

Colwill, K., Pawson, T., Andrews, B., Prasad, J., Manley, J.L., Bell, J.C. & Duncan, P.I. The Clk/Sty protein kinase phosphorylates SR splicing factors and regulates their intranuclear distribution. *The EMBO Journal* **15**, 265–275 (1996). doi:10.1002/j.1460-2075.1996.tb00357.x

Corley, M., Solem, A., Phillips, G., Lackley, L., Ziehr, B., Vincent, H.A., Mustoe, A.M., Ramos, S.B.V., Weeks, K.M., Moorman, N.J., & Laederach, A. An RNA structure-mediated, posttranscriptional model of human α-1-antitrypsin expression. *Proceedings of the National Academy of Sciences* **114** (2017). doi:10.1073/pnas.1706539114

Ding, Y., Tang, Y., Kwok, C., Kit, Z., Yu, B., Philip, C., & Assmann, S. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696–700 (2013). doi:10.1038/nature12756

Fredericks, A.M., Cygan, K.J., Brown, B.A., & Fairbrother, W.G. RNA-Binding Proteins: Splicing Factors and Disease. *Biomolecules* **5**, 893–909 (2015). doi:10.3390/biom5020893

Fu, X.D. Specific commitment of different pre-mRNAs to splicing by single SR proteins. *Nature* **365**, 82–85 (1993). doi:10.1038/365082a0

Gravely, B.R. Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics* **17**, 100–107 (2001). doi:10.1016/s0168-9525(00)02176-4

Gui, J.F., Lane, W.S. & Fu, X.D. A serine kinase regulates intracellular localization of splicing factors in the cell cycle. *Nature* **369**, 678–682 (1994). doi:10.1038/369678a0

Hertel, K.J. & Maniatis, T. Serine–arginine (SR)-rich splicing factors have an exon-independent function in pre-mRNA splicing. *Proceedings of the National Academy of Sciences* **96**, 2651–2655 (1999). doi:10.1073/pnas.96.6.2651

Kent OA & MacMillan AM. Early organization of pre-mRNA during spliceosome assembly. Nature Structural Biology **9**, 576–581 (2002). doi: 10.1038/nsb822

Kondo, Y., Oubridge, C., Roon, A.M., & Nagai, K. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *ELife* **4** (2015). doi:10.7554/elife.04986

Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A., & Burge, C.B. RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Molecular Cell* **54**, 887–900 (2014). doi:10.1016/j.molcel.2014.04.016

Lu, Z., Zhang, Q.C., Lee, B., Flynn, R.A., Smith, M.A., Robinson, J.T., Davidovich, C., Gooding, A.R., Goodrich, K.J., Mattick, J.S., Mesirov, J.P., Cech, T.R., & Chang, H.Y. RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell* **165**, 1267–1279, (2016). doi:10.1016/j.cell.2016.04.028

Makarov, E.M., Owen, N., Bottrill, A., & Makarova, O.V. Functional mammalian spliceosomal complex E contains SMN complex proteins in addition to U1 and U2 snRNPs. *Nucleic Acids Research* **40**, 2639–2652 (2011). doi:10.1093/nar/gkr1056

Michaud, S. & Reed, R. A functional association between the 5' and 3' splice site is established in the earliest prespliceosome complex (E) in mammals. *Genes & Development* **7**, 1008–1020 (1993). doi:10.1101/gad.7.6.1008

Modrek, B. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research* **29**, 2850–2859 (2001). doi:10.1093/nar/29.13.2850

Mount, S.M., Pettersson, I., Hinterberger, M., Karmas, A., & Steitz, J.A. The U1 small nuclear RNA-protein complex selectively binds a 5' splice site in vitro. *Cell* **33**, 509–518 (1983). doi:10.1016/0092-8674(83)90432-4

Pandit, S., Zhou, Y., Shiue, L., Coutinho-Mansfield, G., Li, H., Qiu, J., Huang, J., Yeo, G.W., Ares, M. Jr, & Fu X.D. Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Molecular Cell* **50**, 223–235 (2013). doi:10.1016/j.molcel.2013.03.001

Roca, X., Krainer, A.R., & Eperon, I.C. Pick one, but be quick: 5′ splice sites and the problems of too many choices. *Genes & Development* **27**, 129–144 (2013). doi:10.1101/gad.209759.112

Sapra, A.K., Anko, M.L., Grishina, I., Lorenz, M., Pabis, M., Poser, I., Rollins, J., Weiland, E.M., & Neugebauer, K.M. SR protein family members display diverse activities in the formation of nascent and mature mRNPs in vivo. *Molecular Cell* **34**, 179–190 (2009). doi:10.1016/j.molcel.2009.02.031

Singh, R.K. & Cooper, T.A. Pre-mRNA splicing in disease and therapeutics. *Trends in Molecular Medicine* **18**, 472–482 (2012). doi:10.1016/j.molmed.2012.06.006.

Smith, C.W., & Valcárcel, J. Alternative Pre-MRNA Splicing: the Logic of Combinatorial Control. *Trends in Biochemical Sciences* **25**, 381–388 (2000). doi:10.1016/s09680004(00)01604-2

Spitale, R.C., Flynn, R.A., Zhan, Q., Crisalli, P., Lee, B., Jung, J.W., Kuchelmeister, H.Y., Batista, P.J., Torre, E.A., Kool, E.T., & Chang, H.Y. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **527**, 264–264 (2015). doi:10.1038/nature15717

Taliaferro, J.M., Labert, N.J., Sudmant, P.H., Dominguez, D., Merkin, J.J., Alexis, M.S., Bazile, C., & Burge, C.B. RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation. *Molecular Cell* **64**, 294–306, (2016). doi:10.1016/j.molcel. 2016.08.035

Wang, J., Smith, P.J., Krainer, A.R., & Zhang, M.Q. Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. *Nucleic Acids Research* **33**, 5053–5062 (2005). doi:10.1093/nar/gki810

Zhou, Z. & Fu, X.D. Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma* **122**, 191–207 (2013). doi:10.1007/s00412-013-0407-z

Zhu, J. & Krainer, A.R. Pre-mRNA splicing in the absence of an SR protein RS domain. *Genes & Development* **14**, 3166–3178 (2000). doi:10.1101/gad.189500

Zhu, J., Mayeda, A., & Krainer, A.R. Exon Identity Established through Differential Antagonism between Exonic Splicing Silencer-Bound hnRNP A1 and Enhancer-Bound SR Proteins. *Molecular Cell* **8**, 1351–1361 (2001). doi:10.1016/s1097-2765(01)00409-9

Zillmann, M., Rose, S.D., & Berget, S.M. U1 small nuclear ribonucleoproteins are required early during spliceosome assembly. *Molecular and Cellular Biology* **7**, 2877–2883 (1987). doi:10.1128/mcb.7.8.2877