Deep sequencing of pre-translational mRNPs reveals hidden flux through evolutionarily conserved AS-NMD pathways

A Dissertation Presented By

Carrie A. Kovalak

Submitted to the Faculty of the University of Massachusetts Graduate School of Biomedical Sciences, Worcester in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

January 06, 2020

Program in Biochemistry and Molecular Pharmacology

Deep sequencing of pre-translational mRNPs reveals hidden flux through evolutionarily conserved AS-NMD pathways

A Dissertation Presented By Carrie A. Kovalak

This work was undertaken in the Graduate School of Biomedical Sciences

Program in Biochemistry and Molecular Pharmacology

Under the mentorship of

Melissa J. Moore PhD, Thesis Advisor

Andrei Korostelev PhD, Member of Committee

Athma A. Pai PhD, Member of Committee

Sean P. Ryder PhD, Member of Committee

William G. Fairbrother PhD, External Member of Committee

Nick Rhind PhD, Chair of Committee

Mary Ellen Lane PhD, Dean of the Graduate School of Biomedical Sciences

January 06, 2020

Acknowledgements

First and foremost, I would like to thank my parents, Deborah and Robert, for nurturing and supporting my love of science for as long as I can remember. Even when I have doubted myself, their unwavering belief that I would achieve my dreams someday has motivated me to keep pushing ahead. Though my mother cannot be here to see the dream finally become reality, I know she would be front row and beaming with pride if she could.

I also owe an incredible amount of gratitude to my thesis advisor, Melissa Moore. My journey through grad school has been far from typical, and she has continuously done whatever she possibly could to help me succeed. Words really can not do justice to express how thankful I am for her support and guidance.

Many thanks also to my committee - Nick Rhind, Sean Ryder, Andrei Korostelev, and Athma Pai - for pushing me to think critically and teaching me how to be a better researcher (even when I resisted).

When I walked into Melissa's office in 2012 I was searching for a lab to study RNA, but I realized years later that what I had found instead was an extended family. The Moore lab became my second home and I will always cherish the lab parties, our late night venting sessions, and so much more. Special thanks to Joerg, Andrew, Eric, Ami, Kelly, Weijun, Mihir, Harleen, Erin, Alicia, Laureen, Lingtao, Makoto, and Amrit for everything over the years.

I would not have made it through without leaning on my peers at UMass. To Beth, Cansu, Will, Brian, Pak, Carolyn, Jon, Livio, and many more: thank you for always being there. You turned this journey into something special.

And lastly, to everyone else that has helped shape me into the person I am today. To Kate and Vinny for opening my eyes to the wonders of research. To Jeff,

Kristian, TJ, and the rest of the Coller/Baker labs for giving me a chance and making me realize that grad school was not beyond me. To Ashley for teaching me to not believe everything I think. To Jesse for keeping me smiling even on the worst days. And finally, to Kelsey, Todd, Jonathan, Rob, John, Lauren, and Josh for always seeing the best in me.

Abstract

Deep sequencing of mRNAs (RNA-Seq) is now the preferred method for transcriptome-wide quantification of gene expression. Yet many mRNA isoforms, such as those eliminated by nonsense-mediated decay (NMD), are inherently unstable. Thus a significant drawback of steady-state RNA-Seq is that it provides marginal information on the flux through alternative splicing pathways. Measurement of such flux necessitates capture of newly made species prior to mRNA decay. One means to capture nascent mRNAs is affinity purifying either the exon junction complex (EJC) or activated spliceosomes. Late-stage spliceosomes deposit the EJC upstream of exon-exon junctions, where it remains associated until the first round of translation. As most mRNA decay pathways are translation-dependent, these EJC- or spliceosome-associated, pre-translational mRNAs should provide an accurate record of the initial population of alternate mRNA isoforms.

Previous work has analyzed the protein composition and structure of pre-translational mRNPs in detail. While in the Moore lab, my project has focused on exploring the diversity of mRNA isoforms contained within these complexes. As expected, known NMD isoforms are more highly represented in pre-translational mRNPs than in RNA-Seq libraries. To investigate whether pre-translational mRNPs contain novel mRNA isoforms, we created a bioinformatics pipeline that identified thousands of previously unannotated splicing events. Though many can be attributed to "splicing noise", others are evolutionarily-conserved events that produce new AS-NMD isoforms likely involved in maintenance of protein homeostasis. Several of these occur in genes whose overexpression has been linked to poor cancer prognosis.

Table of Contents

List of figures

Chapter 1

Chapter 2

Previously published figures

License to republish included, if required.

Figure	Publisher	License
1.1	Elsevier	4753710197872
1.2	National Academy of Sciences	
1.3	Elsevier	4753710197872
1.4	Elsevier	4753701098578
1.5	Elsevier	4753701098578
1.6	National Academy of Sciences	
1.7	Wiley	4753730547391
1.8	Wiley	4753720921126
1.9	Elsevier	4753730716033
1.10	Nature Publishing Group	4753720115182
1.11	Nature Publishing Group	4753720278029
1.12	Nature Publishing Group	4753720278029
1.13	Annual Reviews	1014632-1
1.14	Nature Publishing Group	4753720278029
1.16	BioMed Central	
1.17	Oxford University Press	
1.18	Elsevier	4753720986535
1.19	Elsevier	4753701299251
1.20	Elsevier	4753711495698
1.21	Elsevier	4753711495698
3.1	Elsevier	4753710197872
3.2	Elsevier	4753701440974
3.3	Elsevier	4753701440974
3.4	Elsevier	4753701299251

List of tables

Abbreviations

APA	**A**lternative **P**oly**a**denylation
AS	**A**lternative **S**plicing
AS-NMD	**A**lternative **S**plicing coupled to **NMD**
cEJC	**c**anonical **EJC**
CHESS	**C**omprehensive **H**uman **E**xpressed **S**equence**S**
CLIP	**C**ross**L**inking and **I**mmuno**P**recipitation
EJC	**E**xon **J**unction **C**omplex
ER	**E**ndoplasmic **R**eticulum
GENCODE	**G**enome research at **ENC**yclopedia **o**f **D**NA **E**lements
GINI	**G**ene **I**dentification by **N**MD **I**nhibition
GO	**G**ene **O**ntology
GTEx	**G**enotype-**T**issue **Ex**pression
HAVANA	**H**uman **a**nd **V**ertebrate **An**alysis and **A**nnotation
hnRNP	**h**eterogeneous **n**uclear **R**ibo**N**ucleo**P**rotein
INSDC	**I**nternational **N**ucleotide **S**equence **D**atabase **C**onsortium
IP	**I**mmuno**P**recipitation
lncRNA	**l**ong **n**on-**c**oding **RNA**
mRNP	**m**essenger **RN**A **P**articles
NCBI	**N**ational **C**enter for **B**iotechnology **I**nformation
nEJC	**n**oncanonical **EJC**

NGD	**N**o-**G**o **D**ecay
NMD	**N**onsense-**M**ediated **D**ecay
NSD	**N**on-**S**top **D**ecay
ORF	**O**pen **R**eading **F**rame
PABP	**P**oly(**A**)-**B**inding **P**rotein
PAS	**P**oly**a**denylation **S**ite
pre-mRNA	**pre**mature-**m**essenger **RNA**
PTC	**P**remature **T**ermination **C**odon
RefSeq	**Ref**erence **Seq**uence
RBP	**R**NA **B**inding **P**rotein
RIP	**R**NA **I**mmuno**P**recipitation
RIPiT	**R**NA:protein **I**mmuno**P**recipitation **i**n **T**andem
RIPPLiT	**R**NA **I**mmuno**P**recipitation and **P**roximity **L**igation **i**n **T**andem
RNA-Seq	**RNA-Seq**uencing
siRNA	**s**hort **i**nterfering **RNA**
snRNP	**s**mall **n**uclear **R**ibo**N**ucleo**P**rotein
SRA	**S**hort **R**ead **A**rchive
SS	**S**plice **S**ite
TDD	**T**ranslation-**D**ependent **D**ecay
TSS	**T**ranscription **S**tart **S**ite
uROF	**u**pstream **ORF**
UTR	**Unt**ranslated **R**egion
-1 PRF	(-1) **P**rogrammed **R**ibosomal **F**rameshifting

Chapter 1

Introduction

Living organisms must regulate gene expression. Although much early focus was placed on regulation at the level of transcription, post-transcriptional regulation is just as important, particularly in eukaryotic cells. Once transcribed, precursor mRNAs undergo multiple RNA processing pathways throughout their life cycles. Each of these involves a wide array of protein and RNAs that mediate post-transcriptional gene expression regulation from mRNA processing to eventual cytoplasmic or nuclear decay. Unfortunately, much of this regulation relies on the production of short-lived transcripts that have been largely skipped by traditional sequencing and annotation processes. Thus, we need a means of more comprehensively cataloging the full eukaryotic transcriptome in order to better understand post-transcriptional gene regulation.

1.1 General mRNA Processing

After transcription, *m*essenger *RNAs* (mRNAs) must first go through extensive processing prior to translation. This happens through a series of discrete mRNA

processing pathways that all involve a multitude of RNA-protein and RNA-RNA interactions. Many of these contacts have acute affinities for specific RNA sequences or features, ensuring the proper advancement of an mRNA through each step of processing. Below is a summary of these pathways, and their required RNA and protein components, with a particular focus on pre-mRNA splicing and the requisite macromolecular complexes (i.e., spliceosome and the exon junction complex).

1.1.1 Pre-mRNA Splicing

Faithful transcription of the genome produces *pre*mature-*m*essenger *RNAs* (pre-mRNAs), each a sequence copy of its DNA predecessor before further nuclear processing. In its initial state, a pre-mRNA contains a combination of *intr*agenic non-protein-coding regi*ons*, called introns, and *ex*pressed regi*ons*, or exons. Removal of intronic sequences and subsequent ligation of neighboring exons occurs during a process known as splicing (Figure 1.1). Though depicted here as a singular event, the average human gene contains nine to ten introns and some even have more than 100 (Lander et al. 2001). Further, introns comprise the bulk of a transcribed message, with lengths typically 10 times that of internal exons. Splicing, therefore, generates significantly shorter mRNA products for much of the transcriptome. In fact, length discrepancies between viral DNA and cytoplasmic mRNA motivated the initial research into the then-unknown step in mRNA metabolism (Berget et al. 1977). As sequencing methodology was not yet readily available, authors instead observed mRNA-DNA hybrids under an electron microscope. Disparate regions between the two molecules caused stretches to remain unpaired. Free-floating mRNA ends confirmed expected 5′ and 3′ processing, whereas loops of single-stranded DNA advanced a model of post-transcriptional mRNA processing via splicing.

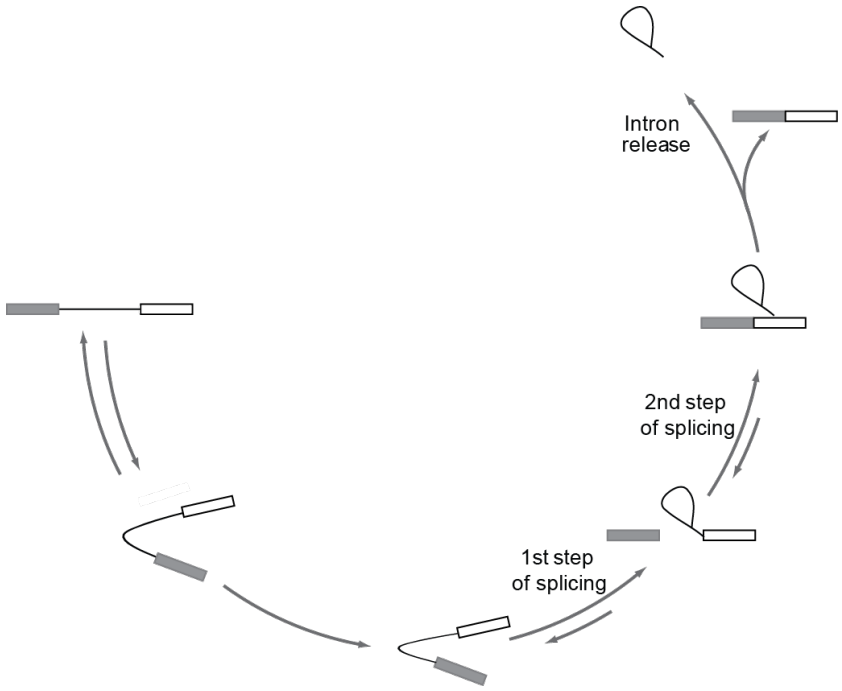

Figure 1.1: Schematic of the splicing cycle including the two steps of splicing and release of the intron lariat product. Boxes indicate exons; line show introns. Figure adapted from Shcherbakova et al. 2013. Copyright 2013 Elsevier.

Efficient intron recognition and removal necessitates that a common feature (i.e., sequence or structure) exists in every intron given their ubiquity across the genome. Such features were first identified when authors compared six intron-exon boundaries within the chicken ovalbumin gene (Breathnach et al. 1978). Sequence juxtaposition revealed distinct patterns at both 5′ and 3′ extremities, later called "splice sites." Though some nucleotides within the splice sites fluctuated, intronic 5′ and 3′ ends invariably coded for G-U and A-G dinucleotides, respectively. Moreover, splice site sequences appeared to be conserved among the few introns identified at the time in other species. More recent advances to sequencing methodology has since made it possible to analyze introns genome-wide. A comparison of thousands of short introns in five different organisms not only validated initial observations of the chicken introns, but also highlighted distinct differences between species (Lim & Burge 2001). In particular, more advanced organisms (e.g., humans) have less stringent sequence requirements at splice sites (Figure 1.2). This was most striking at branch points, the internal sequences used to form an intronic lariat

structure during splicing (Section 1.1.2). In yeast, branch points have very little variability and are easily identifiable by sequence alone (Figure 1.2, top middle). Though computational methods for predicting splice sites *a priori* have since been developed (Desmet et al. 2009), the extent of motif degeneracy in human introns obfuscates the identification of novel introns.

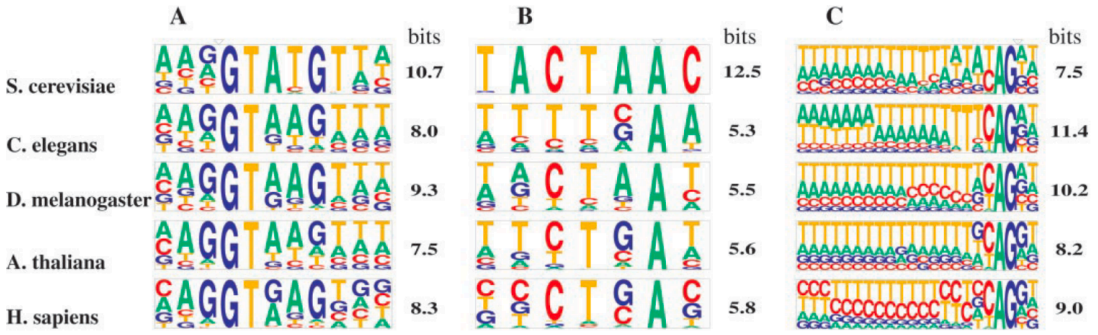

Figure 1.2: Sequence motifs for 5' (A) and 3' (C) splice sites or branch points (B) used in five different species. Letter height signifies the relative abundance of that nucleotide at each position. Figure from Lim and Burge 2001. Copyright 2001 National Academy of Sciences.

Soon after the discovery of introns, research began to uncover the mechanistic details behind splicing. Early observations of intron excision in both yeast and cellular extracts established that splicing proceeds through two transesterification reactions (reviewed in Guthrie 1991). The first of these happens when a 2′ hydroxyl at the branch point adenosine residue nucleophilically attacks the phosphate residue upstream of the G-U dinucleotide in the 5′ splice site (Moore & Sharp 1993). Successful completion of this step produces two splicing intermediates, a free 5′ exon and an intron lariat structure attached to the 3′ exon (Figure 1.1). The second step begins when the new 3′-OH on the 5′ exon attacks the phosphate following the 3′ A-G dinucleotide (reviewed in Wahl et al. 2009). This leads to exon-exon ligation and release of the intron lariat structure. In order for both reactions to occur efficiently, 5′ and 3′ splice sites must be close in spatial proximity. As such, introns with extensive secondary and/or tertiary structures possess the ability to splice in *cis* without the aid of any other factors. The majority of introns, however, lack the necessary conserved sequences to self-splice, and

must rely on a host of trans-acting proteins for excision.

1.1.2 Spliceosome Assembly

Initial experiments in cell-free extracts determined that efficient splicing of nuclear pre-mRNAs required two additional factors: ATP and an unexpectedly large complex of proteins (reviewed in Guthrie 1991). The macromolecular complex, later identified as the spliceosome, is now known to contain five *s*mall *n*uclear *r*ibo*n*ucleo*p*roteins (snRNPs), the Prp19 complex (NTC), and more than 80 associated proteins (reviewed in Wahl et al. 2009). Each individual snRNP consists of a unique non-coding snRNA (U1, U2, U4, U5, or U6) and numerous accessory proteins. To ensure precise and accurate excision of intronic sequences, both protein and RNA components of snRNPs bind to the previously described splice sites and other intronic regions via transient, step-wise interactions (reviewed in Brow 2002; Wahl et al. 2009). As such, the presence or absence of specific snRNPs, and other complex-specific proteins, defines the discrete stages in the spliceosome assembly cycle (Figure 1.3). Hereafter follows a stepwise walkthrough of the *de novo* formation of the spliceosome on a pre-mRNA.

Assembly of the first spliceosomal complex, known as the early (E) complex, begins upon U1 snRNP association with a pre-mRNA substrate. Sequences within the U1 snRNA complement 5′ splice site motifs, and these RNA-RNA interactions help position the snRNP at the exon-intron boundary (reviewed in Wahl et al. 2009). Due to their short length, however, these contacts alone are insufficient to maintain association with U1 snRNP. E complex stability, therefore, relies on further interactions between spliceosomal proteins and the pre-mRNA. Specifically, SF1/BBP and a U2 snRNP auxiliary protein bind both the branch point and downstream polypyrimidine tract in this complex. Thus, U1 snRNP binding and

subsequent E complex assembly defines both the 5′ and 3′ ends of an intron.

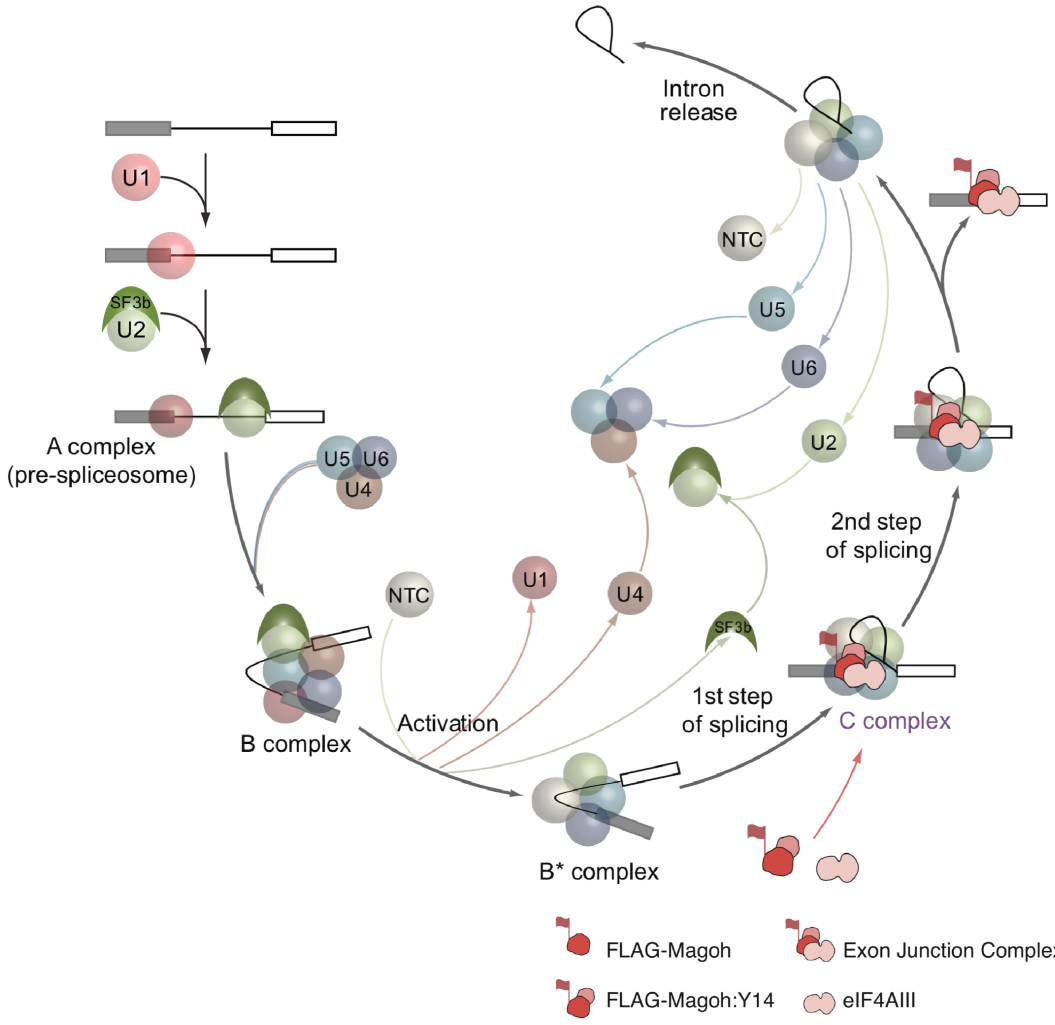

Figure 1.3: Schematic of the spliceosome assembly cycle showing the stepwise composition of snRNP(s) and RNA within each discrete spliceosomal complex. Based on data available prior to single molecule observation of the spliceosome (Section 3.2). Figure adapted from Shcherbakova et al. 2013. Copyright 2013 Elsevier.

The spliceosome transitions to its second stage, the A complex, when U2 snRNA binds the branch point and displaces the aforementioned E complex proteins (Figure 1.3). This rearrangement involves a number of accessory proteins, including two different ATPases, and thus marks the first of many ATP-dependent steps throughout the assembly cycle (reviewed in Wahl et al. 2009). Furthermore, much like the E complex, stable A complex formation requires a combination of RNA-RNA and RNA-protein interactions. Complementary sequences within U2 snRNA form base-pairing contacts within the branch point, forcing the adenosine

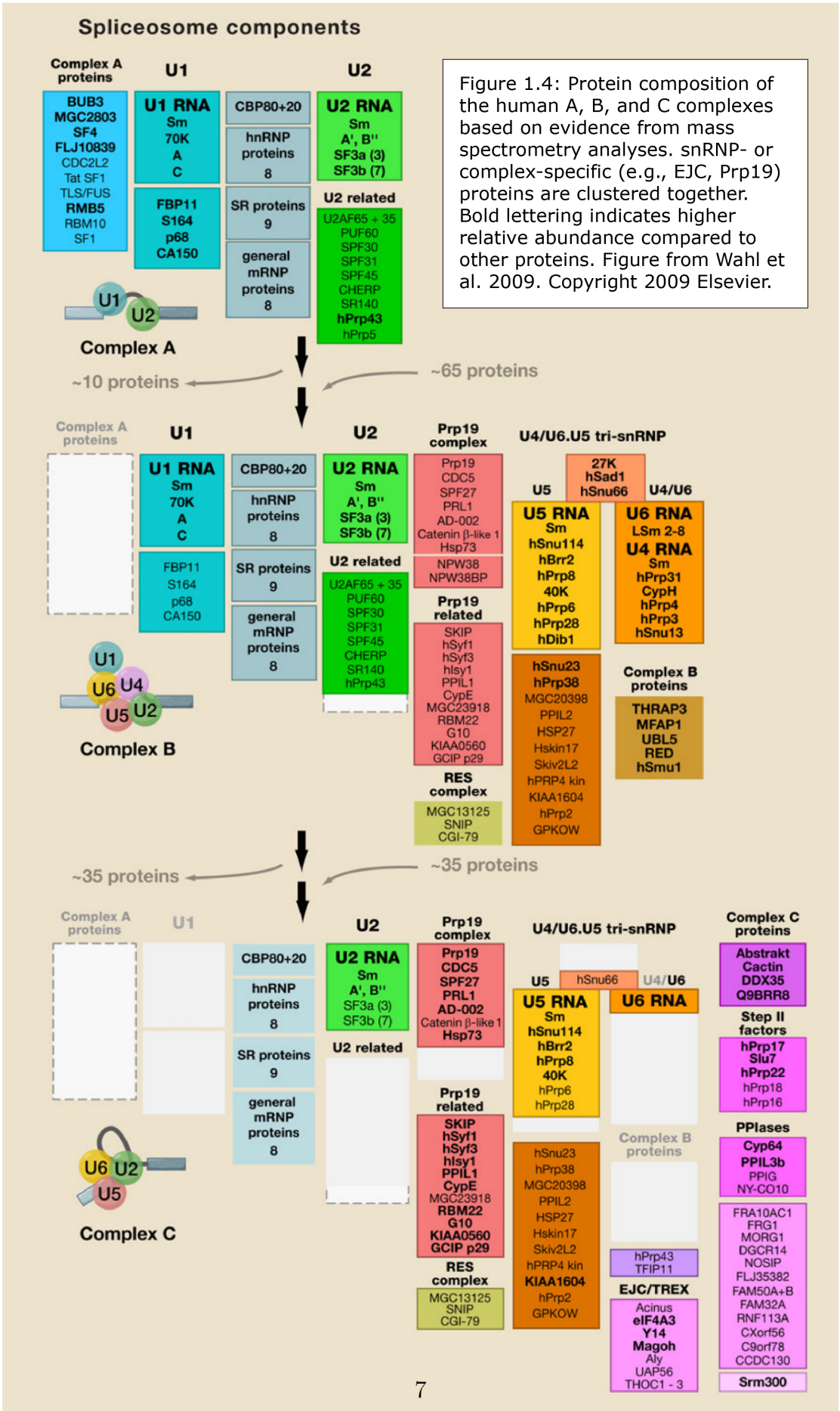

Figure 1.4: Protein composition of the human A, B, and C complexes based on evidence from mass spectrometry analyses. snRNP- or complex-specific (e.g., EJC, Prp19) proteins are clustered together. Bold lettering indicates higher relative abundance compared to other proteins. Figure from Wahl et al. 2009. Copyright 2009 Elsevier.

Figure 1.4

residue to bulge out prior to the first transesterification reaction (Query et al. 1994). Additional interactions between either U2 snRNA or the intron itself and various U2 snRNP proteins help support the structural conformation within the A complex.

Following A complex assembly, the pre-catalytic spliceosome, or B complex, forms upon recruitment of the U4/U6 and U5 snRNPs, including many accessory proteins (Figure 1.3). These three join the spliceosome as a preassembled complex, known as the tri-snRNP, after having been pre-processed in an upstream pathway (reviewed in Wahl et al. 2009). The tri-snRNP starts off as a di-snRNP established through extensive base-pairing interactions between U4 and U6 snRNAs, and later combines with the U5 snRNP by means of several protein-protein connections. Arrangement in this way poises the tri-snRNP for entry into the spliceosome, upon which U6 and U5 snRNAs create RNA-RNA interactions with U2 snRNA and the pre-mRNA, respectively (Figure 1.5, left). Stability of these interactions requires a number of associated proteins, particularly Prp19 and the Nineteen Complex. However, even with these accessory proteins and all five snRNPS, the B complex remains inactive until further structural and protein rearrangements.

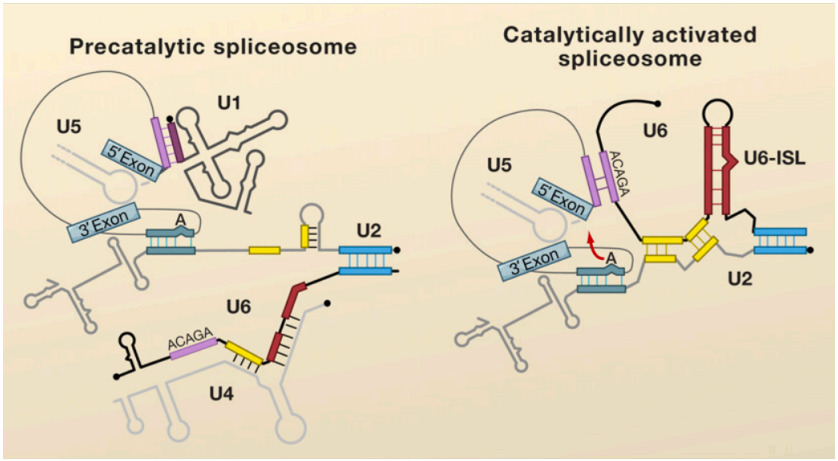

Figure 1.5: Representation of changes in RNA-RNA interactions between pre-mRNA splice site sequences and snRNAs within the precatalytic (left) and catalytically active (right) spliceosomes. Figure from Wahl et al. 2009. Copyright 2009 Elsevier.

The spliceosome continues to be catalytically inactive until the B* complex is

generated (reviewed in Wahl et al. 2009). Similar to the B complex, various ATPases facilitate this transformation, specifically by displacing U1 snRNA from the 5′ splice site. This intronic region is then available to form RNA-RNA contacts with conserved sequences in U6 snRNA once it has been released from U4 snRNA. While in the activated confirmation, additional connections are established between U6 and U2 snRNAs to position the branch point near the 5′ splice site. Such structural rearrangements lead to the first transesterification reaction (Section 1.1.1), which occurs during the transition from the B* complex into the C complex. Further conformational changes within the latter complex bring the 5′ and 3′ splice sites in close proximity, causing the second transesterification reaction (reviewed in Will & Luhrmann 2011). Successful nucleophilic attack in this reaction results in ligation of the two exons and formation of an intron lariat RNA (Section 1.1.1). Following this catalytic step, the splicing cycle concludes with spliceosome disassembly and recycling, and subsequent release of the mature mRNA and intron lariat products.

1.1.2.1 Alternative Spliceosome Assembly

In the pathway outline above, spliceosome assembly occurs across a single intron, establishing its 5′ and 3′ boundaries during A complex formation. This is feasible for short introns, but mammalian introns can span more than a hundred thousand nucleotides (Lander et al. 2001). Internal exons, on the other hand, are substantially shorter ($\leq$ 150 nt). As such, closely spaced splice sites on either end of an exon may be recognized first by the splicing machinery, otherwise known as cross-exon definition (Berget 1995). As with cross-intron spliceosomes, U1 and U2 snRNP binding define the 5′ and 3′ ends of an exon, respectively. The two pathways diverge, however, by the way in which the two snRNPs interact (Braun et al. 2018). Whereas cross-intron A complex assembly relies on contacts between U1

and U2 to promote binding between U2 and the branch point, cross-exon assembly results from U1 snRNP interactions with splicing factors at a nearby 3′ splice site that recruit the U2 snRNP. The two pathways converge during subsequent steps in assembly, which shift the complex to cross-intron interactions and follow the previously described pathway (Section 1.1.2).

1.1.3 Exon Junction Complex

Even before discovering specific details of the splicing cycle and spliceosome assembly, it was evident in early experiments that this mRNA processing event had long-term effects on gene expression. Hamer *et al.* first explored this using a series of modified viral transcripts containing between zero and two introns (Hamer & Leder 1979). The mRNA stability of spliced mRNAs drastically increased compared to the intronless control. However these analyses only examined mRNAs containing introns within the 5′ UTR and/or ORF. In fact, placing introns downstream of an early termination codon had the opposite outcome, and led to increased turnover through the nonsense-mediated decay pathway (Section 1.4.1; Carter et al. 1996). In both studies, completely removing the intron abolished the impact on mRNA stability, suggesting that productive splicing must leave behind something more than ligated exons as a record of its history.

To confirm such a potential marker, Luo and Reed analyzed the protein complexes associated with intronless and intron-containing transcripts (Luo & Reed 1999). Spliced transcripts traveled slower in native gels due to association with multiple *m*essenger *r*ibo*n*ucleo*p*rotein complexes (mRNPs), none of which bound the unspliced controls (Figure 1.6, left). One of the mRNPs was readily identifiable as the intron complex, whereas the second bound to spliced mRNAs. To determine if this spliced mRNP was simply lingering spliceosomes, authors examined the com-

plexes associated with a slower spliced transcript (AdML, Figure 1.6 right). This revealed different migration rates between the spliceosome and the spliced mRNP, thus proving that a complex of proteins remained bound to processed transcripts after the splicing machinery disassembled.

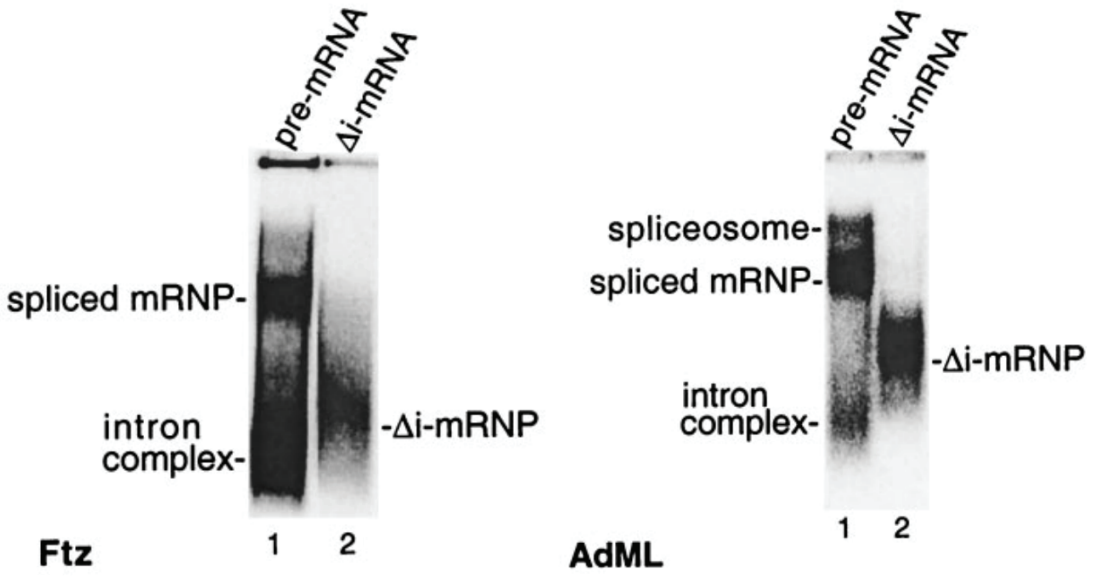

Figure 1.6: Differences in complexes associated with spliced and unspliced (Δi-mRNP) mRNAs on native gels. Splicing of Ftz pre-mRNAs (left) is more efficient than AdML pre-mRNAs (right), and the spliceosome is undetectable in this gel. Figure adapted from Luo and Reed 1999. Copyright 1999 National Academy of Sciences.

Research thereafter continued on a pursuit towards identifying the protein(s) component(s) of these splicing-dependent mRNPs. Le Hir *et al.* first isolated the complex by crosslinking it to a photoreactive modification near the exon-exon junction of a synthesized transcript (Le Hir, Melissa J. Moore, et al. 2000). Using two control mRNAs, the authors compared the spliced mRNP to proteins associated prior to or in the absence of splicing and found several proteins specific to this complex. Many of these proteins, however, could not be named without further experimentation (Le Hir, Izaurralde, et al. 2000). Authors co-immunoprecipitated candidate proteins and analyzed the associated mRNA to confirm their inclusion in the spliced mRNP. As expected, members of the complex pulled down more spliced mRNA than an unprocessed control. Furthermore, all complex proteins

bound the same section of the spliced mRNA, an eight nucleotide region located 24 nucleotides upstream of the exon-exon junction (Figure 1.7), and protected it from RNase H digestion. Thus, the authors identified a ~335 kD complex consisting of SRm160, DEK, RNPS1, REF, and Y14.

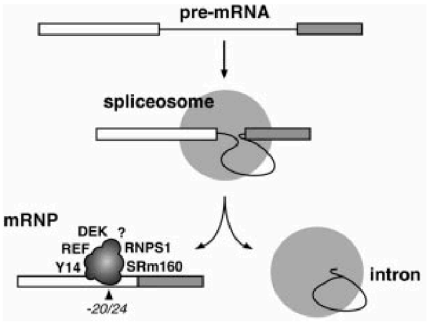

Figure 1.7: Schematic of the location (-24 nt) and initially proposed composition of the exon junction complex. Figure from Le Hir, Izaurralde, et al. 2000. Copyright 2000 Wiley.

These initial studies, however, did not identify all of the proteins now known to be core components of the *e*xon *j*unction *c*omplex (EJC). One of these, Magoh, was later found during a yeast two-hybrid screening of proteins capable of interacting with Y14 (Kataoka et al. 2001). Supporting its role in the EJC, Magoh preferentially bound spliced mRNAs and protected the same -24 nt region identified by Le Hir *et al.* Moreover, Y14 and Magoh distribution after cellular fractionation demonstrated that these EJC factors were unique in their continued association with spliced mRNAs in the cytoplasm. Subsequent mass spectrometry analysis revealed that the cytoplasmic complex of Y14 and Magoh contained a third factor, eIF4AIII (Chan et al. 2004). The DEAD-box RNA-binding protein behaved similarly to Y14 and Magoh in terms of both its RNA footprint (i.e., 8 nt at -24) and cellular localization pattern. Following identification of the final factor, MLN51, a crystal structure of the four EJC core proteins bound to an mRNA mimic was obtained in 2006 (Andersen et al. 2006).

1.1.3.1 EJC Recruitment and Assembly

Given the importance of the EJC in both mRNA expression and stability, it was critical to next understand how and when this complex assembles onto an intron-containing transcript. To track EJC deposition, mRNAs were first radioactively labeled at -24 nt upstream of the exon junction, incubated in splicing reactions, and then digested with multiple RNases (Reichert et al. 2002). Whereas unspliced pre-mRNAs showed no signs of protection from RNase degradation, the EJC footprint appeared on free 5′ exons formed during the first step of splicing (Figure 1.3). Mass spectrometry analysis of both the C complex and final spliced mRNP confirmed that EJC core proteins, specifically Y14 and Magoh, assemble onto mRNAs prior to exon ligation and remain bound long after spliceosome disassembly. Though this evidence established when the fully-assembled EJC bound mRNAs, it did not address whether the complex associates with the spliceosome prior to binding. To answer this question, Merz *et al.* tested whether the EJC interacted with the spliceosome in the absence of binding using a pre-mRNA substrate with a truncated 5′ exon (Merz et al. 2007). Although purified B*/C complexes bound to this transcript contained the EJC, none of the core proteins remained in the final spliced mRNP. These results demonstrated that EJC association does not require an available binding platform because recruitment and assembly happen as independent events.

In fact, further studies showed that recruitment of individual EJC core proteins occurs through independent interactions with the spliceosome (Figure 1.8). CWC22, an essential splicing factor, recruits eIF4AII to the spliceosome prior to its association with other EJC components (reviewed in Woodward et al. 2017). Interactions between the two proteins locks the latter into an inactive conformation until further assembly steps. On the other hand, specifics about Y14 and Magoh recruitment remain largely unknown. Recent evidence suggests that these factors

join as a pre-assembled heterodimer, though it is unclear what facilitates its incorporation into the spliceosome (reviewed in Woodward et al. 2017). Current models propose that stable recruitment and deposition of Y14:Magoh potentially relies on interactions with either eIF4AII or IBP160, a tri-snRNP protein.

After the first step of splicing, the separately recruited proteins assemble into the EJC while associated with the C complex (Figure 1.8). Assembly initiates when CWC22 releases eIF4AII, allowing it to change to an active confirmation capable of binding both ATP and RNA (reviewed in Woodward et al. 2017). The crystal structure of eIF4AII in this conformational state showed that it binds to the sugar-phosphate backbone, allowing for precise localization (i.e., 8 nt at -24) of the EJC independent of mRNA sequence (Andersen et al. 2006). Once in its active state, the heterodimer then binds to eIF4AII and locks it onto the 5′ exon by blocking further ATPase activity. Following release from the spliceosome, MLN51, the last remaining core protein, further stabilizes eIF4AII on the mRNA through contacts with the surrounding nucleotides.

1.1.3.2 EJC Disassembly and Translation

Once assembled, EJCs continue to bind spliced transcripts through nuclear export and in the cytoplasm (Figure 1.8). This association is now known to continue until the first, or "pioneer," round of translation. To investigate the relationship between the EJC and translation, Dostie and Dreyfuss fractionated cytoplasmic material across sucrose gradients (Dostie & Dreyfuss 2002). The mRNP and, to a lesser extent, the monosome fractions both contained Y14, and presumably the rest of the EJC. However, the complex was missing from the polysome fraction. The authors further examined the effect of translation by comparing levels of Y14 across a number of modified mRNAs and found enhanced association on transla-

tionally inactive transcripts. Although these results established that EJC removal requires active translation, mechanistic details behind this relationship continue to be investigated. As of now, it has been suggested that removal requires interactions between the EJC and an accessory protein, PYM (reviewed in Woodward et al. 2017). PYM contacts both the EJC and the translating ribosome through its N- and C-terminus, respectively, thus connecting the two macromolecular complexes. Interactions with between the ribosome and PYM, however, have been shown to be dispensable for EJC disassembly in *Drosophila*. As such, it remains unclear how PYM functions *in vivo*.

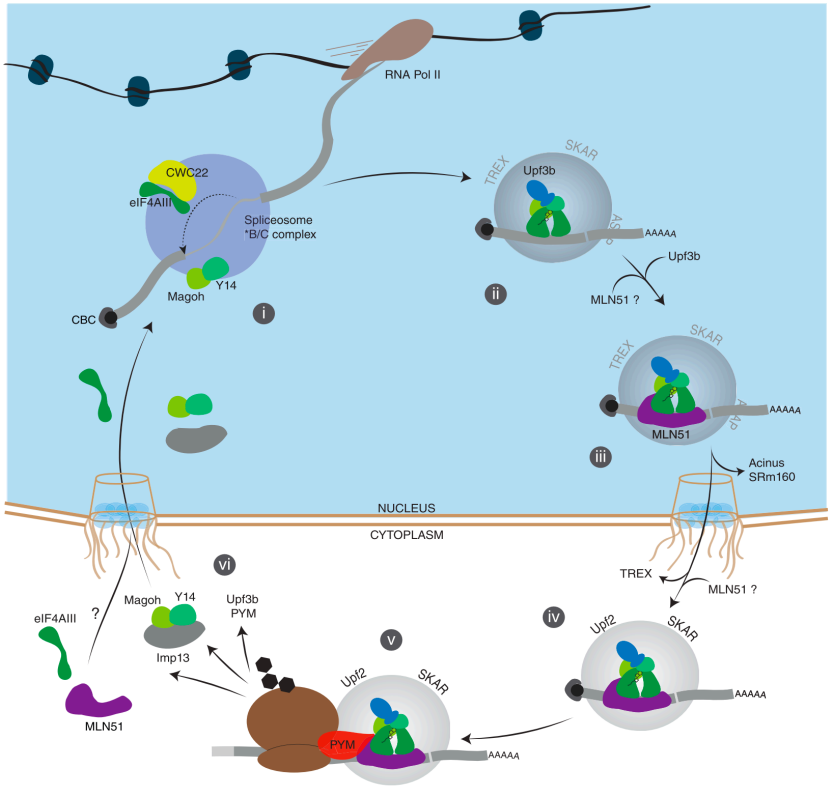

Figure 1.8: Depiction of the exon junction complex assembly cycle. (i) EJC components are recruited to the catalytically active spliceosome complex. (ii) Following release from the spliceosome, the EJC remains stably bound -24 nt upstream of the newly created exon-exon junction. (iii to v) The EJC-containing mRNP is exported to the cytoplasm where it remains associated until the pioneer round of translation (vi). EJC components re-enter the nucleus and are recycled for use in subsequent splicing reactions. Figure from Woodward et al. 2017. Copyright 2017 Wiley.

1.2 Alternative mRNA Processing

Though the mRNA life cycle has been presented thus far as a series of simplified linear pathways, numerous alternative processing steps can occur throughout an mRNA's existence. Without these, each gene would simply produce a single transcript that coded for only one protein. Instead multiple mRNAs and proteins derive from just one gene through a combination of co- and post-transcriptional mechanisms, including: alternative transcription initiation, alternative splicing, *a*lternative *p*oly*a*denylation (APA), and alternative translation initiation.

Not only do these alternative mRNA processing events greatly expand the transcriptome, they also provide multiple opportunities to regulate gene expression. For example, different APA choices change the length of 3′ UTRs between transcripts. As this region often harbors localization signals, these differences can greatly modify the localization pattern of a transcript. Furthermore, shorter 3′ UTRs profoundly affect translation efficiency due to decreased interactions between the polyA tail and initiation factors (Section 1.3.1; reviewed in Klerk & 't Hoen 2015). The reduction in translation leads to increased turnover of the mRNA (Section 1.3.1), thus connecting APA and mRNA stability.

Though the majority of the genome contains APA sites, many are functionally suppressed and, consequently, largely unanalyzed (reviewed in Klerk & 't Hoen 2015). This is not unique, however, to this form of alternative mRNA processing. As many of these mechanisms significantly alter the half-life of a transcript, evidence of their usage is fleeting. Fortunately, recent advancements in sequencing technology have provided better ways of exploring the frequency of many alternative mRNA processing events. For the sake of brevity, the following section will focus solely on alternative splicing and its ramifications on gene expression.

1.2.1 Alternative Splicing

The pre-mRNA splicing cycle in Figure 1.1 depicts a simplistic model in which neighboring exons always join to form the final mRNA product. In fact, this mRNA processing event is often much more complicated. *A*lternative *s*plicing (AS) varies the exonic regions included or excluded when processing each pre-mRNA, thereby expanding the transcriptome and increasing protein diversity. AS events fall into one of the following categories, each of which is depicted in Figure 1.9: (a) alternative $5'$ splice site; (b) alternative $3'$ splice site; (c) cassette exon - skipping or inclusion; (d) mutually exclusive exons; and (5) intron retention.

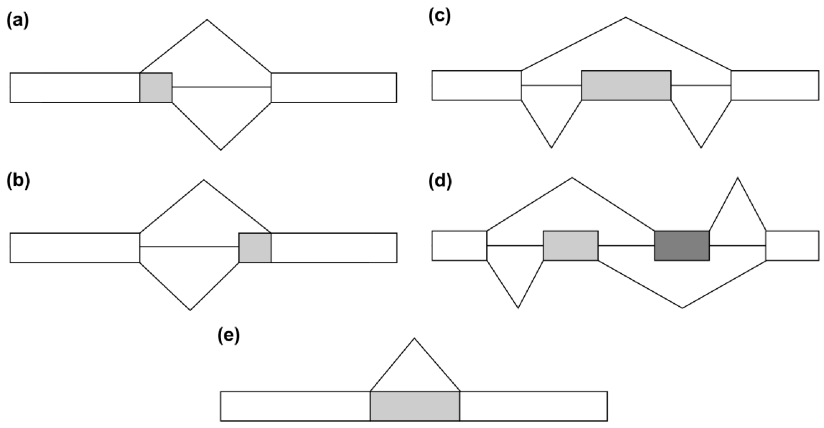

Figure 1.9: Figure 1.4: Major types of alternative splicing: (a) alternative $5'$ splice site; (b) alternative $3'$ splice site; (c) cassette exon - skipping or inclusion; (d) mutually exclusive exons; and (5) intron retention. Constitutive exons appear as open boxes, and shaded boxes represent alternative exons. Lines above and below the depicted pre-mRNA show alternative splicing events. Figure adapted from Graveley 2001. Copyright 2001 Elsevier.

Similar to the discovery of splicing (Section 1.1.1; Berget et al. 1977), the first observation of alternative splicing happened during the analysis of various mRNAs transcribed from the adenovirus genome (Chow et al. 1977; Klessig 1977). None of the transcripts examined completely hybridized to DNA, which indicated that each of the viral gene products underwent splicing. The authors confirmed this using short sequence probes that covered potential exonic regions. However, two mRNAs expressed late in the viral infection cycle failed this secondary hybridization due to discontinuous sequence rearrangements. Chow *et al.* suggested this must result

from alternate processing of adenovirus transcripts, but could also happen for eukaryotic mRNAs. In 1982, Rosenfeld *et al.* observed such an event in rat thyroid tumor cells in response to calcitonin starvation (Rosenfeld et al. 1982). Further analysis revealed that alternative splicing patterns of calcitonin transcripts varied between neural tissue and thyroidal cells, and suggested this processing event requires a high level of regulation (Rosenfeld et al. 1983).

Based on these and similar results, AS was initially suspected to be simply a minor pathway (Stamm et al. 2005), but the prevalence of events grew exponentially as new methods of analyzing the transcriptome emerged. By the early 2000s, multiple analyses based on expressed sequence tags revealed that pre-mRNAs from 35 to 60% of all human genes undergo AS during maturation (reviewed in Stamm et al. 2005). This approximation later grew to 74% with the use of DNA microarrays. These technologies, however, likely underrepresented the true prevalence of AS as both only analyze a limited portion of each pre-mRNA and/or require *a priori* knowledge of its existence. As such, the advent of deep sequencing technology, which circumvents these limitations, significantly improved the ability to detect new pre-mRNA species. Consequently, extensive deep sequencing of RNA (RNA-Seq) has now revealed that more than 95% of human protein-coding genes are subject to AS (Eric T Wang et al. 2008; Pan et al. 2008).

Although this high percentage shows that almost every pre-mRNA must be processed differently by the splicing machinery, it provides little insight into the number of AS isoforms per gene. Current annotations have now identified ~82,000 different protein-coding mRNA isoforms transcribed from ~20,000 protein coding genes (Cunningham et al. 2019). Thus, on average, each gene produces approximately four different mature transcripts. In some cases, however, the amount of isoforms per gene can reach into the thousands. For example, *Drosophila* transcribe more than 38,000 isoforms from the *Down syndrome cell adhesion molecule*

(DSCAM) gene (Nilsen & Graveley 2010). Although this represents an extreme case, this gene alone produces more mRNAs than the number of genes in the fly genome (~14,500) and demonstrates how powerful AS is in diversifying the transcriptome.

1.2.1.1 Functional Alternative Splicing

Based on initial annotations, the vast majority (~75%) of AS events mapped within the *o*pen *r*eading *f*rame (ORF) of mRNAs (Stamm et al. 2005). As such, the functional relevance of each isoform could vary anywhere from a total loss-of-function to a complete overhaul of an mRNA's activity and localization. The former typically results from introducing a premature termination codon into the ORF, either by including intronic sequences containing a stop codon or by causing a frameshift that introduces one. These isoforms most often exist as a means of quickly abolishing further translation of a gene as these transcripts experience rapid turnover by the NMD pathway (Section 1.4.2).

Unlike the loss-of-function isoforms described above, not all of AS events lead to negative outcomes. In fact, many exist as a means of modifying protein function post-transcriptionally. For example, different isoforms of the calcitonin gene not only have tissue-specific expression patterns (Section 1.2.1), but also serve entirely different purposes within these particular tissues (reviewed in Stamm et al. 2005). In the thyroid, full-length calcitonin acts in calcium and phosphorus metabolism, whereas in the nucleus, the shorter variant known as calcitonin-gene-related peptide (CGRP) functions as a vasodilator. Furthermore, other AS events may excise entire binding or localization domains, as is the case with interleukin 4 (reviewed in Stamm et al. 2005). The isoform lacking its transmembrane domain remains soluble but otherwise fully functional, and soaks up growth hormone binding pro-

tein in order to regulate the signal transduction pathway of the membrane-bound form. In addition to these examples, AS can also control post-transcriptional gene expression by affecting protein and/or mRNA stability, signaling activity, and post-translational modifications.

1.2.1.2 Unproductive Splicing Events / Splicing Noise

However, the sheer number of transcripts identified by these deeper analyses makes it implausible for every mRNA in the cell to have functional relevance. As evaluating the biological purpose of each isoform presents an impractical task, authors have long relied on the lack of conserved splicing patterns to identify aberrant splicing events, or "splicing noise." AS events conserved across different species, particularly those that evolutionarily diverged millions of years ago, likely represent functionally important transcripts. On the other hand, those lacking function face no such evolutionary constraint and are not expected to appear in multiple species. Surprisingly, a comparison of EST-based transcriptomes of mice and humans revealed that only a small percentage of exon-skipping events were conserved (Yeo et al. 2005). Subsequent RNA-Seq experiments uncovered a much greater overlap between the two species and even more so between humans and primates, but many AS exons continued to be species-specific (Barbosa-Morais et al. 2012).

Thereafter, it next became a question whether these unconserved transcripts had any functionality within the cell or if they simply resulted from aberrant splicing events. By examining alternative and constitutive exons transcriptome-wide, Dou *et al.* found that alternative $5'$ and $3'$ splice sites most often occurred within a few nucleotides of the dominant splice site (Dou et al. 2006). These results suggested that the splicing machinery could use nearby AG/GT dinucleotides during catalysis. However, further investigations determined that although weak consen-

sus sequences drove the frequency of alternative splicing events, the spliceosome acts with a high degree of fidelity (i.e., one error per 100,000 events) (Fox-Walsh & Hertel 2009). Nonetheless, this study only addressed the error rates of two genes. Extensive deep sequencing experiments later found a transcriptome-wide error rate of 0.7%, and that longer intronic regions significantly increased this rate (Pickrell et al. 2010). In spite of that, the authors acknowledged these rates likely underestimate the true quantity of splicing noise due to rapid turnover by the NMD pathway (Sections 1.4.1 and 1.4.1.1). Therefore, the transcriptome must be evaluated in the absence of this pathway in order to observe the true rate of splicing noise.

1.2.1.3 Regulating Alternative Splicing Across the Transcriptome

Whereas the NMD pathway can remove aberrantly spliced transcripts once produced, a number of other regulatory mechanisms control mRNA expression before this point by influencing spliceosome assembly. Efficient assembly depends on accurate splice site recognition, and pre-mRNAs encode additional information that aids or inhibits this process. In fact, one of the first studies into the regulation of alternatively spliced transcripts from adenovirus revealed that secondary structure of the pre-mRNA heavily influenced splice site choice (Solnick 1985). It is not feasible, however, for every transcript to contain sequences amenable to forming strong hairpins. Rather pre-mRNAs contain short sequences, much like splice sites, called *s*plicing *r*egulatory *e*lements (SREs) that can be split into four categories: exonic/intronic splicing enhancers (ESEs and ISEs) or exonic/intronic splicing silencers (ESSs and ISSs). Depending on the intended functionality, these elements either stimulate or repress spliceosome assembly at nearby splice sites through

interactions with accessory regulatory proteins (Will & Luhrmann 2011).

Among the most well studied of these auxiliary spliceosome factors are two families of ubiquitously expressed RNA binding proteins, SR proteins and *h*eterogeneous *n*uclear *r*ibo*n*ucleo*p*roteins (hnRNPs). SR proteins bind to enhancer elements and then recruit members of U1 and U2 snRNPs through direct protein-protein interactions (Busch & Hertel 2012). Contrarily, hnRNPs downregulate spliceosome assembly, though the exact mechanistic details behind this regulation remain unclear (reviewed in Busch & Hertel 2012). Binding motifs for both proteins appear near both constitutively and alternatively utilized splice sites, and the location and frequency of these sites impacts both proteins' effectiveness. However, this effect seems to be more nuanced than what is currently understood. Although exon inclusion rates correlate well with the distance between ESEs and splice sites, adding more of these elements failed to further promote spliceosome assembly (reviewed in Busch & Hertel 2012). Even so, increased binding potential does not necessarily translate into increasing the amount of bound SR proteins. In fact, the cellular concentration of many regulatory proteins is tightly controlled through autoregulated feedback loops, typically by producing inactive isoforms (Kelemen et al. 2013).

In addition to these proteins, more recent work discovered an abundance of *trans*-acting RNAs can also influence splice site selection. Similar to SREs, *mi*cro *RNAs* (miRNAs) do not directly interact with the spliceosome, but rather act on accessory regulatory proteins. In one such example, miR-124 regulates the abundance of PTB, a well known hnRNP, thus leading to increased exon inclusion rates due to decreased levels of PTB (reviewed in Kelemen et al. 2013). Due to their own unique expression profiles, other miRNAs target SR and hnRNP proteins to indirectly regulate AS events in a tissue-specific manner. Furthermore, although many are not spliced themselves, long non-coding RNAs (lncRNAs) have now

been linked to AS regulation through control over the phosphorylation state of SR proteins (reviewed in Kelemen et al. 2013). Additional roles for other non-coding RNAs (e.g., tRNAs) have since been identified, though the breadth of this method of regulation continues to be investigated.

1.2.1.4 Splicing Mutants and Disease

Considering the functional importance of many alternatively spliced isoforms, it is not shocking that disrupting typical splicing patterns often leads to widespread deleterious effects. Perturbation can be caused by mutations within splice sites, introns, or the splicing machinery itself. An early instance of this was discovered while investigating the aberrant expression profile of the β-globin polypeptide associated with β-thalassaemia diseases (Treisman et al. 1983). Compared to normal patients, those afflicted with the disease produced four extra isoforms that either reduced or abolished polypeptide synthesis. Sequencing revealed that three of these transcripts resulted from single nucleotide mutations surrounding the $5'$ splice site. Though two merely weakened the splice site, the third mutation fell within the $5'$ dinucleotide and abolished splicing of the intron all together. The final mutation, however, modified sequences within an upstream intron, generating a novel $5'$ splice site (Figure 1.10). This is sufficient to activate a nearby cryptic $3'$ splice site and produce an isoform containing a disease-specific cassette exon.

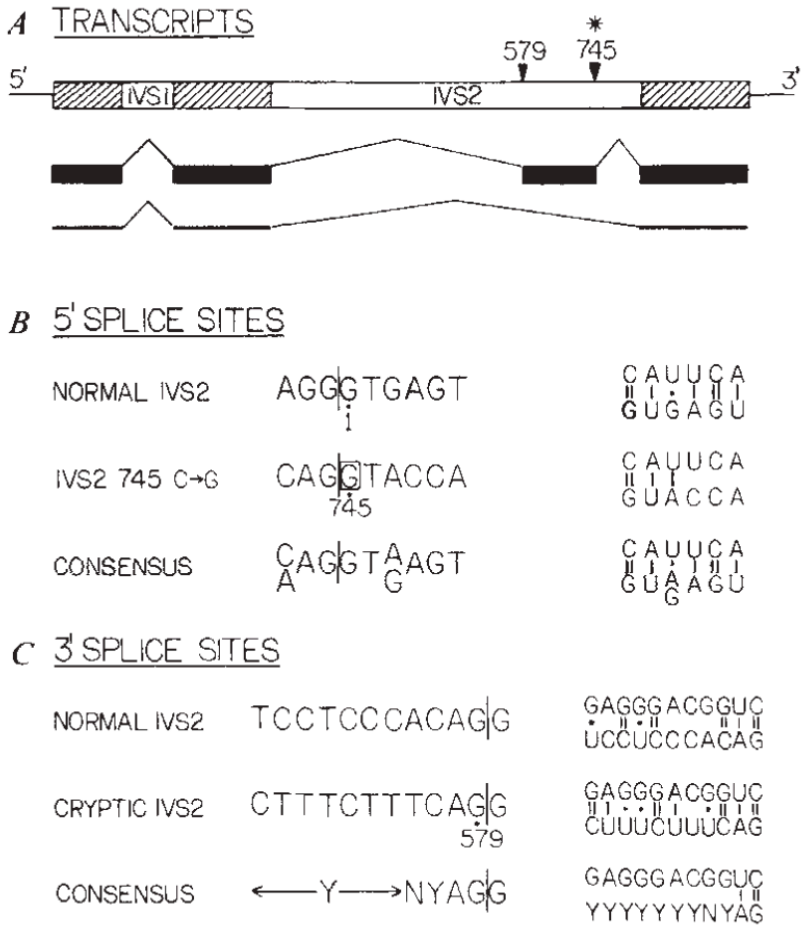

Figure 1.10: Schematic of the alternative spliced transcripts produced when the ß-thalassaemia gene is mutated at position 745. This mutation creates a new 5' splice site, shown in (B) compared to the normal splice site and consensus sequences. This activates a nearby cryptic 3' splice site at position 579, which is shown in (C). Figure from Treisman et al. 1983. Copyright 1983 Nature Publishing Group.

Subsequent to this discovery, alternative splicing changes have since been connected to countless diseases, including a wide range of cancers. Cancer-causing mutations may occur in exons or introns, similar to B-thalassaemia, but can also result from mutations in core and accessory spliceosome proteins (reviewed in Climente-González et al. 2017). Many of these have been identified through large-scale deep sequencing and bioinformatics analyses of cancerous tissues, and have since been catalogued in *T*he *C*ancer *G*enome *A*tlas (TCGA) (Weinstein et al. 2013). By comparing more than 4,500 samples from 11 different cancer types from TCGA, Climente-González *et al.* identified thousands of AS isoform switches recurring amongst the tumors (Climente-González et al. 2017). The vast majority

of these switches generated transcripts containing added or removed protein domains and those with impaired or gained protein-protein interactions. However, the authors could not pinpoint *cis*-acting somatic mutations responsible for these switches, and suggested instead that many originated from *trans*-acting changes within splicing factors. At this time, research continues to dissect the relationship between these factors and observable phenotypic changes.

1.3 General mRNA Decay

Much like the aforementioned nuclear processing, clearance of mRNA from the cell (i.e., general mRNA degradation or decay) is also accomplished through tightly controlled processes involving numerous protein competents. Some of these proteins, though not all, participate in two general mRNA decay pathways (Figure 1.11). Both require targeted mRNAs to first be deadenylated (Section 1.3.1), but are then differentiated by whether degradation proceeds in the 5′-to-3′ (Section 1.3.2) or 3′-to-5′ (Section 1.3.3) direction. Although depicted as branched pathways, decay of some mRNAs happens in both directions (Garneau et al. 2007). Collectively, the majority of mRNAs are removed by one or both of these pathways. A subset of the transcriptome, however, undergoes accelerated decay by other means, and these quality control processes will be discussed further in Section 1.4.

1.3.1 Deadenylation

General mRNA decay starts with progressive removal of the 3′ polyadenylated tail, a process known as deadenylation (Figure 1.11, top). As is the case with much of what has been discovered about mRNA decay, initiation via deadenylation was

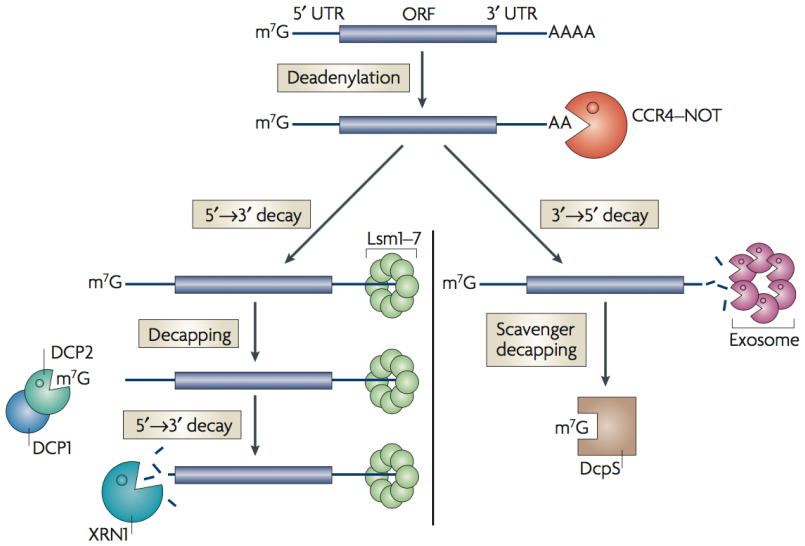

Figure 1.11: Deadenylation-dependent decay pathways. Once the major deadenylase complex, Ccr4-Not, shortens the poly(A) tail, mRNAs may be degraded from either side. Left, digestion at the 5' end begins when the cap is removed by Dcp2/Dcp2, exposing the mRNA to Xrn1. Right, the exosome degrades mRNAs 3'-to-5' and DcpS metabolizes the remaining cap structure. Figure adapted from Garneau et al. 2007. Copyright 2007 Nature Publishing Group.

initially explored using reporter mRNAs in yeast (Decker & Parker 1993). By measuring the length and abundance of various transcripts following transcription inhibition, the authors calculated the relative rates of deadenylation and decay, respectively. Although degradation of most analyzed mRNAs held off until deadenylation began, this was not the case for all transcripts. (Additional discussion on deadenylation-independent decay pathways can be found in Section 1.3.4.) Notably, subsequent steps in mRNA turnover did not appear to require the complete and absolute removal of the poly(A) tail. Captured decay intermediates maintained short tails that were typically long enough to accommodate a single copy of *p*oly(*A*)-*b*inding *p*rotein (PABP) (Sachs et al. 1987).

Prior to deadenylation, PABP binds along the full length tail, potentially limiting its exposure to cytoplasmic exonucleases during translation. Some such exonucleases form the major deadenylation complex, hereafter referred to as Ccr4-Not (Wiederhold & Passmore 2010). Though evolutionarily conserved, the role of the Ccr4-Not complex in decay has been largely examined using knock-out yeast

strains. When mutations impaired either exonucleolytic subunits (i.e., Ccr4 and Pop2), reporter mRNAs deadenylated less efficiently (Tucker et al. 2002). This effect could be rescued by overexpressing Ccr4p in either strain, suggesting this protein serves as the primary exonuclease within the complex. In *in vitro* experiments, Ccr4p activity was further tested in the presence and absence of purified Pab1p, which slowed deadenylation rates in a concentration-dependent manner. Even so, this inhibition was only observable when Pab1p significantly exceeded the amount of Ccr4-Not complex. Whereas at equilibrium, Pab1p-bound mRNAs deadenylated faster than naked transcripts, which suggests that PABP actually stimulates Ccr4 (Webster et al. 2018). This stimulation appears to be particularly vital in the turnover of short-tailed ($\leq$ 15 As) transcripts, and could exist as a final moment to control mRNA stability.

In fact, not all acts of deadenylation lead to decay. The poly(A) tail plays a critical role in regulating gene expression, much of which is mediated by PABP. For example, direct interactions between PABP and initiation factors bound to the 5′ cap promote translation initiation (Tarun & Sachs 1996). When the Ccr4-Not complex breaks these contacts by removing PABP during deadenylation, translation efficiency decreases in parallel. To understand the extent of PABP's role in gene regulation, short- ($\leq$ 22 As) and long-tailed mRNAs in yeast were characterized transcriptome-wide (Beilharz & Preiss 2007). Shorter tailed mRNAs were bound by less copies of PABP, as expected, and less ribosomes, confirming translation repression occurs as a byproduct of deadenylation. In higher eukaryotes, the relationship between tail length and ribosomal occupancy is used to regulate gene expression in the absence of transcription (Novoa et al. 2010). This occurs in both meiosis and early embryogenesis, during which cycles of tail shortening followed by cytoplasmic polyadenylation control translation. Readenylation makes this initial step of decay unique, and any further action commits an mRNA to destruction.

1.3.2 5′-to-3′ Decay

One such irreversible path is known as 5′-to-3′ mRNA decay (Figure 1.11, left). Following deadenylation, this pathway begins with the removal of the 5′ cap structure. When present, the cap is bound by eIF4E, a component of the translation initiation complex (reviewed in Kapp & Lorsch 2004). This complex interacts with PABP prior to deadenylation to load pre-translational ribosomal subunits (Section 1.3.1). As one would expect, loss of the cap-binding activity of eIF4E abolished these interactions and negatively impacted translation rates (Schwartz & Parker 1999). mRNAs in this mutant background also degraded faster, more so than the other translation factor mutations tested. Further analysis revealed that cap-bound eIF4E inhibits the onset of decay by blocking decapping enzymes during translation (Schwartz & Parker 2000). When eIF4E is removed, the cap becomes accessible and is rapidly cleaved by decapping proteins, Dcp1p and Dcp2p (reviewed in Coller & Parker 2004). This cleavage event generates two products, m^7GDP and 5′-monophophaste mRNA, the latter of which is a preferred substrate of the 5′-to-3′ exoribonuclease, Xrn1p.

In addition to the decapping enzymes and Xrn1p, there are a number of accessory proteins involved in 5′-to-3′ decay. One such example is the decapping activator complex, consisting of Pat1p and a heptameric ring of Lsm1p through Lsm7p, that preferentially binds near the 3′ end of deadenylated mRNAs (Tharun & Parker 2001). In the absence of this complex (i.e., knock-out yeast strains), short-tailed, capped mRNAs accumulated due to a reduction in decapping activity. Dcp2 is now known to be recruited to mRNAs by direct interactions with Pat1p, but mechanistic details beyond this continue to be investigated (Lobel et al. 2019). Recent deep sequencing experiments, however, determined that the decapping activator complex targets only a subset of the yeast transcriptome (He et al. 2018). Some of these mRNAs are co-regulated by Dhh1p, another decapping activator,

but still others responded only to one decay factor or the other (Dhh1p and its role in regulating decay will be discussed later in Section 1.3.5). Furthermore, knocking out any of the decapping activators did little to stabilize other transcripts that undergo accelerated turnover through quality control pathways (Section 1.4). Therefore, decay factors can trigger the onset of 5′-to-3′ degradation in varied ways, providing different means of fine tuning gene expression.

1.3.3 3′-to-5′ Decay - The Exosome

An alternative pathway of mRNA decay can also occur at the 3′ end of the mRNA (Figure 1.11, right). Like its 5′-to-3′ counterpart, decay via this pathway requires a decapping enzyme, an exoribonuclease, and a number of associated protein factors (reviewed in Coller & Parker 2004). The exosome, the functional equivalent to Xrn1p, is a large complex of exonucleases that digests both nuclear and cytoplasmic mRNAs. Exosome-mediated degradation produces a residual 5′ cap that is then digested by DcpSp. Called "the scavenging decapping enzyme," DcpSp preferentially hydrolyzes shorter ($\leq$ 15nt) RNA substrates, which precludes it from 5′-to-3′ decay (Liu et al. 2002). As this pathway concludes with cap processing, decay factors must interact with the exosome to trigger degradation. Multiple Ski proteins act in lieu of decapping activators, and the loss of even one inhibited turnover of most analyzed transcripts (Anderson & Parker 1998). However, many are dispensable for other exosome-mediated events (e.g., nuclear pre-rRNA processing). This suggests that these proteins primarily function in identifying decay substrates and recruiting the exosome for processing. In fact, they have since been shown to be particularly vital to the rapid turnover of aberrant mRNAs lacking a terminating codon (Section 1.4.3).

Though more clearly discussed as distinct RNA processing events, these two decay

pathways often function simultaneously inside the cell. To tease apart the overlap between them, either direction of decay can be inhibited experimentally. For example, 5′-to-3′ degradation can be blocked by deleting *XRN1*, inserting RNA sequences that block exonuclease progression, or adding translation inhibitors (Section 1.3.5). Analysis of reporter mRNAs showed that the exosome is able to compensate for the loss of 5′-to-3′ decay in any of these conditions (Muhlrad et al. 1995). In wild type cells, however, only products of 5′-digestion were readily visible. Though these results allude to exosome-mediated degradation as simply a secondary pathway, this is not always the case. Decay in the nucleus predominantly occurs via the exosome. This was first shown to be true for pre-mRNAs in yeast, which are stabilized in exosome mutant strains and less so by the absence of the nuclear Xrn1p homologue, Rat1p (Bousquet-Antonelli et al. 2000). Later studies confirmed this is the case for many nuclear non-coding RNAs (reviewed in Moore 2002). Therefore, while pathways are favored differently in the cytoplasm and nucleus, they functionally overlap in both compartments, thereby allowing cells to remain viable when one is impaired.

1.3.4 Deadenylation-Independent Decay Pathways

Though the pathways described above are responsible for degrading the majority of the yeast and human transcriptomes, other pathways can dispose of mRNAs prior to deadenylation. The requirement to remove the poly(A) tail is bypassed when substrates are made vulnerable in some way to the decay machinery. As deadenylation otherwise limits the rate of Xrn1p and exosomal activity (Muhlrad et al. 1994), deadenylation-independent pathways provide a means for rapidly changing gene expression. An example of this happens in yeast during the auto-regulation of *RPS28B* (reviewed in Garneau et al. 2007). At higher concentrations, Rps28B protein binds its own 3′ UTR and recruits Edc3p, a decapping activator, to trigger

decay of the transcript (Figure 1.12, top). Moreover, poly(A)-tailed mRNAs can be rendered susceptible to accelerated decay by several cellular endonucleases. Endonucleolytic cleavage events create substrates for both Xrn1p and the exosome at once (Figure 1.12, bottom). This is often used to limit translation of specific mRNAs during cellular stress responses (reviewed in Garneau et al. 2007). In other instances, endonucleases act on mRNAs when translation is stalled, which accelerates their turnover through the No-Go Decay response (Section 1.4.4).

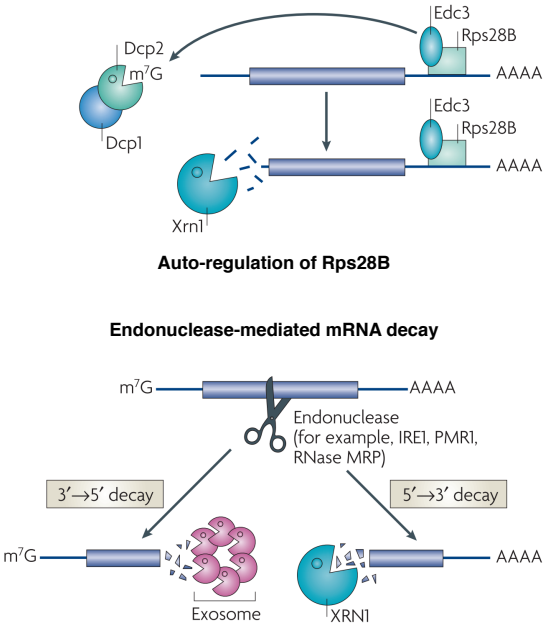

Figure 1.12: Examples of deadenylation-independent mRNA decay. Top, RPS28B recruits Edc3p, a decapping activator, to its 3' UTR while still adenylated. Bottom, endonucleolytic cleavage events lead to bi-directional mRNA decay. Figure adapted from Garneau et al. 2007. Copyright 2007 Nature Publishing Group.

1.3.5 Transitioning From Translation to Decay

Before either the mechanisms or components of mRNA degradation were known, the connection between translation and decay had already been well established. This relationship was first studied by treating *E. coli* with different translation inhibitors (Cremer et al. 1974). mRNAs in the presence of chloramphenicol, an antibiotic that blocks ribosome translocation, were more stable than in wild type

conditions. However, the addition of puromycin, which promotes ribosomal release, caused mRNAs to be more rapidly degraded. Similar changes in mRNA turnover also developed when kasugamycin treatment blocked translation initiation (Schneider et al. 1978). As a whole, these results illustrated that though the ribosome does not function directly in the process of decay, both ribosomal occupancy and active translation prevent mRNA degradation.

Once the decay machinery was identified, the focus shifted to understanding how mRNAs transition from a state of active translation into decay substrates. Within yeast, many of the factors involved in 5′-to-3′ digestion (e.g., Dcp2p, Dhh1p) colocalize with decay intermediates in discrete cytoplasmic mRNPs, called "P bodies" (Sheth & Parker 2003). Deleting any of these proteins individually caused an increase in mRNA stabilization as the foci grew in both number and size. However, P bodies disappeared within minutes of cycloheximide treatment, which locks translocating ribosomes in place, illustrating that only ribosome-free mRNAs enter into the mRNPs. Further studies confirmed this when reduced translation initiation, either from a strong 5′ stem loop or inhibitors, led to enhanced P body formation (reviewed in Franks & Lykke-Andersen 2008). Though translational repression is a prerequisite for mRNP assembly, this is not a permanent fate. Many mRNAs have since been found to exit P bodies and re-enter an active translational state, particularly in response to cellular stress (reviewed in Franks & Lykke-Andersen 2008). As such, mRNAs within P bodies appear to exist at a transitional point between translation and degradation (Figure 1.13).

In order for P bodies to be the primary location for mRNA turnover, it would be essential for decay intermediates to first exist in a ribosome-free state. To assess this possibility, authors enriched for these mRNAs and their associated proteins in Δ*dcp2* and Δ*xrn1* mutant strains. The collected cellular material was separated by sucrose density gradients into lighter (i.e., ribosome-free mRNPs) and heavier

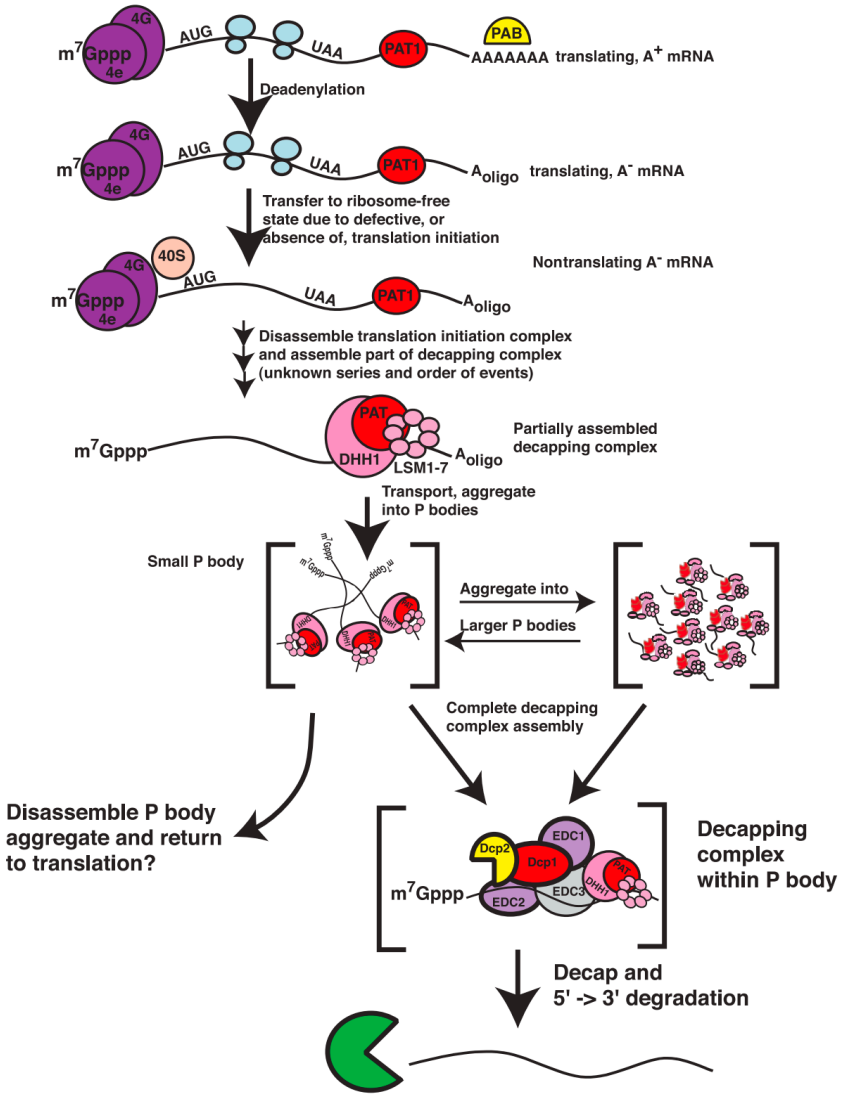

Figure 1.13: An early model for an mRNA's progression from active translation to P body formation and subsequent 5'-to-3' decay. Though some specifics have since been proven incorrect, this figure illustrates the proposed role of P bodies in mRNA degradation. Figure adapted from Coller and Parker 2004. Copyright 2004 Annual Reviews.

(i.e., multiple ribosomes or "polysomes") fractions (Hu et al. 2009). Surprisingly, decay intermediates sedimented with the heavier polysome fraction. Similar sedimentation profiles were later observed for mRNAs harboring early termination codons (Hu et al. 2010), which are known targets of accelerated decay (Section 1.4.1). These results conflicted with the requirement for P body formation prior to degradation, and instead established that the majority of mRNAs undergo decay co-translationally. And yet, precisely how these two processes were connected remained unclear.

Prior to this finding, studies that explored the relationship between translation and decay predominantly concentrated on initiation. However, even the earliest observations on mRNA stability noted that the wide variability in transcript half-lives was only partially attributable to reduced initiation rates (reviewed in Brawerman 1987). Given that transcripts are digested co-translationally, it follows then that changes in elongation and/or termination could also trigger decay. To observe the consequences of modifying elongation rates, rare codons were inserted into the ORF of a reporter mRNA (Sweet et al. 2012). These codons code for infrequently expressed tRNA molecules, and cause stalls in ribosome translocation until the correct tRNA is inserted. Adding even ten rare codons led to a significantly reduced mRNA half-life, but removing Dhh1p (i.e., Δ*dhh1* strain) rescued the effect. It was determined thereafter that codon optimality influences mRNA half-lives transcriptome-wide (Presnyak et al. 2015), and that mRNAs coding for sub-optimal codons are preferentially associated with Dhh1p (Radhakrishnan et al. 2016). Together, these results implicate Dhh1p in monitoring elongation rates in order to initiate decay on translationally inefficient mRNAs. As of now, additional details on the link between translation and normal mRNA decay continue to be investigated.

1.4 Quality Control Pathways

In addition to the general mRNA decay pathway discussed above (Section 1.3), several other surveillance mechanisms exist that target specific subsets of the mRNA population for accelerated decay upon translation. These pathways (Figure 1.14), which include *n*onsense-*m*ediated *d*ecay (NMD, Section 1.4.1), *n*on-*s*top *d*ecay, (NSD, Section 1.4.3) and *no*-*go* *d*ecay (NGD, Section 1.4.4), are collectively known as translation-dependent mRNA decay (TDD). Each is characterized by pathway-

specific mRNA modifications that alter the typical translation rate (Figure 1.14).

These pathways were initially identified as a means by which cells could eliminate problematic mRNAs before they can adversely impact the cell at the protein level. If left unchecked, aberrant mRNAs can cause widespread phenotypic consequences, including disease and cancer (Section 1.2.1.4). More recent work, however, has elucidated the key roles that these mRNA turnover processes also play in post-transcriptional gene regulation. In the sections below is a review of the history of mRNA quality control and how these pathways mediate gene regulation.

1.4.1 Nonsense-Mediated Decay (NMD)

Premature stop codons, introduced either by genomic mutation (nonsense mutation or frameshift) or altered pre-mRNA splicing, can result in the production of dominant negative truncated proteins. In eukaryotic cells, however, most nonsense and frameshift mutations are recessive, loss-of-function alleles. This is due to elimination of the aberrant mRNA by a translation-dependent mRNA decay pathway known as *n*onsense-*m*ediated *d*ecay (NMD).

Rapid decay of mRNAs containing nonsense mutations, which introduce a chain-terminating codon, was first studied in yeast (Losson & Lacroute 1979). Pulse labeling experiments revealed that the wild type mRNA and an isoform with a *p*remature *t*ermination *c*odon (PTC) were synthesized at the same rate. The decay rate of the PTC-containing transcript, however, was greatly increased. Moreover, this change in stability showed a position-dependent effect, as isoforms with PTCs near the 5′ end of the *o*pen *r*eading *f*rame (ORF) were turned over faster than those with PTCs by the 3′ end. So fast, in fact, that an isoform with an early PTC was undetectable at steady state.

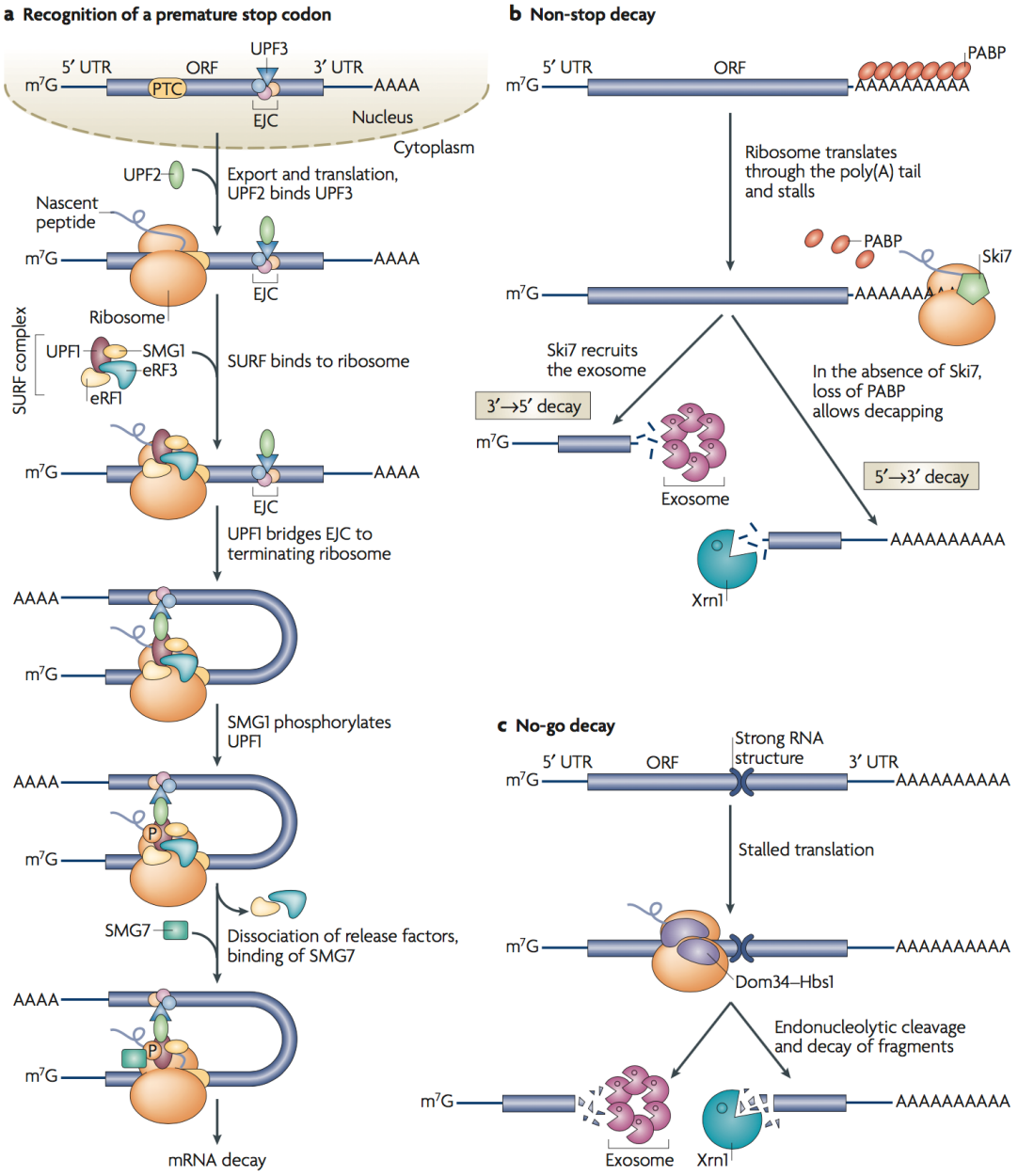

Figure 1.14: Mechanisms of mammalian translation-dependent mRNA decay (TDD): (a) Nonsense-mediated decay (NMD), Section 1.4.1; (b) Non-stop decay (NSD), Section 1.4.2; (c) No-go decay (NGD), Section 1.4.3. Figure from Garneau et al. 2007. Copyright 2007 Nature Publishing Group.

NMD transcripts are identified by a number of conserved protein factors, many of which were initially determined through genetic screening experiments in *C. elegans* and *S. cerevisiae* (Hodgkin et al. 1989; Pulak & Anderson 1993; Leeds et al. 1991, 1992). The NMD machinery is inessential in both yeast and nematodes, providing ideal models to map out the pathway and proteins responsible. In nematodes, however, loss of activity does lead to adverse cellular effects. Conse-

quently, NMD factors were first discovered in *C. elegans* because loss of function mutants led to physical abnormalities while stabilizing known PTC-containing mRNAs (Hodgkin et al. 1989). As such, the six genes involved were named *smg-1* through *smg-6* for *s*uppressor with *m*orphological effect on *g*enitalia. As for yeast, the core proteins involved in the NMD pathway are Upf1, Upf2, and Upf3 – named after the *Up-f*rameshift mutant isoforms that were stabilized upon their deletion (Leeds et al. 1991). Homologous proteins were later identified in higher order species (reviewed in Hug et al. 2015).

In yeast, Upf1, an RNA-dependent ATPase, associates with release factors, eRF1 and eRF3, and Upf2 on nonsense-containing isoforms in the cytoplasm (Figure 1.14a) (Mühlemann et al. 2008). Translation termination leads to interaction with Upf3 and the phosphorylation of Upf1 by Smg1, causing dissociation of both release factors and the ribosome. Additional Smg proteins subsequently recruit general mRNA decay factors responsible for decapping and 5′-to-3′ exonucleolytic digestion (Hug et al. 2015). Recent studies in yeast have shown that these isoforms remain associated with polyribosomes even after decapping occurs, indicative of the strong relationship between translation and NMD (Hu et al. 2010).

Along with the identification of the responsible protein factors, early genetic experiments in *C. elegans* revealed that PTC-containing transcripts are not the sole target of the NMD pathway (Pulak & Anderson 1993). Loss of *smg* proteins stabilized the *unc-54(r293)* mutant mRNA, which is nearly 2kb longer than its wild-type counterpart. This extension is the result of a deletion in the *unc-54* 3′ *un*t*r*anslated *r*egion (UTR) that removes the wild-type polyadenylation site. This phenomenon was later studied in yeast (Muhlrad & Parker 1999), where the destabilization of extended mRNAs was linked to the length of the 3′ UTR rather than the presence of a specific sequence. Studies of mammalian 3′ UTR variants have determined this regulation is conserved (Bühler et al. 2006). Through var-

ious tethering experiments, it was shown that this particular NMD response is controlled by interactions between *p*oly(*A*)-*b*inding *p*rotein (PABP) and eRF3, a release factor, at the termination codon (reviewed in Mühlemann 2008). If the 3′ UTR is extended, PABP is deposited too far to compete with Upf1 to bind eRF3 during termination, leading to rapid degradation of the transcript by the NMD machinery.

Though designation as an NMD-regulated transcript typically happens during mRNA processing (i.e., mutations giving rise to PTCs, errors in splicing extending UTRs), nonetheless an mRNA's fate can be changed during translation itself. While stalled at a pseudoknot or other pause-inducing secondary structure, the ribosome may translocate at a "slippery site" of nucleotides to a different frame through a single tRNA slippage (reviewed in Dinman 2012). Much like alternative splicing (Section 1.2.1), *−1 P*rogrammed *R*ibosomal *F*rameshifting (-1 PRF) is a means of expanding the transcriptome given a limited genome size. Though first identified in multiple viral genomes, computational analyses showed that at least one strong -1 PRF signal exists in nearly 10% of all yeast genes (Jacobs et al. 2007). Moreover, the use of $\geq$ 95% of these signals would result in termination at a previously out-of-frame PTC, providing an opportunity for rapid decay of mRNAs prone to erroneous translocation. Conservation of these rates was seen across more than 20 eukaryotic transcriptomes (Belew et al. 2008).

As genome-wide experiments become more ubiquitous, it has become increasingly obvious that the NMD-regulated transcriptome is not strictly limited to mRNAs as initially believed. Loss-of-function mutants *upf1-1* and *upf3-1* in *Arabidopsis* stabilized a class of mRNA-like non-coding RNAs in a genome-wide tiling array (Kurihara et al. 2009). Similar results were obtained through deep sequencing experiments performed in both mouse embryonic stem cells and yeast (Hurt et al. 2013; Smith et al. 2014). mRNAs containing *u*pstream-*ORFs* (uORFs) and *l*ong

*n*on-*c*oding *RNAs* (lncRNAs) increased in abundance when NMD was impaired in these systems. NMD regulation of lncRNAs suggests that these transcripts are present in the cytoplasm where they undergo some level of translation; this was confirmed by ribosome profiling and polyribosome analysis (Smith et al. 2014). It is now believed that mRNAs with uORFs or lncRNAs containing short ORFs mimic transcripts with extended 3′ UTRs and may be regulated by NMD in a similar fashion (Hurt et al. 2013; Smith et al. 2014).

1.4.1.1 Role of the Exon Junction Complex in NMD

Although the NMD machinery is conserved in all eukaryotes, the pathway in higher order species involves a major protein complex not present in *S. cerevisiae*: the EJC (Section 1.1.3). While expendable for NMD in *C. elegans* and *Drosophila* (Mühlemann et al. 2008), the EJC is an essential part of mammalian NMD. Depletion of eIF4AIII, one of the core EJC proteins described above, stabilized known NMD targets in HeLa cells (Shibuya et al. 2004); MLN51 depletion showed a similar effect (Palacios et al. 2004). Moreover, loss of the conserved C terminal end of Upf3, where Y14 binds, causes a loss of NMD in tethering experiments (Gehring et al. 2003), indicating this bridge between the NMD machinery and the EJC is required. The mammalian EJC proteome was analyzed via mass spectroscopy and Western blots, confirming association with numerous regulatory and splicing factors (including Upf3) (Singh et al. 2012).

One of the major contributions of the EJC to NMD is in the identification of "premature" termination codons. EJCs within the coding region are necessarily removed by the ribosome during the first, or "pioneer," round of translation (Section 1.1.3.2). As the ribosome translocates along the mRNA, it collides with the EJC and other RNA-binding proteins, thus only proteins bound downstream of the

first in-frame termination codon remain (Dostie & Dreyfuss 2002). Any remaining EJCs bound more than 50 to 55 nucleotides downstream of the termination codon serve as strong enhancers for NMD (reviewed in Maquat 2004). This significantly expands the NMD-regulated transcriptome in higher-order species beyond mRNAs simply containing a nonsense mutation, long UTR, or a strong -1 PRF signal. In the absence of all of these aforementioned triggers of NMD, splicing events that leave bound EJCs in the 3′ UTR can shuttle mRNAs into rapid decay by the NMD machinery.

1.4.2 NMD as a General Post-Transcriptional Regulatory Pathway

Thus far, NMD has been narrowly discussed as simply a means of clearing the cell of aberrantly translated RNAs, but it is also a key post-transcriptional gene expression regulation pathway. The simplest example of this occurs when -1 PRF is coupled to NMD. The wild-type and NMD-regulated transcript are one and the same, until a substantial pause in translation causes abrupt decay. Gene expression of certain cytokine receptors in mammalian cells is believed to be regulated in this way (Belew et al. 2014). When a microRNA binds to the cytokine receptor mRNA, a stable pseudoknot is formed downstream of a slippery stretch of nucleotides, consequently introducing a PTC through -1 PRF. A rise in the transcription of the microRNA increases clearance of the mRNA through this NMD-causing binding event, thereby regulating the immune response to viral cytokines.

Nevertheless, NMD-regulated gene expression most often results from alternative splicing events that introduce PTCs during nuclear mRNA processing. As in the initial discovery of the NMD machinery, *C. elegans* provided the ideal system to explore questions about how post-transcriptional gene expression is regulated

through *a*lternative *s*plicing coupled to *NMD* (AS-NMD). Products of the *smg* genes were first linked to alternative pre-mRNA processing of two SR protein mRNAs (Morrison et al. 1997). Genes encoding these mRNAs also each code for an alternatively spliced isoform with an in-frame stop codon; these AS-NMD transcripts are stabilized in *smg(-)* mutants. Similar results were observed in later studies of the alternative splicing of ribosomal proteins (Mitrovich & Anderson 2000). In both cases, when *smg* genes were mutated, AS-NMD transcript abundance was more than or equal to that of the wild type isoform, indicative of a high level of unproductively spliced mRNAs in wild type cells.

Conservation of gene regulation via AS-NMD was evident even in early constructions of the human transcriptome (Lewis et al. 2003). Based on alignments of EST sequences to a RefSeq annotation, AS-NMD was initially estimated to regulate one-third of the approximately 16,000 mRNAs. Homologs of the aforementioned *C. elegans* SR and ribosomal proteins were among this pool, suggesting conserved regulation of their expression. In fact, subsequent EST studies showed that all human SR protein genes contain ultraconserved regions that act as poison cassette exons when included (Lareau et al. 2007). Exon inclusion has since been found to act as a conserved method of autoregulating the homeostatic gene expression of many *trans*-acting splicing factors, including SR proteins (Ni et al. 2007).

The true extent of post-transcriptional regulation via AS-NMD was not known until transcriptomes were assembled from deep sequencing datasets. Analysis based on EST sequences is inherently biased towards stable mRNAs, precluding many short-lived regulatory transcripts (Lewis et al. 2003). *RNA-Seq*uencing (RNA-Seq) experiments introduce their own issues (discussed further in Section 1.6), but the sheer depth of information allows for improved recovery of AS-NMD-regulated transcripts. Genome-wide analysis after knocking down *p*olypyrimidine *t*ract-*b*inding *p*rotein 1 (Ptbp1), a splicing repressor, revealed its role in regulating

expression by controlling 5′ and 3′ *s*plice *s*ite (SSs) choice on many pre-mRNAs (Hamid & Makeyev 2014). Among these genes, AS-NMD controls the regulation of proteins involved in organelle biogenesis (Hamid & Makeyev 2014), neural development (Zheng et al. 2012), and chromatin modification (Yan et al. 2015). Further transcriptome-wide experiments have since revealed numerous pathways regulated by AS-NMD in *Arabidopsis* (Drechsel et al. 2013), *C. elegans* (Longman et al. 2013), and *S. cerevisiae* (Kawashima et al. 2014).

1.4.3 Non-Stop Decay (NSD)

Post-transcriptional regulation is also achieved through *N*on-*S*top *D*ecay (NSD), another TDD pathway conserved from yeast to humans (reviewed in Klauer & Hoof 2012). NSD specifically degrades transcripts that fail to terminate translation. This typically arises from mutations in the normal stop codon, products of aborted transcription events, and premature polyadenylation due to cryptic poly(A) signals. Like NMD, NSD was first identified in *S. cerevisiae* (Frischmeyer et al. 2002; Van Hoof et al. 2002). A reporter mRNA without a stop codon was rapidly degraded in mutant strains lacking either the NMD machinery (*upf1*Δ) or the general 5′-to-3′ decay machinery (*xrn1*Δ or *dcp1-2*) (Frischmeyer et al. 2002). Deleting components of the exosome (*ski2*Δ, *ski3*Δ, *ski7*Δ, or *ski8*Δ), however, stabilized the NSD transcript (Van Hoof et al. 2002). The C terminal half of Ski7 is required for this stabilization. This region resembles two GTPases, EF1A and eRF3, that interact with the A site of the ribosome during translation. When ribosomes reach the end of the poly(A) tail rather than properly terminating, Ski7 is thought to bind to the empty A site and trigger decay by the exosome (Figure 1.14b). As Ski7 is not conserved, the mammalian exosome instead interacts with Hbs1L and Pelota (the mammalian homologs of Hbs1 and Dom34, respectively) to recognize and eliminate NSD transcripts (Saito et al. 2013).

1.4.4 No-Go Decay (NGD)

The three phases of translation are initiation, elongation, and termination. Early termination events trigger NMD, and failure to terminate prompts degradation by NSD. The rate of elongation is monitored by the final major TDD pathway, *N*o-*G*o *D*ecay (NGD). During translation, ribosomes may encounter strong secondary structures (i.e., stem-loops and pseudoknots) or rare codons that stall further translocation. Introducing any of these features into reporter mRNAs decreased their steady-state abundance in yeast (Doma & Parker 2006). Full-length mRNAs containing a stem-loop in the ORF were rescued by impairing either the 5′-to-3′ or the 3′-to-5′ decay pathways. The mutants also stabilized RNA fragments 3′ or 5′ of the stem-loop, respectively. Both fragments, however, were missing in *dom34*Δ and *hbs1*Δ strains. In addition to their role in mammalian NSD (Section 1.4.2), Pelota/Dom34 and Hbs1L/Hbs1 bind stalled ribosomes and cause subunit dissociation, thereby activating NGD (Shoemaker et al. 2010). Neither, however, is the endonuclease responsible for creating the aforementioned RNA fragments (Passos et al. 2009). Although this protein has yet to be identified, recent experiments showed that endonucleolytic cleavage only occurs when ribosomes are stacked within the ORF (Simms et al. 2017). Once cleaved, mRNA fragments are degraded by both the exosome and general decay machinery (Figure 1.14c).

1.4.5 NSD/NGD as General Post-Transcriptional Regulatory Pathways

Unlike NMD, evidence of NSD and NGD regulating post-transcriptional gene expression is limited. Due to the overlapping functions of Pelota/Dom34 and Hbs1L/Hbs1, the identification of naturally occurring transcripts targeted by these pathways has largely occurred in parallel. Transcriptome-wide studies in *dom34*Δ

and *hbs1*Δ yeast strains identified *HAC1* mRNA as a regulated substrate (Guydosh & Green 2014). *HAC1* is exported to the cytoplasm while still containing its first intron. The mRNA is spliced under cellular stress, allowing translation of a stress-specific transcription factor (Harigaya & Parker 2012). Otherwise, under normal conditions, endonucleolytic cleavage at the 5′ splice site leaves a truncated *HAC1* transcript. This short mRNA is up-regulated in the knockout strains. Either improper termination at the 5′ end of the exon or the resulting ribosome stalling upstream may trigger degradation via NSD or NGD, respectively (Guydosh & Green 2014).

Though *HAC1* is the only known mRNA to be regulated in this manner in *S. cerevisiae*, many such substrates have been identified in *S. pombe*. In this yeast, a stress response occurs when unfolded or misfolded proteins accumulate at the *e*ndoplasmic *r*eticulum (ER) (Guydosh et al. 2017). The unfolded protein response activates signal transduction pathways that fold proteins while also limiting further translation. Like *HAC1*, mRNAs at the ER are cleaved and rapidly degraded. Of the nearly 500 mRNAs regulated in this manner, the vast majority (91%) contain an ER-specific signal sequence or transmembrane domain. This indicates that NSD and NGD play a significant role in homeostasis of this organelle. Further studies of similar stress responses are needed to determine if this regulation is conserved mammalian cells.

1.5 Transcriptome Annotation

A full understanding of how TDD contributes to the regulation of gene expression requires a transcriptome annotation that is both complete and accurate. Although the human transcriptome is mentioned here and elsewhere as a specific entity, there is striking variability between annotations. This reflects the diversity

of gene expression, much of which is dictated by factors such as cell type and environmental conditions. Further, transcriptome annotation is highly dependent on the processes used to derive the annotation and in the decisions of what are "true" or "functional" transcripts and what are not.

There are currently four major human transcriptome annotations or "references": RefSeq (Pruitt et al. 2005), Ensembl (Curwen et al. 2004), GENCODE (Harrow et al. 2012), and CHESS (Pertea et al. 2018). Although there is substantial overlap, each has its own idiosyncrasies, resulting in thousands of reference-specific annotations. To illustrate the extent of overlap and differences between references, the number of exon junctions with shared or unique annotations is depicted in Figure 1.15. Notably, of the approximately 550,000 annotated junctions, only 45% are found in all four references reviewed below.

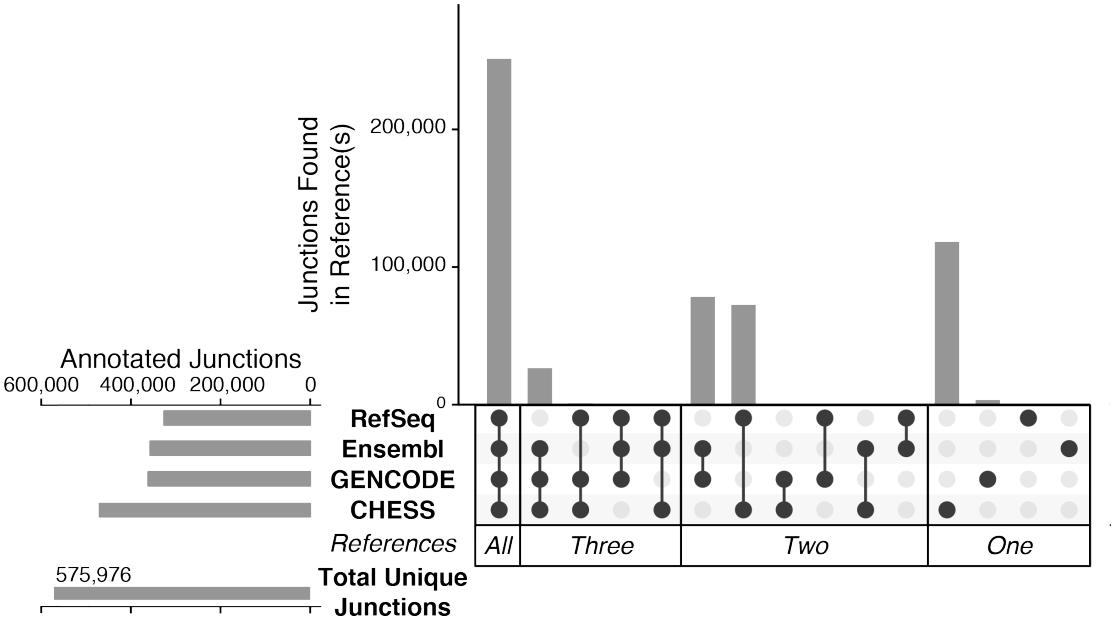

Figure 1.15: Comparison of annotated exon junctions among the transcriptomes sourced from RefSeq (hg38), Ensembl (GRCh38.p12), GENCODE (v29), and CHESS (v2.1). Horizontal bars: total junctions in each reference set; vertical bars: intersections of indicated reference sets. Bar graphs created with UpSetR.

1.5.1 RefSeq

The reference human transcriptome with the smallest number of exon junctions (Figure 1.15) is RefSeq (for *Ref*erence *Seq*uence), curated by the *N*ational *C*enter for *B*iotechnology *I*nformation (NCBI) at the US National Institutes of Health (Pruitt et al. 2005). The RefSeq database contains a collection of the latest genome, transcript, and protein annotations, which are frequently updated in each "release." As of release 93 (available in March 2019), RefSeq covered 88,816 organisms, including but not limited to: bacteria, fungi, viruses, vertebrates, and plants (O'Leary et al. 2016). Hereafter, my focus will be on the human annotations provided by RefSeq.

Transcriptome assembly by RefSeq involves a combination of automatic and manual annotation processes (Pruitt et al. 2014). Much of the RefSeq human transcriptome is based on publicly available RNA-Seq datasets processed through the NCBI Eukaryotic Annotation Pipeline. Datasets are sourced from both the *I*nternational *N*ucleotide *S*equence *D*atabase *C*onsortium (INSDC) and, more recently, the *S*hort *R*ead *A*rchive (SRA), and chosen based on depth and lack of tissue-specificity or treatment to ensure individual-to-individual consistency. Annotation of the non-coding transcriptome (e.g., pseudogenes, micro-RNAs), however, has typically been performed by other sources, such as miRBase, and then incorporated into RefSeq (Pruitt et al. 2014).

RefSeq is the most conservative in annotation among the discussed references, a result of their unique goal to provide a non-redundant collection of full-length transcript and protein sequences (O'Leary et al. 2016). Consequently, an mRNA annotated in RefSeq is likely true, however many real transcripts are omitted. Therefore the absence of a transcript from RefSeq should not be interpreted as proof it does not exist. This is readily visible when comparing RefSeq to

other databases (Table 1.1). One reason for the lower number of transcripts in RefSeq is its systematic under-representation of alternative isoforms. Rather than annotate multiple short isoforms, transcripts with mutual first and final exons are extended to share the same 5' and 3' ends (Frankish et al. 2015) (Figure 1.16, top transcripts). Thus RefSeq under-reports alternative transcription start and polyadenylation sites (TSS and PAS, respectively). With regard to protein-coding transcripts, RefSeq often labels many as "non-coding" (or "NR") due to features such as 3′ UTR exon junctions regardless of evidence indicating their productive translation (reviewed in Bicknell et al. (2012)). Other examples of transcripts that were previously left out or labeled as NR include those encoding proteins containing selenocysteine (which is encoded by UGA within a particular sequence context) and transcripts that engage in -1 ribosomal frameshifting. Recent efforts by manual curators have begun to address some of these deficiencies (Rajput et al. 2019), but this remediation work remains incomplete.

Table 1.1: Features of the human genome and transcriptome in each annotation source.

Feature	RefSeq	Ensembl	GENCODE	CHESS
Protein-coding Genes	20,070	20,418	19,940	21,306
Non-coding Genes	17,710	22,107	23,643	21,856
Transcripts	113,224	206,762	206,694	323,824
Exon Junctions	325,855	356,956	361,387	469,743

1.5.2 Ensembl and GENCODE

Though maintained by different sources, Ensembl and GENCODE [a *g*enome research project at ENCODE (*ENC*yclopedia *O*f *D*NA *E*lements)] are best described

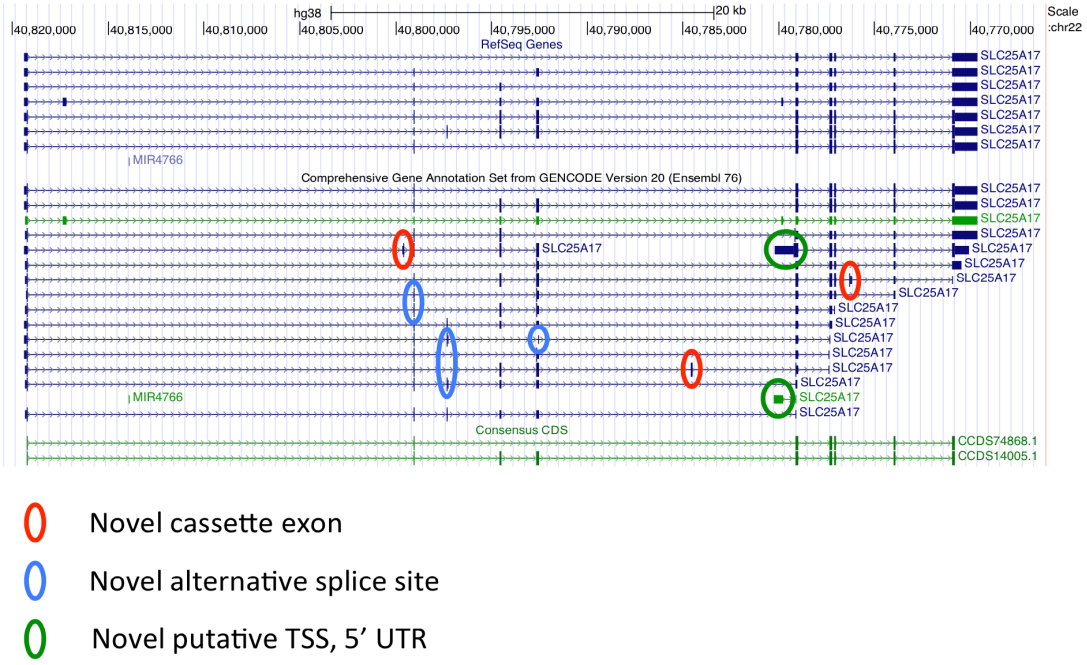

Figure 1.16: Comparison of GENCODE and RefSeq Annotation in the UCSC genome browser displaying the SLC25A17 locus. Novel GENCODE splicing features are highlighted in red (novel cassette exons), blue (novel alternative splice site, or shifted splice site) and green (novel putative TSS and 5' UTR). Figure from Frankish et al. 2015. Copyright 2015 BioMed Central.

together because of the high degree of overlap in their annotation process and resulting transcriptomes. Similar to RefSeq, Ensembl provides annotations for a large variety of organisms, however GENCODE focuses solely on humans and mice. The European Bioinformatics Institute curates the Ensembl database through an automatic pipeline, known as their Ensembl Annotation Process (Down et al. 2002). Once assembled, Ensembl annotations are merged with manual annotations from the HAVANA (*H*uman *a*nd *V*ertebrate *An*alysis and *A*nnotation) group to form the GENCODE reference (Harrow et al. 2012). The two transcriptomes end up being nearly identical, with differences primarily located in duplicated regions and pseudogenes.

Although data for the Ensembl Annotation Process pipeline is sourced from RNA-Seq libraries in the INSDC (Aken et al. 2016), just as is RefSeq, the resulting Ensembl and GENCODE transcriptomes are much larger than RefSeq (Table 1.1). As previously mentioned (Section 1.6), some level of variation between the an-

notations is expected. A major reason for this difference is how each database represents alternative events. Rather than extending an isoform to the full-length transcript, annotation requires support from biological evidence at each nucleotide (Frankish et al. 2015). This results in many more short isoforms in the Ensembl- and GENCODE-annotated transcriptomes, increasing the number of transcripts per gene to 4.8 and 4.7, respectively, compared to 3.0 in RefSeq (from Table 1.1). Included in these isoforms are a number of alternative processing and splicing events (Figure 1.16). Although this may include some false positives from shorter RNA-Seq reads, it allows these transcriptomes to more accurately depict the full extent of alternative splicing in the human transcriptome.

1.5.3 CHESS

The Salzberg lab recently attempted to address the discrepancies between RefSeq, GENCODE, and other databases by assembling more complete and accurate human genome and transcript annotations, dubbed CHESS (*C*omprehensive *H*uman *E*xpressed *S*equence*S*, Pertea et al. (2018)). The annotation process, again, hinges on RNA-Seq data, this time originating from a massive *G*enotype-*T*issue *Ex*pression (GTEx) study (The GTEx Consortium et al. 2015) instead of the INSDC. CHESS was built from 9,725 sequencing libraries with a total depth just shy of 900 billion reads (Pertea et al. 2018). Predicted annotations were strictly filtered by several criteria, including conservation and reproducibility, to eliminate potential alignment artifacts and splicing noise. Though the genome increased a marginal amount, the CHESS transcriptome nearly tripled compared to RefSeq (Table 1.1). Still, more than 80,000 exon junctions annotated by other references, predominantly Ensembl and GENCODE, did not appear in the CHESS transcriptome (Figure 1.15). Many of these junctions are found in non-coding transcripts, and were likely lost to computational filters designed to remove non-

functional, low abundance transcripts (Pertea et al. 2018). This indicates, however, that even the most comprehensive analysis of RNA-Seq data cannot capture all known splicing events.

1.5.4 Other Transcriptome Annotations

In addition to RefSeq, Ensembl, GENCODE, and CHESS, there are other human transcriptome annotations available. Other studies have compared these references to the four described in detail here. The UCSC Known Genes annotation, the basis for the UCSC Genome Browser, performs similarly to RefSeq in mapping analysis (Zhao & Zhang 2015). The Moore lab has historically used RefSeq for bioinformatic analyses so it was chosen between the two. A broader study of annotations included transcriptomes from the H-InvDB Genes, AceView Genes and VegaGenes databases (Wu et al. 2013). VegaGenes is now merged with Ensembl/GENCODE. The others are no longer updated.

1.5.5 Impact of Annotation Choice on Data Analysis

With the advent of deep sequencing in the 2000's, RNA-Seq rapidly emerged as the method of choice for assessing differential gene expression. The outcome of any differential gene expression experiment is highly dependent, however, on what reference is used to align the RNA-Seq reads. For short RNA-Seq reads ($\leq$ 50 nt), alignment to the genome of interest can be sufficient. On the other hand, reads that span exon junctions will have low mapping scores and so will likely be excluded from the analysis. With reads of 75 nts, for example, Zhao reported a 33 - 37 percent difference in the fraction of reads mapping when using only the human genome as a reference versus the genome plus the RefSeq transcriptome (Zhao 2014). Almost all of the reads that failed to map in the genome-only align-

ment spanned one or more exon junctions. As new RNA-Seq technologies allow for increasingly longer sequenced reads, this discrepancy becomes even larger and the likelihood that a read crosses one or more exon junctions rises dramatically with read length. That is not to say that merely having a reference transcriptome is sufficient. Incomplete or inaccurate annotations are known to affect the quantification of alternatively spliced isoforms (Pyrkosz et al. 2013). Fortunately, the sources detailed above (Sections 1.5.1-.3) continue to be updated with the most recent annotations ("releases"), so the question when studying the human transcriptome has become a matter of *which* database to use.

Unfortunately, the question of which database to use is rarely answered on an experiment-specific basis, though recent work has shown that it ought to be (Wu et al. 2013). In that study, both RefSeq and Ensembl annotations (and others mentioned in Section 1.5.4) were ranked by their complexity, a calculation based on the number of genes, transcripts, and exons. Ensembl ranked at higher complexity than RefSeq in every category. The latest numbers, including counts of exon junctions, agree with this finding (Figure 1.15 and Table 1.1). Based on the effect complexity had on both mapping and differential gene expression, the authors concluded that less complex annotations (i.e., RefSeq) are sufficient when studying differential expression at the gene-level (Wu et al. 2013). Accurate estimates at the transcript-level, particularly for non-coding and/or regulatory RNAs, however, required more complex transcriptomes like Ensembl. A later comparison between RefSeq and GENCODE led to similar conclusions (Frankish et al. 2015). Therefore, reference annotations need to be selected based on overall experimental design rather than simply a matter of convenience.

Nevertheless, recall that the majority of exon junctions are absent from at least one reference (Figure 1.15), meaning that no one current annotation is complete. Merging multiple transcriptomes, a crude attempt at a more "complete" anno-

tation, was shown to improve transcript quantification in comparison to either parent reference (Chen et al. 2013). Still, filters based on abundance thresholds during the annotation process have led to the absence of known regulatory RNAs with cell- and/or condition-specific expression profiles, including many related to disease (Morillon & Gautheret 2019). In fact, analysis of 21,500 RNA-Seq libraries from SRA identified more than three million exon junctions (each with ≥ 20 reads per junction) in the human transcriptome (Nellore et al. 2016). This is nearly a ten-fold increase compared to RefSeq, which also uses data from SRA (Pruitt et al. 2014). Additional data from single-molecule and short-read RNA-Seq experiments targeting weakly expressed RNAs suggest that GENCODE could be missing half of the alternative isoforms transcribed from non-coding gene loci (Deveson et al. 2018). While these counts may be inflated by false positive exon junctions, many unannotated regulatory RNAs unquestionably exist and should be taken into account when considering the human transcriptome.

1.6 Identifying TDD Transcripts in Mammalian Systems

As discussed in Section 1.4, TDD of alternatively processed transcripts has been well established as a key post-transcriptional regulatory process in higher eukaryotes. When included in annotations, these transcripts are typically labelled by their method of degradation (e.g., "nonsense-mediated decay"), which provides one way to estimate the prevalence of TDD regulation. In the latest human transcriptome from Ensembl, for example, nearly 15,000 transcripts are marked as known NMD substrates (Cunningham et al. 2019). Yet this value cannot represent the full extent of TDD regulation as many transcripts have expression profiles linked to specific conditions (e.g., development, cell differentiation, disease) and are thus are often missing from transcriptome annotations altogether. Many stud-

ies have therefore attempted to identify additional TDD targets by manipulating conditions to slow their decay.

One of the first methods for identifying TDD substrates was depletion of selenium, an essential trace element. When selenium is present, UGA codons upstream of a ~60 nt hairpin (known as a selenocysteine insertion sequence, or SECIS, element) are decoded as selenocysteine codons and a full length selenoprotein is translated (Figure 1.17) (Moriarty et al. 1998). When the cell is deficient in selenium, however, these codons are recognized as stop codons. When early in the coding region, they are essentially PTCs that lead to NMD of the mRNA. Selenium depletion has been exploited to determine/confirm key mechanistic details of NMD, including but not limited to: (1) NMD occurs in the cytoplasm (Moriarty et al. 1998); (2) NMD requires translation; (3) NMD depends on intron position near PTCs (Sun et al. 2000); and (4) degradation can be cell type-specific (Sun et al. 2001). Although the regulation of mRNAs encoding selenoproteins has been shown to play a role in cellular processes such as immune function, muscle development, and fertility, there only about 300 such mRNAs annotated in the human transcriptome (Rajput et al. 2019).

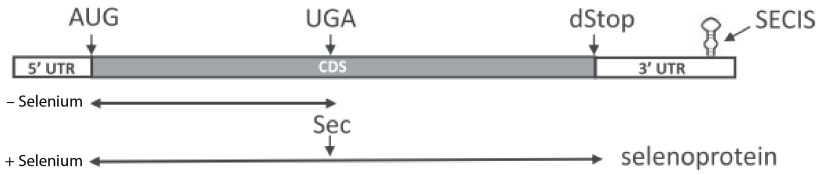

Figure 1.17: Schematic of mRNA that can be translated as either a selenoprotein or a NMD-regulated transcript depending on how the ribosome decodes the indicated UGA codon. Figure adapted from Rajput et al. 2019. Copyright 2019 Oxford University Press.

As translation is required for decay, any inhibitor that blocks protein synthesis will also stabilize TDD-regulated mRNAs. An early study demonstrated that expression of a NMD reporter substrate was up-regulated in the cytoplasm upon addition of any of six different translation inhibitors, regardless of their inhibition mechanism (Carter et al. 1995). These experiments helped to confirm the link

between translation and NMD. Since then, a number of small molecule inhibitors that directly affect translation and/or decay protein factors have been identified. For instance, phosphorylation of Upf1, a required step to initiate NMD, can be blocked by the addition of either NMDI-1 or caffeine in human cell lines (Keeling et al. 2013). Both small molecules have been employed in a method dubbed *G*ene *I*dentification by *N*MD *I*nhibition (GINI) (Noensie & Dietz 2001). Potential cancer-related NMD substrates were identified using GINI by comparing the transcriptomes of normal and disease cell lines to find caffeine-induced stabilization (Johnson et al. 2012). However, the high degree of discordance between findings from different studies and a high number of false positives (reviewed in Johnson et al. 2012) has hindered the wide-spread adoption of GINI as a general NMD substrate identification method.

Eliminating one or more protein factors required for decay (Figure 1.14) is a powerful approach to identifying TDD targets. Investigations into NMD as a regulatory pathway in mammalian systems were initially limited by the fact that the Upf proteins are essential in these species. This eliminates the option of simply deleting these genes and monitoring which transcripts are consequently stabilized, a common approach used to study NMD in yeast and nematodes. In cell lines, mRNAs encoding the NMD machinery can, however, be depleted by transfecting targeted *s*hort *i*nterfering *RNAs* (siRNAs). Cells remain viable as siRNA-mediated knockdown is efficient yet not absolute. Using microarrays, reduced expression of Upf proteins was shown to affect ($\geq$ 1.9-fold change) between 4 and 9% of transcripts analyzed (Mendell et al. 2004; Wittmann et al. 2006). Common features of NMD-regulated transcripts (e.g., long 3′ UTRs, uORFs) were made apparent by comparisons of the microarray-based transcriptome profiles after transfecting different siRNAs (Yepiskoposyan et al. 2011). Moreover, *G*ene *O*ntology (GO) analysis showed genes involved in alternative splicing or the NMD pathway itself

were consistently up-regulated across all NMD-knockdown transcriptomes.

All of the above transcriptome-wide studies relied on microarrays. While microarrays are adaptable and can be designed to target specific mRNAs, such as PTC-containing splice variants (Pan et al. 2006), the reach of this method is limited to known and abundant transcripts. The switch from microarrays to RNA-Seq in the late 2000's expanded not only the scope of detectable transcripts, but also the questions that can be answered. For example, reporter mRNAs with extended 3′ UTRs are known to be regulated by NMD (discussed in Section 1.4.1). It was not known, however, what causes these mRNAs to undergo NMD rather than the general decay pathway. Data from CLIP-Seq, a targeted immunoprecipitation followed by RNA-Seq, revealed UPF1 binding events are, in fact, biased transcriptome-wide towards 3′ UTRs that are longer than average (Hurt et al. 2013). This same bias is seen amongst up-regulated transcripts in Upf1 knockdown cells (Colombo et al. 2017). Further bioinformatic analysis of these UTRs identified a GC-rich motif thought to stall Upf1 translocation, increasing the likelihood of it becoming phosphorylated by Smg1 and triggering NMD (Imamachi et al. 2017).

Although much important insight has been gained from the approaches discussed above, they all fall short of a true representation of the initial or "pre-translational" transcriptome (i.e., the complete set of transcripts generated in the nucleus prior engagement by ribosomes). Under wild-type conditions, mRNA isoforms vary greatly with regard to cytoplasmic degradation rate. As such, the currently available methods can only provide a static snapshot of the transcriptome while revealing little about the flux through mRNA processing pathways. While individual decay pathways can be slowed by the methods discussed above, long-term inhibition can result in confounding pleiotropic effects. For example, simply inhibiting NMD activates autophagy in mammalian cells, which in turn leads to a marked increase in cell death in less than 48 hours (Wengrod et al. 2013). Yet siRNA

treatments typically take much longer than this for adequate depletion (Mendell et al. 2004; Wittmann et al. 2006). Furthermore, all of the commonly used translation inhibitors are known to activate other signaling pathways (Sidhu & Omiecinski 1998). Transcriptomes assembled from these experiments, therefore, are influenced to an unknown degree by TDD-independent changes in gene expression.

1.7 The Pre-translational Transcriptome

TDD-regulated transcripts exist for a very limited timeframe in wild-type cells. To better capture these transcripts, methods are needed to specifically select for the population of mRNAs that have completed splicing but have not yet been translated. This eliminates the influence of varied translation-dependent decay rates. Transcriptomes assembled from pre-translational mRNAs should thus provide a more accurate record of the flux through various alternative processing pathways.

1.7.1 Isolation of Specific RNA Populations

Specific populations of RNAs have been previously purified using isolation methods that enrich for either RNA-RNA or RNA-protein interactions. Full length mRNAs, for example, can be selected using oligo(dT) primers that hybridize to long stretches of adenosines. As general decay initiates upon deadenylation and other RNAs (i.e., ribosomal RNAs) are not adenylated, these probes will primarily bind to the poly(A) tails of fully processed mRNAs not yet subject to deadenylation. Contaminating RNAs (i.e., non-mRNAs) are often depleted during RNA-Seq library preparations with this method (Sultan et al. 2014; Lykke-Andersen et al. 2014; Saudemont et al. 2017). Although the oligonucleotide sequence can be

changed to purify or deplete other RNA species, a significant number of probes are necessary unless the targeted sequence is shared (e.g., RiboZero).

An alternate approach to enriching for RNA species with a particular property, but whose sequences are unknown is to pull down an *R*NA *b*inding *p*rotein (RBP) along with its associated RNAs. In addition to sequence motifs, different RBPs recognize secondary structures (e.g., hairpins) and other structural elements (e.g., 5' cap or poly(A) tail). Many of these proteins are involved in post-transcriptional gene regulation, from pre-mRNA processing (Section 1.1) to decay (Section 1.4). As such, knowing the mRNAs targeted by these proteins can help answer many questions about how regulation is mediated. Bound mRNAs can be isolated and identified using *R*NA *i*mmuno*p*recipitations (RIP) (Tenenbaum et al. 2000). The combination of RIP with high-throughput sequencing (RIP-Seq) expands this analysis to a transcriptome-wide scale (Zhao et al. 2010; reviewed in Singh et al. 2014). For information on precisely where RBPs bind RNA, RIP experiments may include a nuclease treatment step, which destroys unbound and exposed RNA. Remaining RNA fragments are known as "footprints."

However, the interaction of many RBPs with RNA is dynamic. Consequently, RBPs may dissociate during RIP. Though some remain unbound, many will re-associate with other RNA fragments in the cell lysate and create new footprints that were not present in intact cells (Mili & Steitz 2004). Crosslinking agents (i.e., UV or formaldehyde) induce covalent bonds between RNAs and proteins at their site of interaction. The links formed by UV crosslinking prior to RIP, known as *c*ross*l*inking and *i*mmuno*p*recipitation (CLIP), also allow for the use of more stringent purification conditions (Ule et al. 2003). For example, harsher washing of bound material removes indirect protein-RNA interactions, increasing the signal-to-noise ratio. Since its creation in 2003, many other CLIP-based techniques (e.g., CLIP-Seq, PAR-CLIP, iCLIP) have been developed, each modifying

specific step(s) in the methodology (reviewed in Hannigan et al. 2018). CLIP-Seq, the deep sequencing of CLIP RNA, helped identify the mechanism behind Upf1-mediated decay of long 3′ UTRs (Section 1.6; Hurt et al. 2013; Imamachi et al. 2017).

CLIP methods provide exquisitely precise information as to the location of RNA-protein interactions that are amenable to the crosslinking approach used. However, they can only reveal the crosslinking locations for individual proteins, and not multi-protein complexes. While many RBPs do function as individuals [e.g., Nova (Ule et al. 2003); Staufen (Ricci et al. 2014)], many others function as part of larger complexes (e.g., spliceosomes). These larger complexes contain both proteins that directly contact the RNA and other more distal components whose locations cannot be captured by CLIP. In addition, because the composition of many macromolecular complexes that interact with RNA is highly dynamic (e.g., the spliceosome cycle), just knowing where a particular protein binds provides no information as to which complex that protein is contained in. One method for purifying these macromolecular complexes is to perform sequential RIPs using different complex-specific proteins. This method, called *R*NA:protein *i*mmuno*p*recipitation *i*n *t*andem (RIPiT) (Singh et al. 2012, 2014), has been used to study RBPs and their complexes in both yeast and human cell lines (Singh et al. 2012; Ricci et al. 2014; Chen et al. 2014, 2018; Yang et al. 2014; Mabin et al. 2018). The differences in purified material between a RIP-based approach and RIPiT are shown in Figure 1.18.

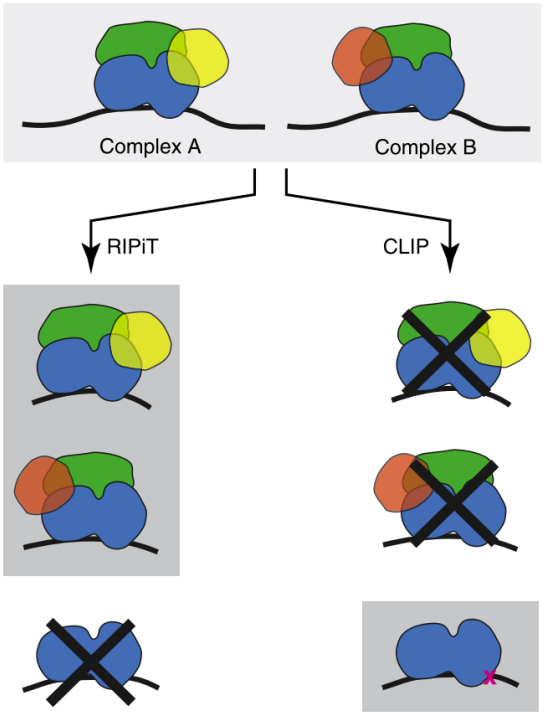

Figure 1.18: Differences in outputs from RIPiT and CLIP experiments. Blue: RNA-binding proteins; green: non-RBPs; red/yellow: complex-specific proteins. RIPiT reveals complex-specific information, whereas CLIP can only show sites of RNA-protein interactions. Figure from Singh et al. 2014. Copyright 2014 Elsevier.

1.7.2 Accessing the Pre-Translational Transcriptome through the EJC

The ideal target for accessing the pre-translational transcriptome is the exon junction complex (EJC). Late in the splicing cycle, EJCs are deposited on mRNAs upstream of exon-exon junctions in a sequence-independent manner (Le Hir, Izaurralde, et al. 2000; Shibuya et al. 2004). The EJC remains stably bound until a translating ribosome forces its removal in the cytoplasm (Section 1.1.3). Thus, the EJC-associated transcriptome records nuclear processing events prior to translation-dependent regulation.

The Moore lab has previously used RIPiT to preferentially purify pre-translational mRNPs from mammalian cell lines (Figure 3.4; Singh et al. 2012; Metkar et al. 2018). To do so, inducible transgenes encoding individual EJC core proteins with

FLAG epitope tags were stably integrated (Flp-in) into HEK293 cells. Tetracycline titration allows for FLAG-tagged protein expression at near endogenous levels. Complexes of interest can then be purified via RIPiT. To accumulate EJC-bound mRNAs, cells can be treated with a translation inhibitor (i.e., cycloheximide) for a short time (1 hour) prior to lysis. EJC-containing mRNPs are then isolated with an initial IP targeting the FLAG-tagged protein followed by a second IP using an antibody for a different endogenous EJC protein. Between the two purification steps, limited RNase I treatment of the precipitated material digests unprotected RNA, leaving EJC footprints. The final eluate contains both the proteins and mRNA regions associated with pre-translational mRNPs for further study.

The first transcriptome-wide analysis of EJC-associated RNAs confirmed many known properties of the EJC, and also provided surprising insight into EJC deposition (Singh et al. 2012). As expected, the EJC footprint is centered around -24 nt upstream of exon-exon junctions with a width of approximately 14 nt. This area is referred to as the *c*anonical *EJC* (cEJC) deposition site. At least 80% of exon junctions were protected by bound EJCs. GO analysis confirmed mRNAs with fewer occupied cEJC sites were not tied to any particular gene class. Conversely, cEJC occupancy was highly enriched (~8-fold) on mRNAs known to be regulated by AS-NMD (Section 1.4.1.2). These same transcripts were also more likely to be bound by EJCs at *n*oncanonical *EJC* sites (nEJCs) in nearby exonic locations.

In addition to short (~12-25 nts) footprints at cEJC and nEJC locations, longer (~30-150 nt) RNA fragments were protected from RNase I digestion even under very stringent digestion conditions. These longer footprints were associated with larger complexes that contained both core EJC proteins and other mRNA binding proteins (Singh et al. 2012). Notably, the broader EJC proteome contained an

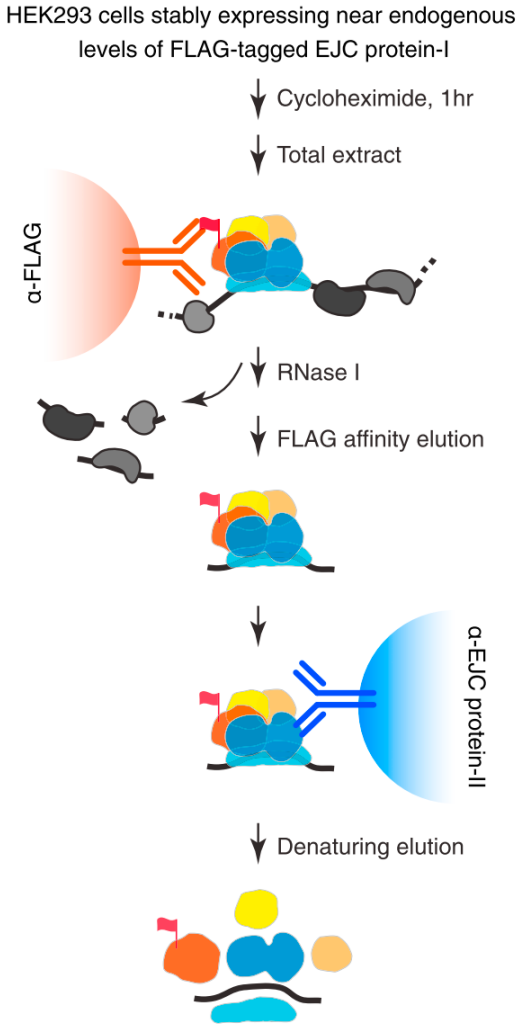

Figure 1.19: RIPiT strategy for isolating EJC-associated mRNPs. Figure from Singh et al. 2012. Copyright 2012 Elsevier.

unexpected number of SR and SR-like proteins with known functional roles in splicing, export, and translation (reviewed in Singh et al. 2012). Associations with these proteins are believed to stabilize the larger EJC interactome that forms the pre-translational mRNP.

The organization of pre-translational mRNPs was further explored using a modified version of RIPiT called RIPPLiT for *R*NA *i*mmuno*p*recipitation and *p*roximity *l*igation *i*n *t*andem (Figure 1.20, Metkar et al. 2018). For RIPPLiT, RNase digestion conditions during the first IP are adjusted to produce a broad fragment distribution between 30 and >500 nts. During the second IP step, tandem phos-

phatase and kinase reactions produce 5' and 3' ends appropriate for ligation by T4 RNA ligase 1, which joins spatially-adjacent RNA ends, creating chimeras. Deep sequencing of the resulting RNA (~200-550 nt fragments) and alignment of resulting reads to the genome using a custom algorithm, ChimeraTie (Metkar et al. 2018), revealed the locations of chimeric junctions. For abundant ncRNAs with known 3D structures (e.g., contaminating 5.8S, 18S and 28S rRNAs), these chimeric junctions were consistent with previously determined structures. On mRNAs, chimeric junctions were exclusively intramolecular and evenly distributed, even across long exons. No interactions, however, were found between the 5′ and 3′ ends of mRNAs. Combined with scaling analysis, which examines the relationship between junction frequency and nucleotide distance, these observations suggest pre-translational mRNPs form flexible rod-like structures (Figure 1.21).

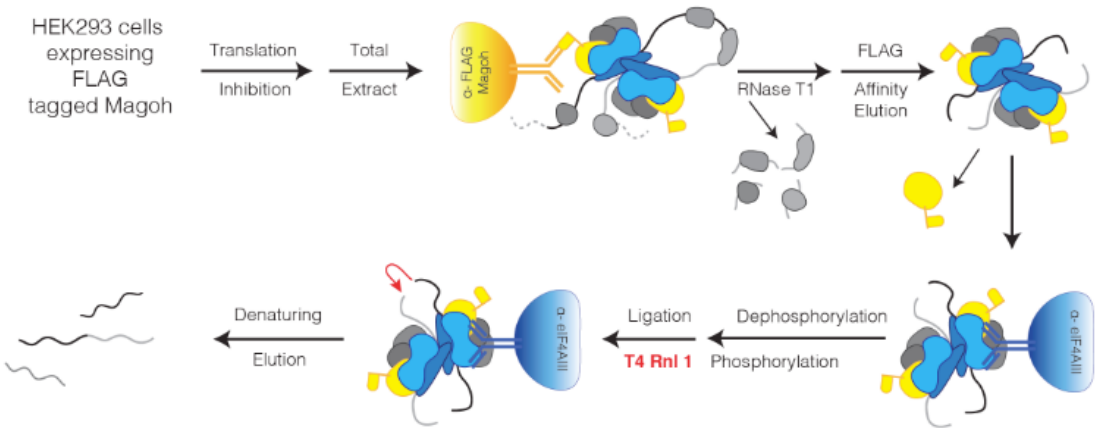

Figure 1.20: RIPPLiT strategy for capturing pre-translational mRNPs. Yellow/blue: EJC proteins; grey: non-EJC proteins. Figure from Metkar et al. 2018. Copyright 2018 Elsevier.

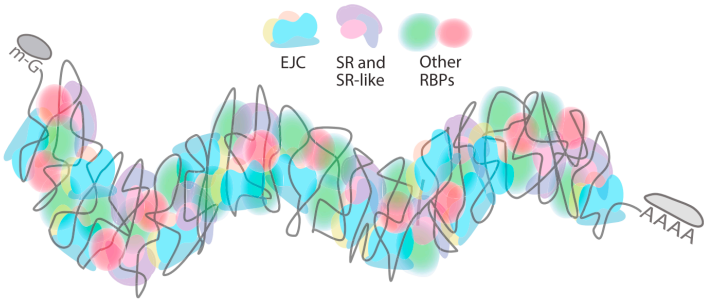

Figure 1.21: Predicted structure of the pre-translational mRNP. Figure from Metkar et al. 2018. Copyright 2018 Elsevier.

A key control in RIPPLiT-Seq experiments is the construction of libraries from RNase-treated samples not subsequently incubated with RNA ligase. The EJC RIPPLiT-Seq study included three (-) ligase control replicates (Metkar et al. 2018). In my dissertation work, I used these - ligase control libraries, along with published RNA-Seq libraries (Sultan et al. 2014), for an in-depth exploration of the pre-translational HEK293 transcriptome. As the median internal human exon length is 187 nts (Ensembl), the long insert (~200-550 nt) and read (220 nt paired end) lengths of the (-) ligase libraries were sufficient for quantifying known splicing events and searching for unannotated ones. Chapter 2 within this disseration contains the manuscript describing my analysis of these pre-existing libraries to assess how the pre-translational record differs between EJC RIPPLiT and standard RNA-Seq.

In addition to this body of work, I also spent a considerable amount of time analyzing libraries derived RNAs associated with from late-stage spliceosomes. Much like when isolating EJC-containing mRNPs, the spliceosome was selected for by first targeting FLAG-tagged Magoh followed by an endogenous core spliceosome protein. Although this work was not published, early analysis determined that the spliceosome-associated transcriptome contained numerous examples of previously unannotated splicing events similar to the pre-translational EJC-associated transcriptome. Our cursory investigation into this transcriptome and these events is described in detail in Chapter 3.

Chapter 2

Deep sequencing of pre-translational mRNPs reveals hidden flux through evolutionarily conserved AS-NMD pathways

2.1 Preface

The contents of this Chapter have been published previously as:

Kovalak C., Metkar M., and Moore M. J. (2019). Deep sequencing of pre-translational mRNPs reveals hidden flux through evolutionarily conserved AS-NMD pathways. bioRxiv. doi: https://doi.org/10.1101/847004.

2.2 Introduction

A central mechanism underlying metazoan gene expression is alternative pre-mRNA processing, which regulates the repertoire of mRNA isoforms expressed in various tissues and under different cellular conditions. Extensive deep sequencing of RNA (RNA-Seq) has revealed that ~95% of human protein-coding genes are subject to alternative splicing (AS) (Eric T Wang et al. 2008; Pan et al. 2008), with current estimates suggesting ~82,000 different protein-coding mRNA isoforms generated from ~20,000 protein coding genes (Cunningham et al. 2019). Thus, production of alternative mRNA isoforms massively expands the protein repertoire that can be expressed from a much smaller number of genes (Nilsen & Graveley 2010; Kelemen et al. 2013). But cells also need to control how much of each protein is made. Although transcriptional control is often considered the predominant mechanism for modulating protein abundance, emerging evidence indicates that post-transcriptional regulatory mechanisms are crucial as well.

Not all mRNA variants are protein-coding. Nearly 15,000 human mRNAs in the Ensembl database (release 93) are annotated as nonsense-mediated decay (NMD) targets (Cunningham et al. 2019). NMD is a translation-dependent pathway that both eliminates aberant mRNAs with malformed coding regions (i.e., those containing premature termination codons due to mutation or missplicing) and serves as a key mechanism for maintenance of protein homeostasis (Kurosaki et al. 2019). This protein homeostasis function is mediated by AS linked to NMD (AS-NMD), wherein the flux through alternate splicing pathways that result in protein-coding and NMD isoforms is subject to tight control (Lewis et al. 2003). These NMD isoforms harbor a premature termination codon either due to frameshifting or inclusion of a poison cassette exon. Because NMD isoforms are rapidly eliminated after the first or "pioneer" round of translation, only protein-coding isoforms re-

sult in appreciable protein production (Figure 2.1, bottom). Thus increasing or decreasing flux through the NMD splicing pathway decreases or increases protein production, respectively. Although AS-NMD was originally described as a mechanism by which RNA binding proteins (e.g., SR and hnRNP proteins) could autoregulate their own synthesis, recent work indicates that AS-NMD is much more pervasive, tuning abundance of many other proteins such as those involved in chromatin modification and cellular differentiation (Nasif et al. 2018).

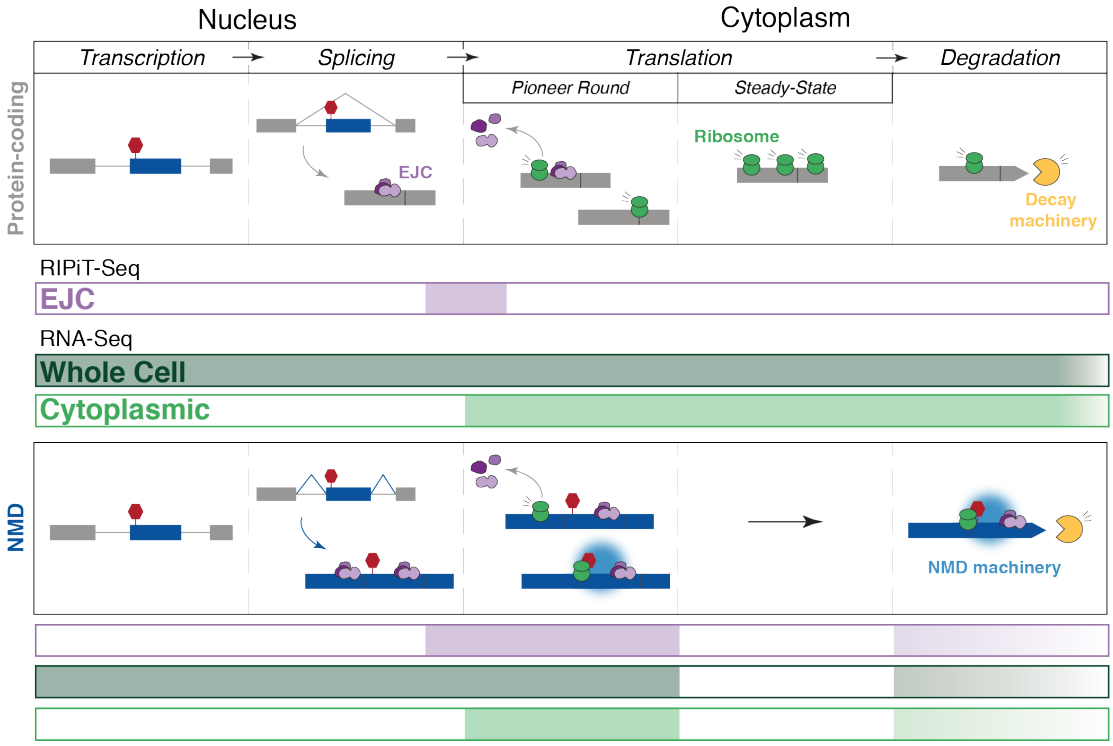

Figure 2.1: (Top) mRNA metabolism from transcription to degradation. In this illustration, poison exon skipping and inclusion lead to a Protein-coding isoform (grey) and NMD isoform (blue), respectively, with the NMD isoform containing a premature stop codon (red). EJCs (purple) deposited upstream of exon junctions are cleared by ribosomes during the pioneer round of translation. While Protein-coding isoforms are subject to multiple rounds of translation prior to decay, NMD isoforms are rapidly eliminated. (Bottom) Libraries analyzed in this paper: EJC-bound RIPiT-Seq (purple), whole cell (dark green) and cytoplasmic (light green) RNA-Seq. Colored bars indicate RNA populations captured in each library type. Figure from Kovalak et al. 2020.

The true extent to which AS-NMD contributes to protein homeostasis can only be appreciated by determining the flux through the protein-coding and NMD splicing pathways. Transcriptome-wide assessment of mRNA isoform abundance generally relies on RNA-Seq of whole cell or cytoplasmic RNA. Such methods provide a static

snapshot of the species present in the sample at the time of collection. Because NMD isoforms are so rapidly decayed, they are generally underrepresented in RNA-Seq datasets. Thus a single RNA-Seq snapshot is generally uninformative as to synthetic flux through protein-coding and NMD splicing pathways.

An alternate means to assess protein-coding and NMD pathway flux is to capture newly synthesized mRNAs after splicing completion but prior to translation. Late in the splicing cycle, the exon junction complex (EJC) is deposited upstream of at least 80% of exon-exon junctions (canonical; cEJCs) and multiple other sites throughout the length of spliced exons (noncanonical; ncEJCs) (Singh et al. 2012; Saulière et al. 2012). Upon nucleocytoplasmic export, the pioneer round of translation removes EJCs within the 5′ UTR and CDS regions, with EJCs remaining downstream of stop codons being key mediators of NMD (Maquat et al. 2010). Pre-translational mRNPs can be selectively isolated by tandem immunoprecipitation of epitope-tagged and untagged EJC components, a technique known as RNA:protein immunoprecipitation in tandem (RIPiT) (Singh et al. 2014). Deep sequencing library preparation from RIPiT samples (RIPiT-Seq) has previously enabled us to map the positions of canonical and noncanonical EJCs on spliced transcripts (Singh et al. 2012) and to investigate the RNA packing principles within pre-translational mRNPs (Metkar et al. 2018).

Here, we compare libraries from pre-translational mRNPs (EJC RIPiT), unfractionated RNA (whole cell RNA-Seq) and RNA post subcellular fractionation (cytoplasmic RNA-Seq) (Figure 2.1). As expected, EJC RIPiT libraries are enriched for transcript isoforms destined for translation-dependent decay. By providing a window into the repertoire of transcripts generated by splicing but prior to translation-dependent decay, EJC RIPiT libraries provide a more accurate record of the flux through various alternative processing pathways than does standard RNA-Seq. Importantly, EJC RIPiT libraries enabled us to identify numerous new

evolutionarily-conserved poison cassette exons that had previously eluded annotation based on even highly extensive RNA-Seq data analyses.

2.3 Results

2.3.1 EJC, whole cell and cytoplasmic libraries

In our recent study investigating the organizing principles of spliced RNPs (Metkar et al. 2018), we generated three biological replicates from HEK293 cells of EJC-bound RNAs partially digested with RNase T1 during RNP purification (Figure 2.1). Paired-end deep sequencing of these EJC RIPiT libraries resulted in 19-25 million mate pairs each (Table 2.1). For comparison to RNA-Seq libraries, we chose rRNA-depleted whole cell and cytoplasmic HEK293 RNA-Seq datasets (two biological replicates each) previously published by Sultan et al. (Sultan et al. 2014). We chose these particular libraries based on their similarity in cell treatment and library preparation to our EJC libraries, their clean cellular fractionation, and sequencing depth (51-57 million mate pairs each).

Library			Sequencing				Alignment			
Name	*Fraction*	*Replicate*	*Type*	*Insert Size*	*Sequenced Pairs*	*Repeats*	*Aligned Pairs*	*MAPQ ≥ 5*	*Unique Pairs*	*Spliced Reads*
EJC	Total	1	Paired End, 150bp	220 - 500	19 Million	– 3 M	6 M	5.6 M	5.1 M	3.3 M (32%)
		2		220 - 500	25 M	– 4 M	10 M	9.0 M	8.2 M	5.6 M (34%)
		3		220 - 500	23 M	– 4 M	7 M	6.6 M	6.1 M	4.3 M (35%)
RNASeq	Whole Cell	1	Paired End, 51bp	100 - 200	57 M	– 15 M	30 M	28.0 M	24.9 M	5.8 M (12%)
		2		100 - 200	55 M	– 15 M	38 M	36.0 M	29.7 M	7.4 M (12%)
RNASeq	Cytoplasm	1	Paired End, 51bp	100 - 200	56 M	– 8 M	46 M	33.4 M	33.4 M	11.7 M (18%)
		2		100 - 200	51 M	– 6 M	43 M	32.6 M	32.6 M	11.5 M (18%)

Table 2.1: Sequencing and alignment information for each replicate of the analyzed libraries. Figure from Kovalak et al. 2020.

All libraries were downloaded from their respective repositories (see Declarations) and processed in parallel. Reads were aligned to the Genome Reference Consortium Human Build 38 (GRCh38.p12) (Cunningham et al. 2019) using STAR (v2.5.3a) (Dobin et al. 2013) after first filtering out those mapping to repeat RNAs. To minimize the effect of misalignment in ensuing analyses, mismatches were lim-

ited to three per read, with gaps caused by deletions or insertions being strongly penalized. These strict mapping parameters resulted in 6-10 million and 30-43 million aligned pairs for the EJC RIPiT and RNA-Seq libraries, respectively (Table 2.1). For quantification, we limited all analyses to unique reads with high mapping quality (MAPQ ≥ 5). For all libraries, we used Kallisto (v0.44.0) to derive expression values for the ~200,000 annotated transcripts in GRCh38.p12 (Cunningham et al. 2019). Examination of per-transcript abundance revealed high concordance ($\geq$ 0.93 to 0.99) among all biological replicates (Figure 2.2). Therefore, all subsequent quantitative analyses utilized merged biological replicate data.

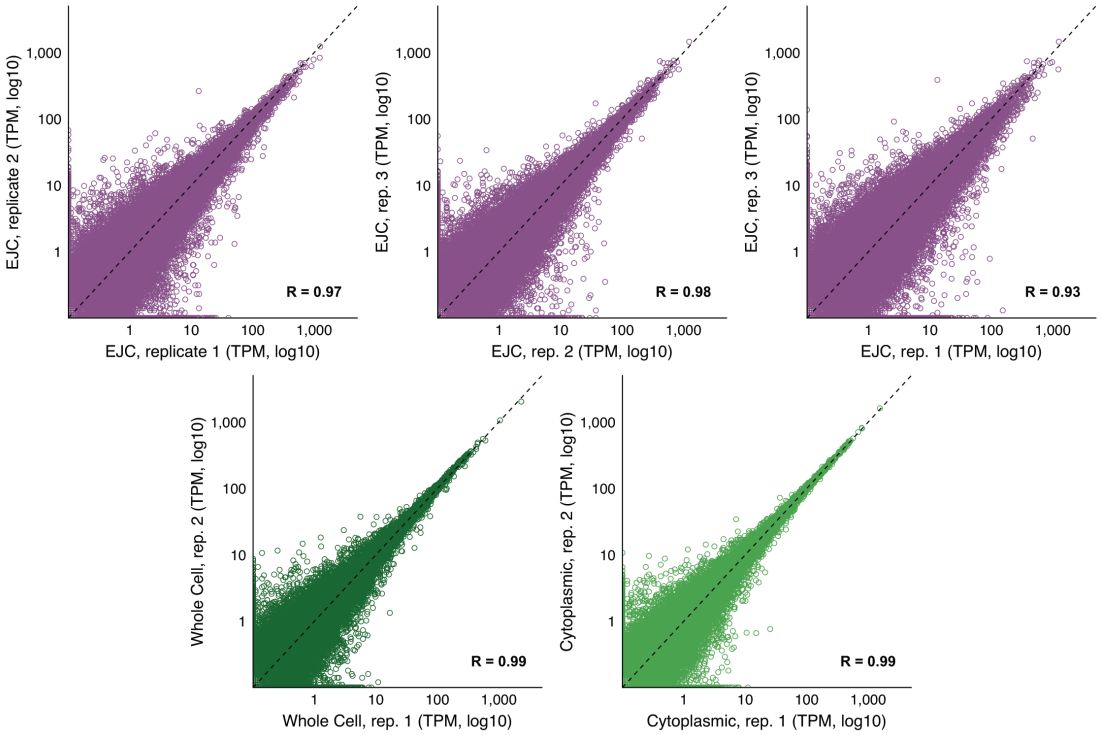

Figure 2.2: Scatterplots comparing transcripts per million (TPM) between replicates of the same library type. R: Pearson's correlation. Figure from Kovalak et al. 2020.

2.3.2 EJC LIBRARIES ARE ENRICHED FOR SPLICED TRANSCRIPTS AND TRANSLATION-DEPENDENT DECAY TARGETS

To assess the relative representation of NMD targets in EJC and RNA-Seq libraries, we first examined read coverage on known AS-NMD genes. The SR pro-

teins TRA2B and U2AF2 negatively regulate their own expression by promoting inclusion of a highly-conserved poison cassette exon containing a premature termination codon (Figure 2.3, A-C). Although these poison exons were detectable in all library types, they were much more abundant in the EJC libraries. As expected due to NMD, cytoplasmic RNA-Seq libraries exhibited the lowest poison exon inclusion (percent spliced in; PSI) values (16% and 4%, respectively), with the whole cell libraries being somewhat higher (29% and 6%, respectively). Yet, the EJC RIPiT libraries indicate much higher inclusion percentages (averaging 94% and 73%, respectively). Thus, for both TRA2B and U2AF2, the predominant splicing pathway in HEK293 cells under standard growth conditions is poison exon inclusion. Similar trends were observed for other known AS-NMD targets (Figure 2.4 and 2.5), including hnRNPA1 where the AS-NMD isoform results from splicing in the 3′UTR as a consequence of alternative polyadenylation (Figure 2.4, A). The substantial differences between the EJC RIPiT and RNA-Seq quantitations for these previously documented AS-NMD isoforms clearly illustrate the advantage provided by the EJC RIPiT libraries for more accurately assessing flux through alternative processing pathways that result in mRNA isoforms with widely different decay rates.

In GRCh38.p12, every transcript isoform is given a specific annotation; relevant annotations in protein-coding genes are “protein-coding”, “NMD”, “NSD”, “retained intron”, and “processed transcript”, with the latter being a catch-all for transcripts not clearly attributable to any other category. NSD (non-stop decay) is another translation-dependent mRNA degradation pathway that eliminates transcripts having no in-frame stop codon (Klauer & Hoof 2012). When exported to the cytoplasm, transcripts containing one or more retained introns are also usually subject to translation-dependent decay due to the presence of in-frame stop codons in intronic regions. For transcripts detectable in our libraries [TPM >0

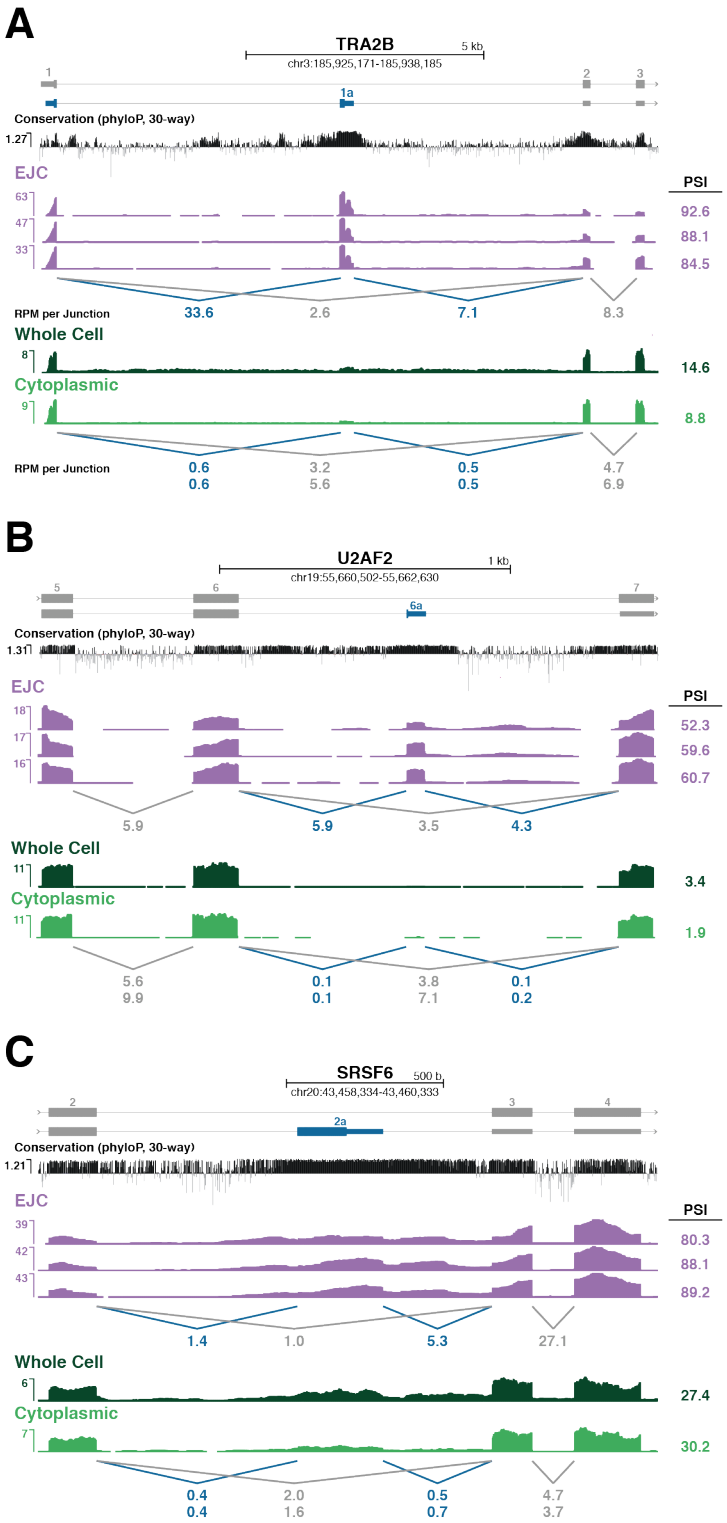

Figure 2.3: Genome browser tracks of library coverage across individual genes (grey: protein-coding isoform(s); blue: NMD isoform) containing poison cassette exons (A, TRA2B; B, U2AF2; C, SRSF6). Shown are all three EJC RIPiT replicates and replicate 1 for whole cell and cytoplasmic RNA-Seq. Conservation tracks show phyloP basewise scores derived from Multiz alignment of 30 vertebrate species. Numbers below tracks indicate mean reads per million (RPM) spanning each exon junction. Numbers to right in B and C are percent spliced in (PSI) values for poison exon inclusion events; PSI values for RNA-Seq libraries are replicate means. R: Pearson's correlation. Figure from Kovalak et al. 2020.

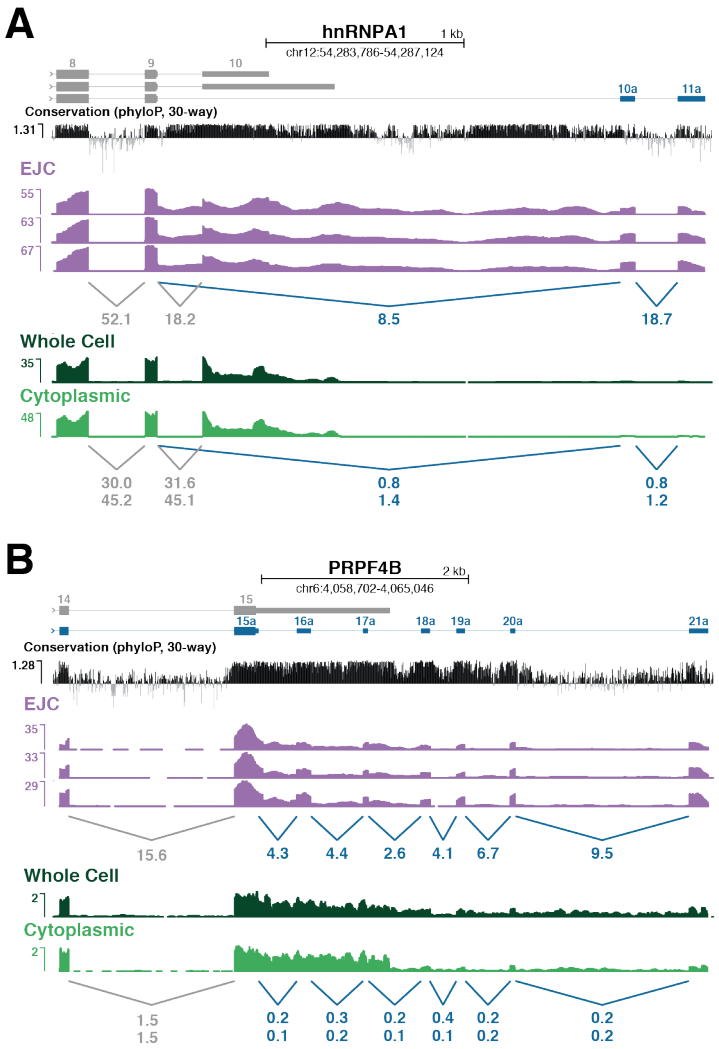

Figure 2.4: Genome browser tracks of library coverage across individual genes (grey: protein-coding isoform(s); blue: NMD isoform) containing 3' UTR introns (A, hnRNPA1; B, PRPF4B).

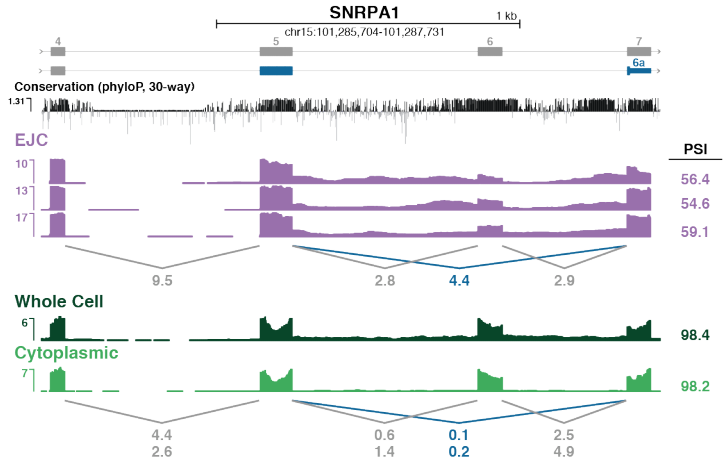

Figure 2.5: Genome browser tracks of library coverage across individual genes (grey: protein-coding isoform(s); blue: NMD isoform) containing an exon skipping event.

in all replicates of a particular library type (EJC, whole cell or cytoplasmic)], the number of exon junctions (i.e., positions at which introns were removed) per protein-coding and NMD isoform ranged from 0 to >100 and 1 to 69, respectively (Figure 2.6A). As expected, protein-coding isoforms having no exon junctions were less abundant in the EJC libraries than in either RNA-Seq library (Figure 2.6B, top). In contrast, spliced protein-coding isoforms containing 5 or more exon junctions were enriched in EJC libraries, with the degree of enrichment increasing with exon junction number. For each exon junction number bin (i.e., 1-4, 5-10 and 10+), NMD isoforms were even more enriched in EJC libraries than were protein-coding isoforms (Figure 2.6B, bottom). EJC library enrichment was also readily discernible in per-transcript scatter plots for NMD, NSD, retained intron, and processed transcript isoforms (Figure 2.7). All of these observations are consistent with the notion that EJC-associated RNAs are enriched for spliced transcripts subject to subsequent elimination by translation-dependent decay.

Because they are not translated, long intergenic non-coding RNAs (lincRNAs) are not subject to translation-dependent decay. As expected, lincRNAs lacking exon junctions (e.g., MALAT1, RMRP, NEAT1 and NORAD) were substantially depleted from EJC libraries, whereas those containing exon junctions were of similar or higher abundance in EJC than RNA-Seq libraries (Figure 2.8). Particularly notable was XIST, the most highly represented Pol II transcript in our EJC libraries (Metkar et al. 2018). XIST is both spliced and exclusively nuclear. Reflecting this, median abundance of the eight major XIST isoforms was five- and forty-fold greater in EJC than in whole cell and cytoplasmic libraries, respectively.

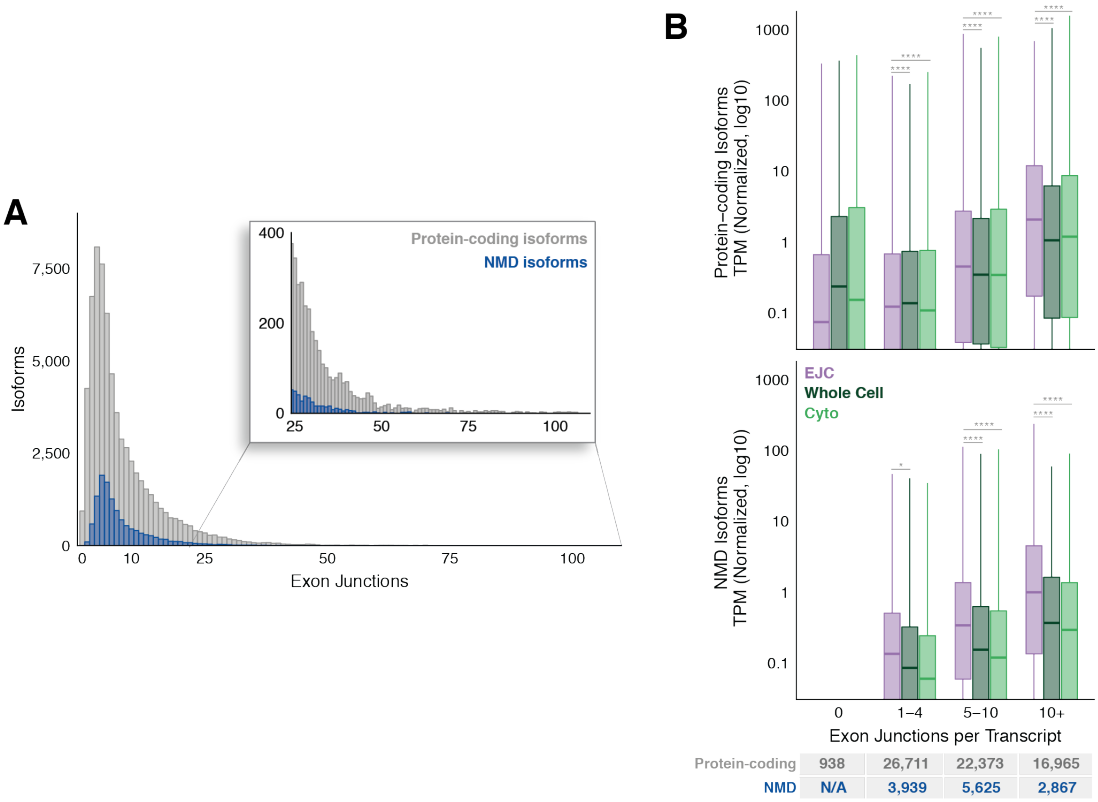

Figure 2.6: (A) Distribution of the number of exon junctions in all annotated protein-coding (grey) or NMD (blue) transcripts. (B) Distribution of protein-coding (top) and NMD (bottom) transcripts per million (TPM) in each library type (colors as in Figure 1A), binned based on indicated number of exon junctions per transcript. Results of one-way ANOVA and Tukey's post hoc significance tests comparing EJC RIPiT-Seq to RNA-Seq libraries are indicated: *P<0.05, **P<0.01, ***P<0.005, ****P<0.0001. Figure from Kovalak et al. 2020.

2.3.3 EJC libraries capture new exon junctions

Having established that spliced transcripts known to be eliminated by translation-dependent decay are enriched in EJC libraries, we next wondered whether EJC libraries might contain new transcript isoforms that had previously eluded detection due to their low abundance in RNA-Seq. Such isoforms should contain previously unannotated exon junctions. To identify all previously annotated exon junctions, we integrated the RefSeq (hg38) (O'Leary et al. 2016), Ensembl (GRCh38.p12) (Cunningham et al. 2019), GENCODE (v29) (Frankish et al. 2019) and Comprehensive Human Expressed SequenceS (CHESS) transcriptome annotations to

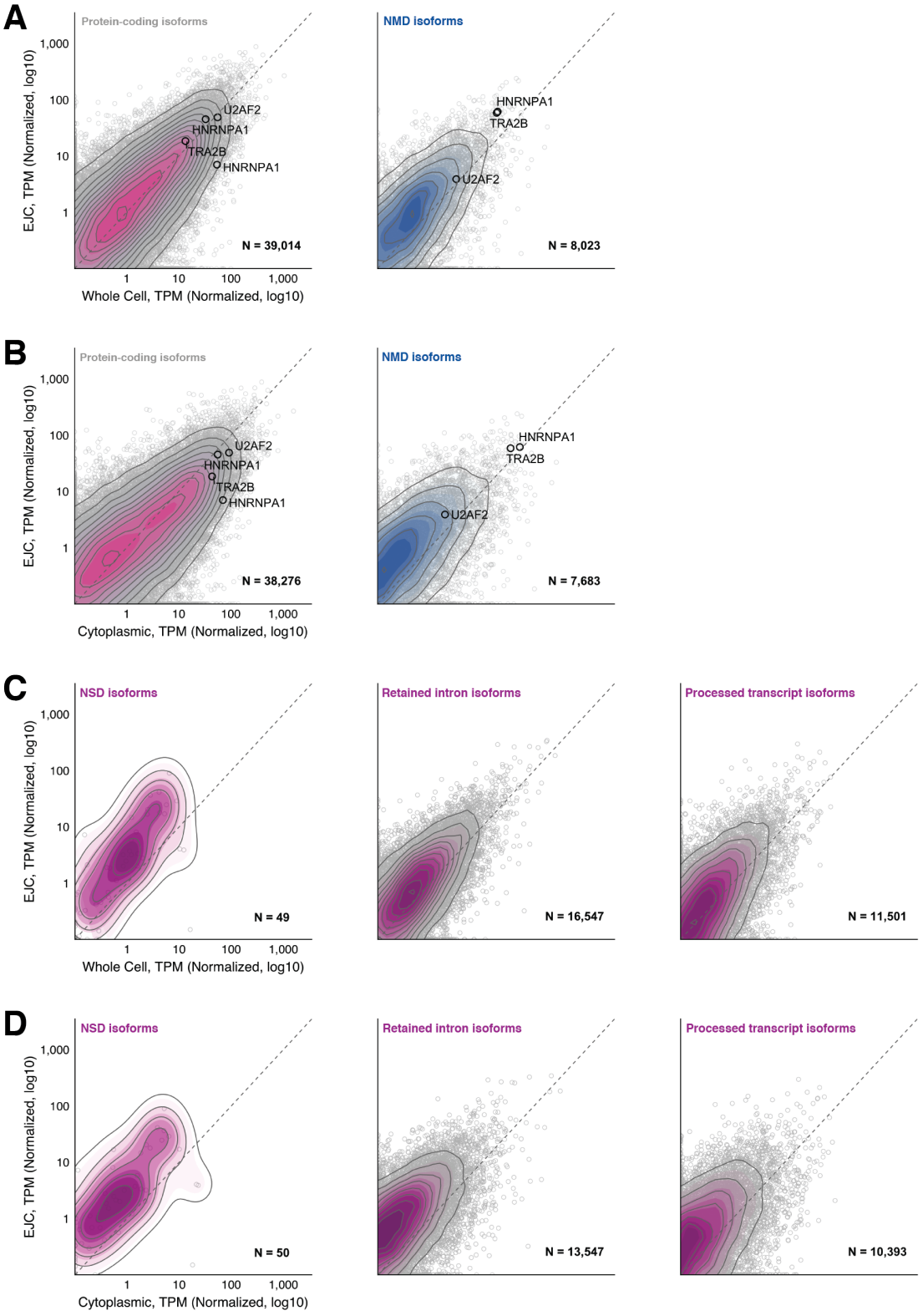

Figure 2.7: Scatterplots comparing TPMs between EJC RIPiT-Seq and whole cell or cytoplasmic RNA-Seq libraries for different isoform types: Protein-coding (A/B, left), NMD (A/B, right), non-stop decay (C/D, left), retained intron (C/D, middle), and processed transcript (C/D, right). In (A and B), transcripts from Figure 2.3 and 2.4 are noted. N: Number of detected transcripts out (of all annotated transcripts of that type). Dashed black line is the x=y line. Figure from Kovalak et al. 2020.

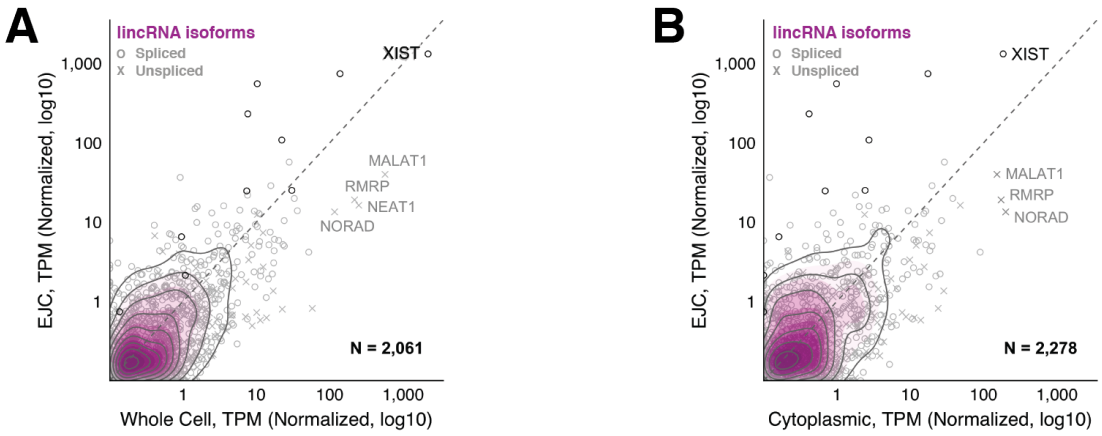

Figure 2.8: Scatterplots comparing TPMs between EJC RIPiT-Seq and whole cell (A) or cytoplasmic (B) RNA-Seq libraries for lincRNAs. O and X indicate spliced and unspliced lincRNA transcripts, respectively; XIST isoforms are indicated as open black circles. N: Number of detected transcripts out (of all annotated transcripts of that type). Dashed black line is the x=y line. Figure from Kovalak et al. 2020.

create a comprehensive reference file containing 575,976 known introns. CHESS is derived from 9,795 RNA-Seq samples from diverse cell types in the GTEx collection, so represents the most complete compendium of human transcripts reported to date (Pertea et al. 2018). Yet while CHESS found 118,183 new exon junctions not previously annotated in RefSeq, Ensembl or GENCODE, 82,918 other junctions present in RefSeq, Ensembl and/or GENCODE were not returned by the CHESS pipeline (Figure 1.15). This lack of concordance with respect to annotated junctions shows that even the most comprehensive RNA-Seq data analyses are unlikely to capture all bona fide splicing events.

To identify annotated and unannotated exon junctions in our EJC, whole cell and cytoplasmic libraries, we considered only those reads that cross an exon junction. The position of an exon junction in an individual read can be found by examining the "N operation" in the CIGAR string, which indicates the locations and lengths of gaps inserted during alignment to genomic DNA (Figure 2.9). We further required that any candidate junction: (1) occur within an annotated gene; (2) have reads with $\geq$ 15 nt aligning on both sides of the junction ($\geq$ 90% exact sequence match on each side); (3) be detectable in all replicates of a particular library type

(EJC, whole cell or cytoplasmic); and (4) have a mean read count ≥ 2 per library type (Figure 2.9). Using these criteria, we identified 151,072 junctions contained in the RefSeq/Ensembl/GENCODE/CHESS reference file (annotated junctions) and 5,917 previously unannotated junctions. MEME analysis of the latter revealed the 5′ and 3′ splice site consensus motifs for the major spliceosome, although at somewhat lesser strength (bits) than annotated junctions (Figure 2.10A). To limit our analysis to events most likely representing real splicing events (as opposed to mapping artifacts), we subsequently only considered the 5,412 previously unannotated junctions where the putative intron began and ended with dinucleotides expected for either the minor (AT-AC) or major (GT-AG) spliceosome. Of these, only three had AT-AC termini, indicating that the vast majority (>99.9%) of the unannotated events we detected are due to intron excision by the major spliceosome.

The majority (73%) of previously-annotated exon junctions meeting our detection criteria in protein coding genes (Figure 2.9) were present in all three library types (Figure 2.10B, left). There was less concordance, however, with respect to unannotated junctions, with the EJC libraries having many more unannotated junctions than either whole cell or cytoplasmic RNA-Seq (Figure 2.10B, right). Consistent with the expectation that EJC libraries should be enriched for exon junctions, both annotated and unannotated junctions were supported by more reads per million mapped (RPM) in the EJC libraries (Figure 2.11A). Also as expected, annotated junctions were generally supported by more reads than unannotated junctions in all library types. The major class (49%) of the new junctions were new alternative 5′ or 3′ splice sites (i.e., that combined a known 3′ or 5′ splice site with a previously unannotated 5′ or 3′ splice site, respectively) (Figure 2.11B). Other categories were previously unannotated exon skipping events (34%), new cassette exons (14%) and new introns (4%).

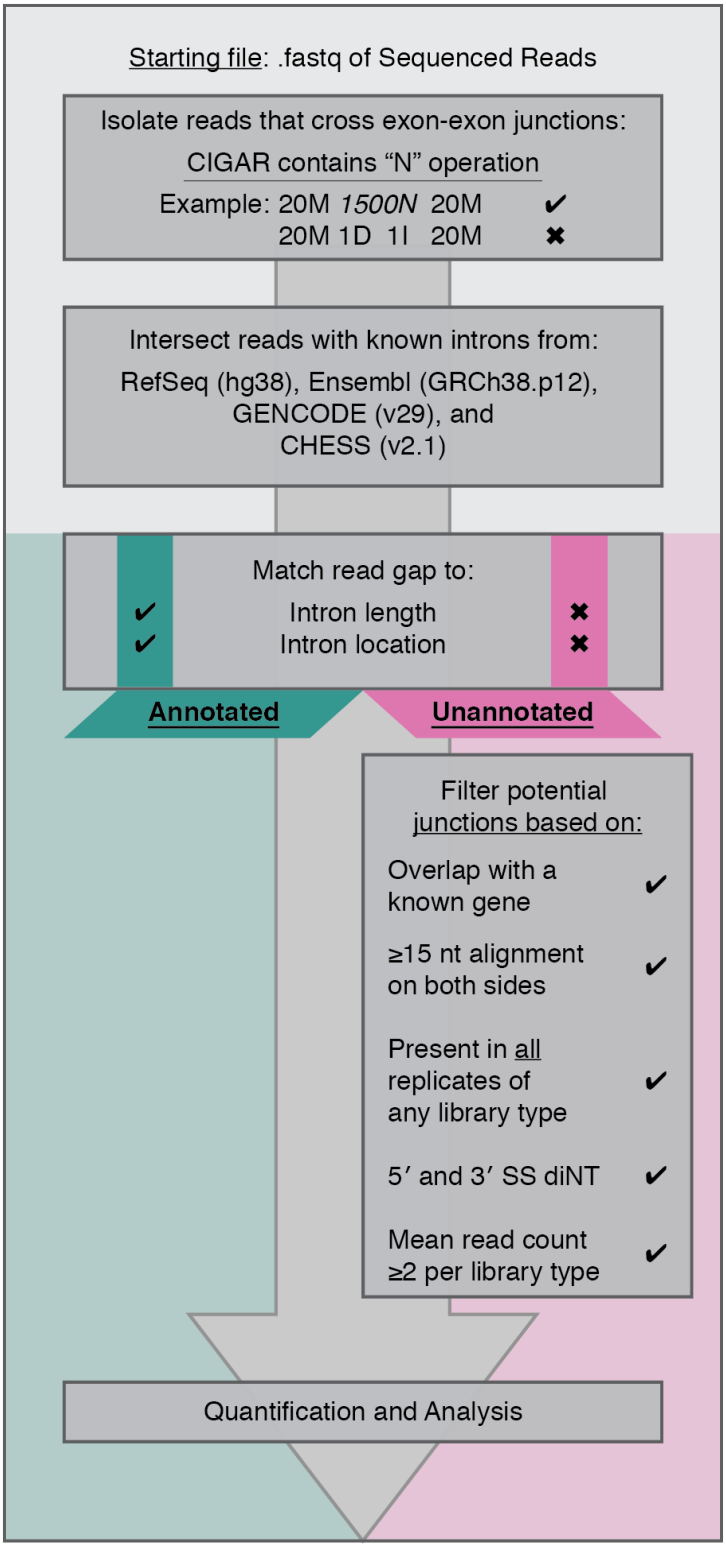

Figure 2.9: Schematic of library processing steps used to identify and analyze reads at annotated and unannotated junctions. Full details of each step are explained within the section above. Figure from Kovalak et al. 2020.

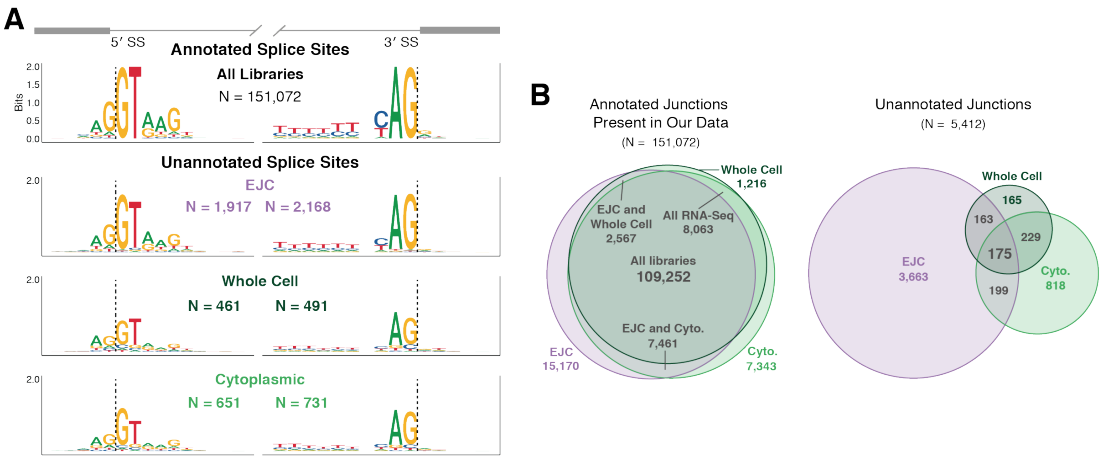

Figure 2.10: (A) Sequence motifs for 5' (left) and 3' (right) splice sites used in annotated junctions observed in at least one analyzed library type (top) and for previously unannotated splice sites in indicated library type (bottom). Sequence logos were generated in R using ggseqlogo; letter height signifies the relative abundance of that nucleotide at each position. N: Number of splice sites contributing to each logo. Note that the number of unannotated junctions (5,917) is greater than the total number of unannotated splice sites because many unannotated junctions combine an annotated and unannotated splice site (i.e., alternative $5'$ or $3'$ splice sites). (B) Venn diagram of annotated and previously unannotated junctions (numbers indicated) shared between library types. Venn diagrams made with eulerr. Figure from Kovalak et al. 2020.

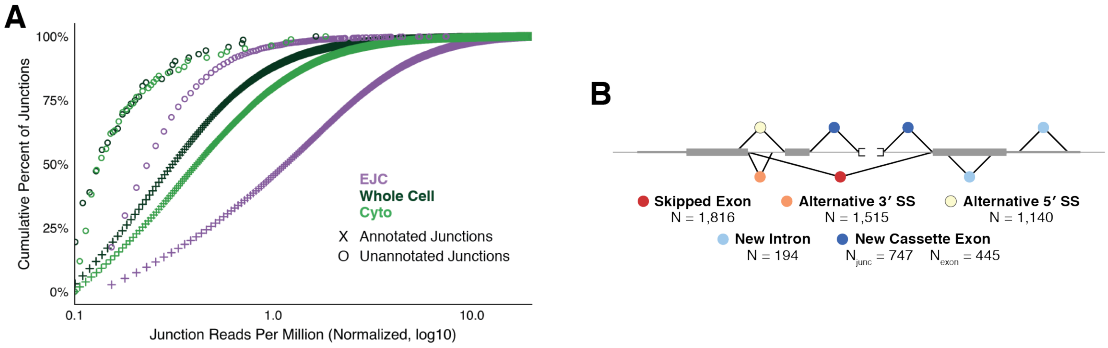

Figure 2.11: (A) Cumulative histogram of exon junction reads (RPM) at annotated (X) and previously unannotated (O) junctions in each library type (colors as in Figure 1A). (B) Schematic of unannotated splicing events separated by event type: Skipped exon (red); alternative $3'$ (orange) or $5'$ (yellow) splice site; new intron (light blue); new cassette exon (dark blue). N: number of observed events; for new cassette exons, both the number of observed unannotated junctions and number of new exons are shown. Figure from Kovalak et al. 2020.

2.3.4 Relationship of new splicing events to reading frame

Previous analyses of low abundance, unannotated splicing events in RNA-Seq data have revealed a strong tendency for such events to maintain reading frame (Dou et

al. 2006; Pickrell et al. 2010). To investigate whether this is due to some inherent ability of the splicing machinery to detect reading frame in the nucleus (Wachtel et al. 2004), or simply due to translation-dependent decay of out-of-frame events in the cytoplasm, we determined the distance from each previously unannotated splice site meeting our selection criteria to the nearest annotated splice site observed in any of our three library types. In all, 126 and 273 unannotated 5′ and 3′ splice sites, respectively, occurred within 15 nts of an annotated 5′ or 3′ splice site. Comparison of unannotated-to-annotated splice site distance aggregation plots between the three library types revealed both similarities and differences (Figure 2.12A). Around annotated 5′ splice sites, all three libraries displayed similar patterns, with the greatest unannotated usage being at intron position +5, consistent with the preference for a G and a T at positions +5 and +6, respectively, in the human 5′ splice site consensus sequence (Figure 2.10A) and the prevalence of GT dinucleotides at this position in this set of 126 5′ splice sites (dotted gray line in Figure 2.12A). More notable was the pattern near 3′ splice sites, where positions +3 and +4 in the downstream exon exhibited the highest unannotated usage. Strikingly, whereas the RNA-Seq libraries were strongly skewed toward position +3, both positions +3 and +4 in the EJC libraries were highly represented, with their usage closely reflecting the number of available AG's at these positions (dotted gray line in Figure 2.12A). Comparison of fractional abundance [unannotated read counts/(unannotated + annotated read counts)] at individual sites confirmed that whereas the EJC and RNA-Seq libraries exhibited similar utilization at position +3, utilization of position +4 was much more prominent in the EJC than either RNA-Seq library (Figure 2.12B). These observations strongly support a model in which out-of-frame splicing events are rapidly eliminated by NMD, resulting in their underrepresentation in both whole cell and cytoplasmic RNA-Seq libraries. Because utilization of AGs at positions +3 and +4 in the EJC libraries so closely paralleled their availability, we conclude that (at least with regard to 3′ splice

sites) the splicing machinery has no ability to read frame.

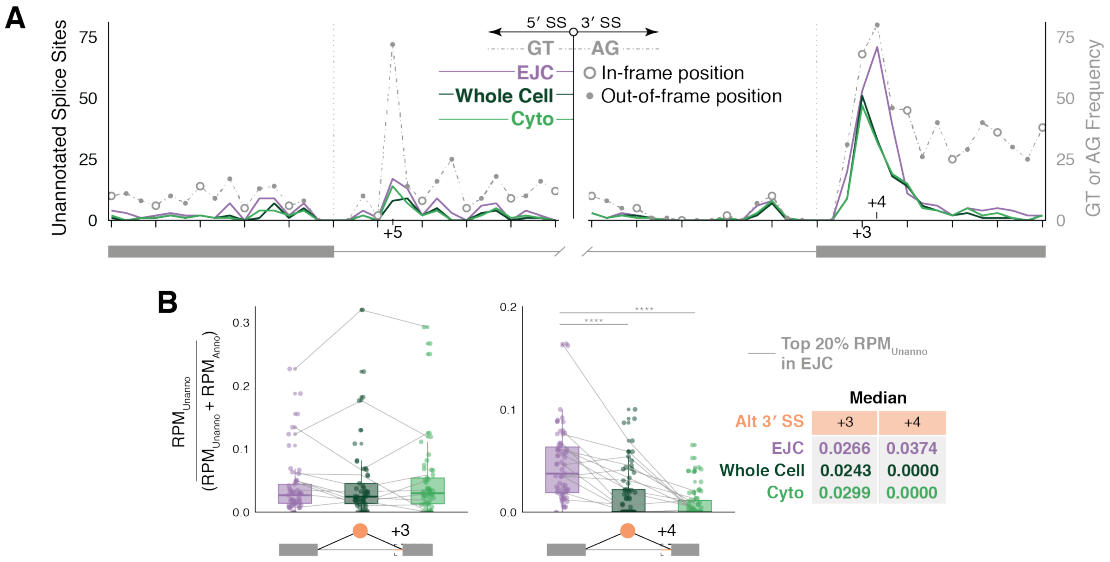

Figure 2.12: (A) Distribution of of unannotated splice sites relative to the closest annotated splice site observed in analyzed libraries (solid colored lines). Grey dotted line: Frequency of available GT or AG dinucleotides surrounding the annotated 5' (left) and 3' (right) splice sites with open circles indicating in-frame positions and solid grey dots indicating out-of-frame positions. (B) Distribution of the ratio of unannotated alternative 3' splice site use (RPMUnanno) over all events using the same 5' splice site (RPMUnanno+ RPMAnno) in each library type. (Left) Unannotated alternative 3' splice sites at the +3 position relative to closest annotated 3′ splice site; (middle) same but at the +4 position. Grey lines show how the top 20 percent (highest RPMUnanno) of unannotated junctions detected in EJC RIPiT-Seq libraries differ between library types. Results of one-way ANOVA and Tukey's post hoc tests comparing EJC RIPiT-Seq to RNA-Seq libraries are indicated; ****P<0.0001. (Right) Median (RPMUnanno/(RPMUnanno+ RPMAnno)) values per library at the +3 and +4 positions. Figure from Kovalak et al. 2020.

2.3.5 Evolutionary conservation versus splicing noise

Regardless of reading frame, most unannotated splicing events are likely due to "splicing error" (Fox-Walsh & Hertel 2009) or "splicing noise" (Pickrell et al. 2010). Splicing noise results from spurious utilization of cryptic splice sites that are not evolutionarily conserved. To assess both evolutionary conservation and splice site strength, we calculated mean basewise phyloP 30-way vertebrate conservation (Pollard et al. 2010) and MaxENT (a generally accepted measure of how well a particular splice site matches the consensus) (Yeo & Burge 2004) scores for both annotated and unannotated splice sites, using the same 5′ and 3′ splice site window sizes (9 and 23 nts, respectively) for both calculations (Figure 2.13). We also

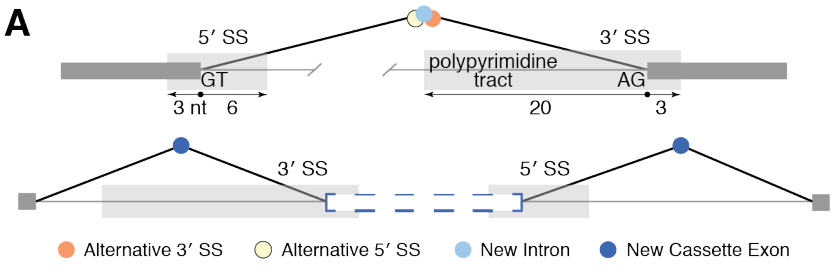

Figure 2.13: Regions used to calculate MaxEnt and mean conservation scores surrounding unannotated alternative 3' and 5' splice sites and new introns (top) or new cassette exons (bottom). Figure from Kovalak et al. 2020.

calculated conservation and MaxENT scores for sequences chosen at random from inside annotated genes and containing either GT or AG at the appropriate position within the 5′ or 3′ splice site window, respectively. Plotting MaxENT versus conservation revealed markedly different distributions between annotated splice sites and random GT- and AG-containing sequences (Figure 2.14A, Figure 2.16), with annotated sites being significantly skewed toward higher values for both measures. In contrast, whereas unannotated splice sites were similarly distributed as annotated splice sites with regard to MaxENT, the majority exhibited conservation scores more similar to random than annotated splice sites (Figure 2.14B). For the random sequences, 95% had 5′ and 3′ splice site conservation scores below 1.03 and 0.63, respectively. Using these values as cutoffs to filter out the majority of events likely due to splicing noise (although this may be unnecessarily conservative for 3′ splice sites due to the high degree of overlap between the annotated and random conservation scores) left us with 252 (12%) and 630 (26%) evolutionarily-conserved unannotated 5′ and 3′ splice sites, respectively. The majority of these occurred within annotated protein-coding exons, so their conservation is likely driven by amino acid conservation and not as a requirement for recognition by the splicing machinery (see Figure 2.17A for an example). Almost all of the new evolutionarily conserved introns (i.e., both the 5′ and 3′ splice sites were previously unannotated, but exhibited high conservation) also fell into this category. For the new introns, calculation of percent intron retention (PIR) in the EJC libraries revealed highly

inefficient splicing (mean PIR = 0.93), and individual examination of those exhibiting the highest number of exon junction reads in the EJC libraries led to no findings of particular note. Thus the new introns likely constitute splicing noise due to low level spliceosome assembly on sites within exons that by happenstance resemble splice site consensus sequences. In contrast, examination of unannotated 3′ splice sites occurring within introns uncovered a conserved alternative splicing event in the HECTD4 (HECT domain E3 ubiquitin protein ligase 4) gene that adds 9 amino acids into the middle of the protein (Figure 2.17B); this spliced isoform is currently annotated in mouse RefSeq and GENCODE, but not in humans. Other alternative 3′ splice sites in the CNOT1 and EEA1 genes generate AS-NMD isoforms (Figure 2.15), the latter due to creation of a new poison cassette exon.

2.3.6 New evolutionarily-conserved poison cassette exons

Having found examples of new AS-NMD isoforms generated by unannotated 3′ splice sites, we were interested to investigate which of the new cassette exons identified here might also function in this capacity. Of the 445 new cassette exons (Figure 2.11A), 412 (93%) occurred in protein-coding genes; the remainder occurred in pseudogenes and ncRNAs. Based on the data in Figure 2.3, poison exons should exhibit higher abundance in EJC than in RNA-Seq libraries. Consistent with this, 315/412 (76%) were solely detectable in the EJC libraries, with the remainder averaging 12- and 13-fold higher abundance in the EJC libraries than in whole cell or cytoplasmic RNA-Seq, respectively (Figure 2.18). Of the 377 new cassette exons detectable in EJC libraries, 70% were frameshifting (i.e., not a multiple of 3 nts long). Individual inspection of the 25 most abundant non-frameshifting exons revealed that 80% contained an in-frame stop codon. Therefore, as expected, the

vast majority of new exons likely function as poison cassette exons.

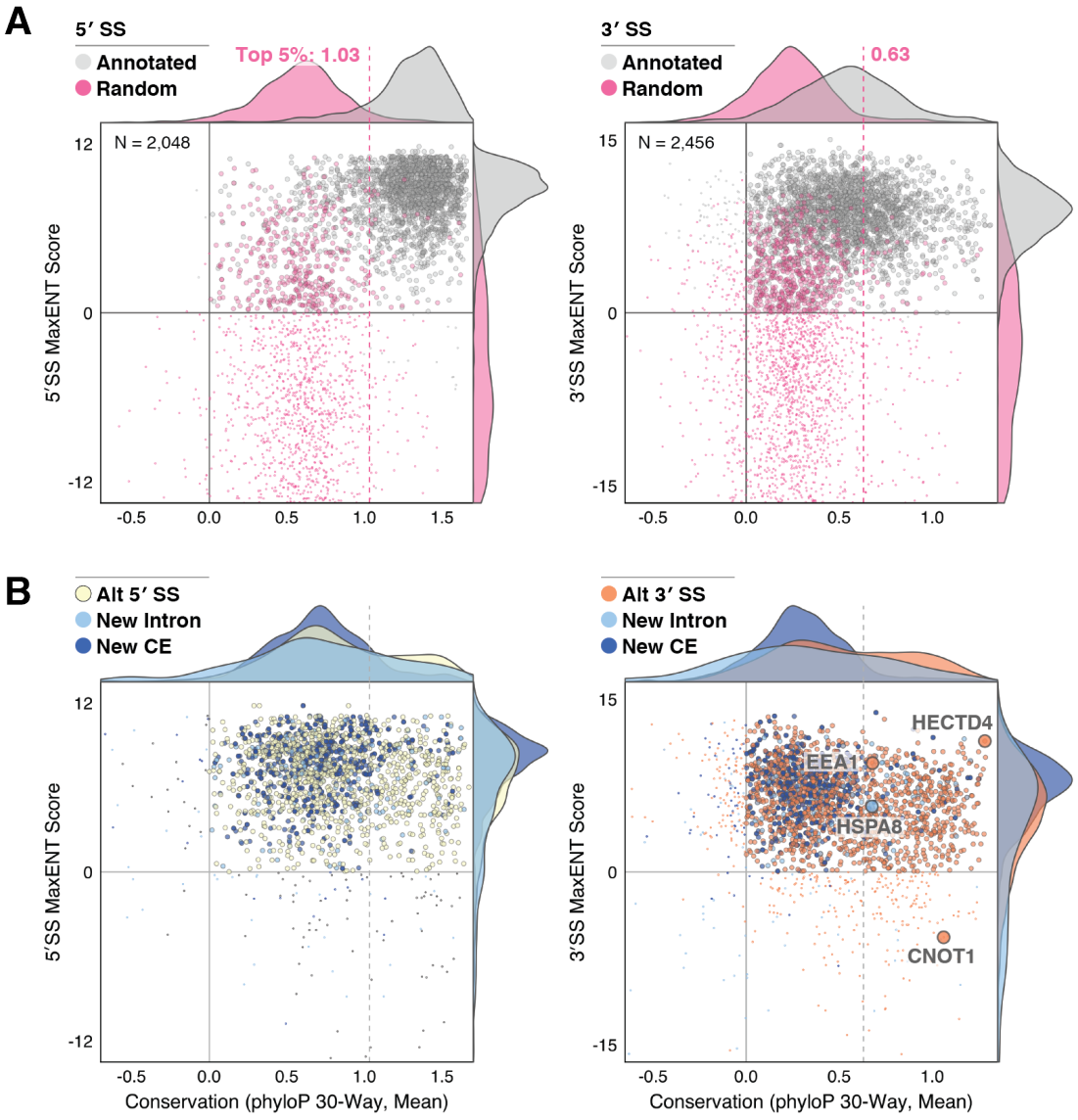

Figure 2.14: Scatterplots comparing MaxEnt scores to mean conservation scores (phyloP, 30-way) at 5' (left) or 3' (right) splice sites for (A) annotated and random or (B) observed unannotated events. Smaller points are used to represent splice sites with either score lower than 0 as these may result from splicing noise. Annotated splice sites were downsampled by randomly selection (5', N = 2,048; 3', N = 2,456; same as unannotated splice site numbers in C) from the 151,072 observed in our libraries. Figure 2.11A shows the same plot for all observed annotated splice sites. (A) also contains 2,048 random GT-containing (left) and 2,456 random AG-containing (right) sites; identical plots for four additional sets of randomized locations are shown in Figure 2.16A. The top 5 percent mean conservation scores of random sites is indicated and marked by a dashed line. Genes for which genome-browser tracks are shown in Figure 2.15 and Figure 2.17 are indicated. Figure from Kovalak et al. 2020.

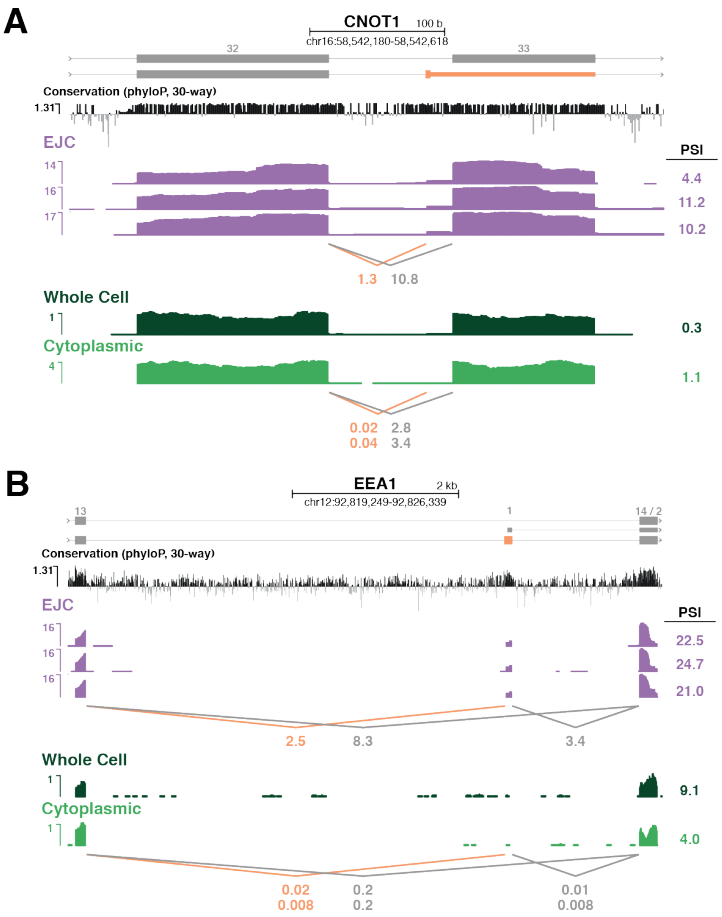

Figure 2.15: Genome browser tracks of library coverage across CNOT1 (A) and EEA1 (B). Annotated transcripts are shown in grey and unannotated alternative 3′ splice site use in orange. Conservation tracks and annotations are as in Figure 2.3. Figure from Kovalak et al. 2020.

To assess whether any of the new cassette exons constitute conserved regulatory elements, we calculated mean phyloP 30-way conservation scores across the entire exon. Combining these exon conservation scores (white to dark blue in Figure 2.19) with the previously calculated 5′ and 3′ splice site conservation scores (Figure 2.14A) revealed a set of 20 previously unannotated cassette exons exhibiting both high internal (phyloP score ≥ 1) and high splice site (≥ 1 for both splice sites) conservation (Figure 2.19 right). Among these, the most highly represented in our datasets was a new 94 nt exon within intron 8 of the 22-intron protein tyrosine phosphatase, receptor type A (PTPRA) gene (Figure 2.20A). Reminiscent of the conserved poison exons in TRA2B and U2AF2 (Figure 2.3A and B), inclusion of

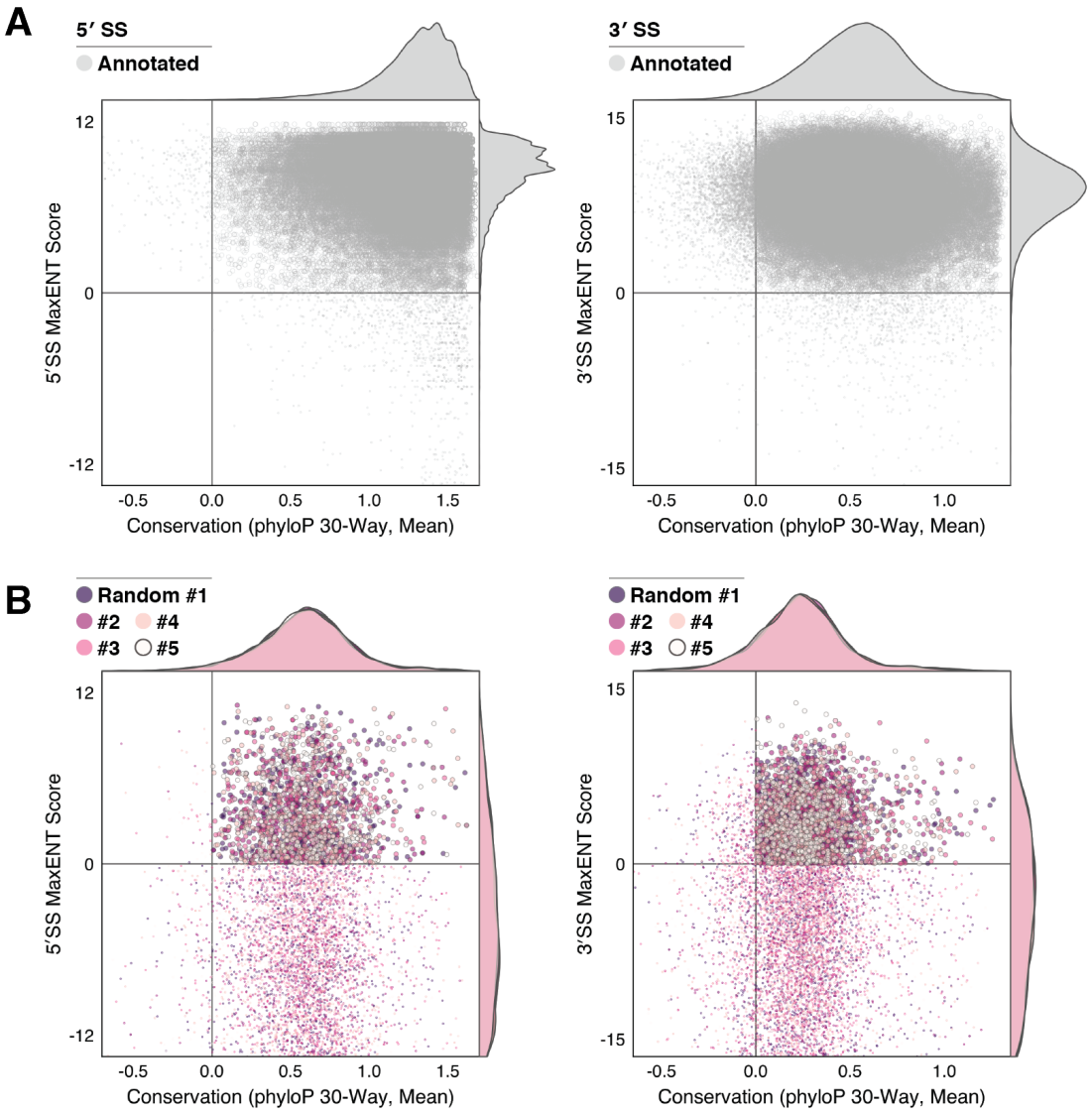

Figure 2.16: (A) Scatterplots comparing the MaxEnt score to mean conservation score (phyloP, 30-way) at 5' (left) or 3' (right) splice sites for all annotated junctions (N = 151,072). (B) Scatterplots comparing the MaxEnt score to conservation (phyloP, 30-way) at 5' (left) or 3' (right) splice sites for multiple sets (N = 5) of randomly selected sequences (5', N = 2,048; 3', N = 2,456). Figure from Kovalak et al. 2020.

Protein Tyrosine Phosphatase Receptor Type A (PTPRA) exon 8a was readily observable in the EJC libraries, but nearly undetectable in the RNA-Seq libraries (Figure 2.20A). Other high abundance examples were a 103 nt exon in intron 3 of the 29-intron DNA Polymerase Theta (POLQ) gene (Figure 2.21) and a 69 nt

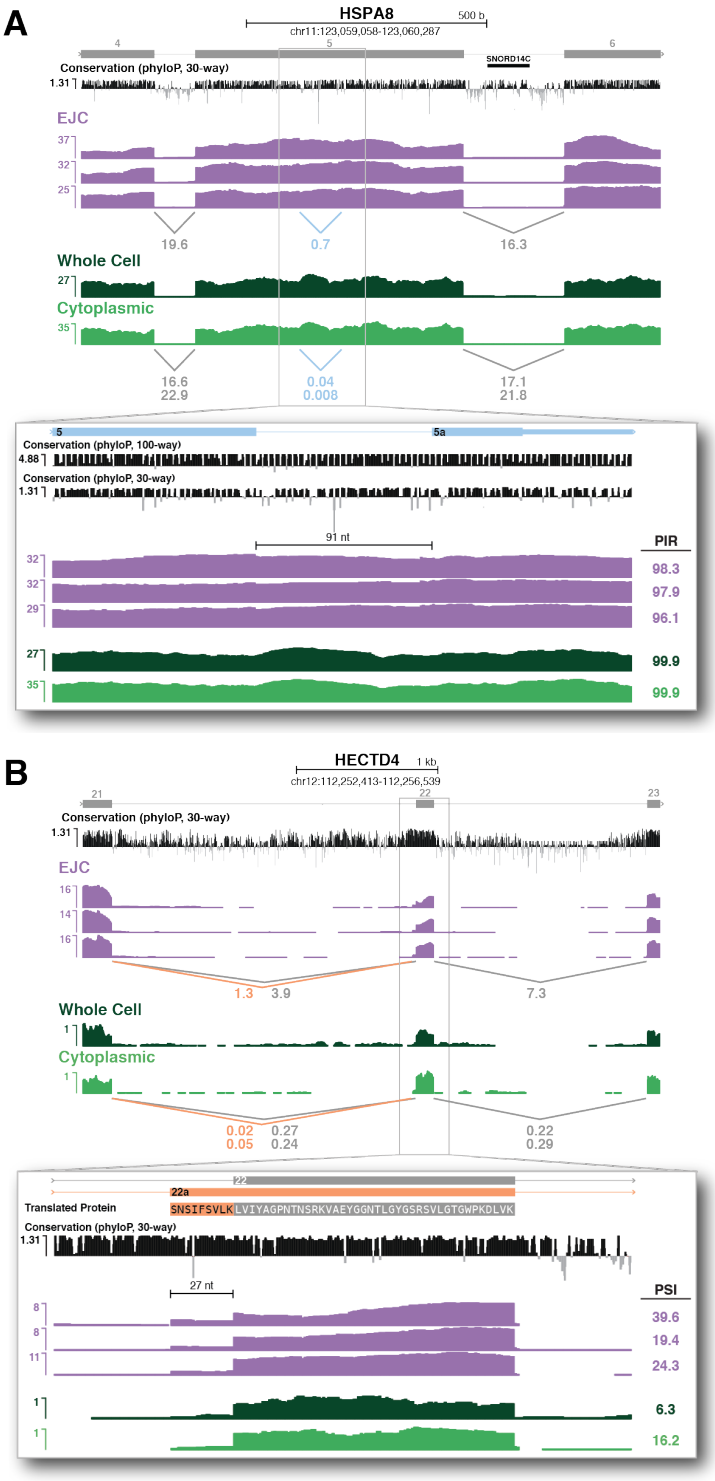

Figure 2.17: (A) Genome browser tracks of library coverage across HSPA8 (C) and HECTD4 (D). Annotated transcripts are shown in grey, unannotated alternative 3' splicing events in orange, and unannotated introns in light blue. Conservation tracks represent phyloP basewise scores derived from Multiz alignment of 30 vertebrate species, as well as 100 vertebrate species in (B). Numbers below tracks indicate mean reads per million (RPM) spanning each exon junction. Numbers to right in C and D are percent intron retention (PIR) and percent spliced in (PSI) values, respectively; PSI and PIR values for RNA-Seq libraries are replicate means. The translated protein sequences of both the annotated and unannotated transcripts are provided in (D). Figure from Kovalak et al. 2020.

exon in intron 37 of the 39-intron pleckstrin homology domain interacting protein (PHIP) gene (Figure 2.20B). Although PHIP exon 37a does not frameshift, it does contain three highly-conserved in-frame stop codons (Figure 2.20B, bottom). Thus all of the new evolutionarily-conserved cassette exons identified here likely function as poison exons to regulate protein expression from their host gene.

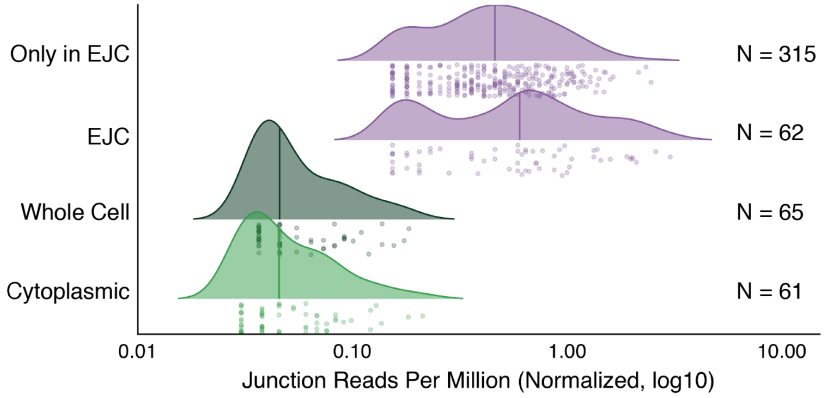

Figure 2.18: Density plot comparing junction-spanning read coverage (RPM) for new cassette exons in EJC and RNA-Seq libraries. Line indicates median expression per library and dots represent individual cassette exons. N: number of observed cassette exons per library. Figure from Kovalak et al. 2020.

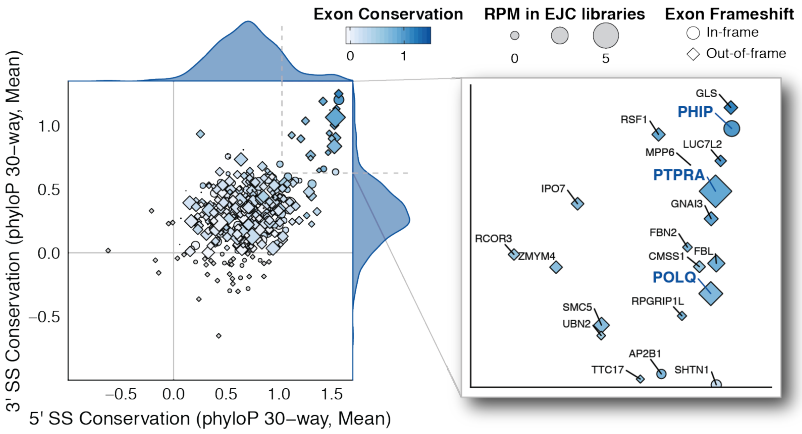

Figure 2.19: (A) (Left) Scatterplot comparing mean conservation (phyloP, 30-way) at 5' and 3' splice sites of new cassette exons. Exons with scores above 0 at both splice sites are colored (white to dark blue) to indicate mean exon conservation and sized by the number of junction-spanning reads supporting that exon in EJC RIPiT-Seq libraries. Diamonds indicate exons that create a frameshift in the resulting mRNA; circles indicate non-frameshifting exons. (Right) Zoomed view of exons with mean 5' and 3' splice site conservation scores above 1.03 and 0.63, respectively. Figure from Kovalak et al. 2020.

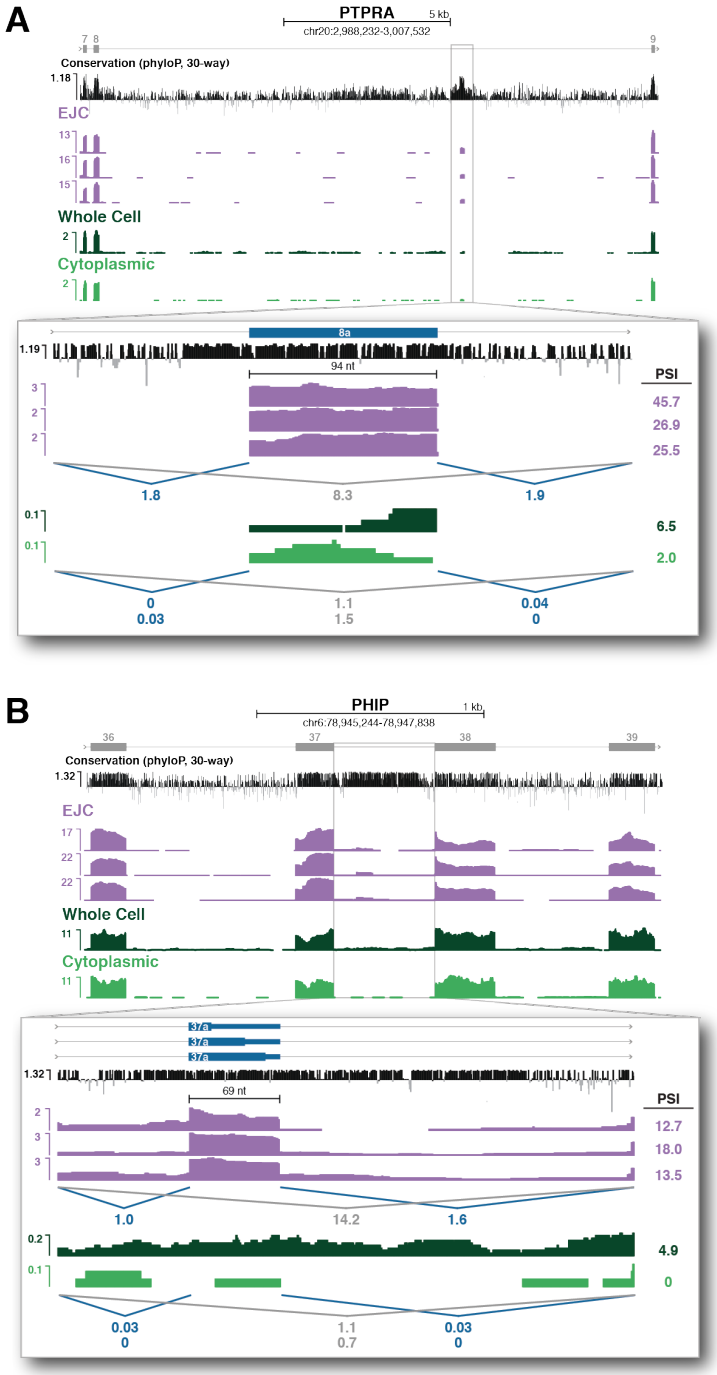

Figure 2.20: Genome browser tracks of library coverage across new poison cassette exons in PHIP (A) and PTPRA (B). New cassette exons are shown in blue and numbered according to their placement in the major isoform observed in all libraries. Conservation tracks and annotations are as in Figure 2.3. Figure from Kovalak et al. 2020.

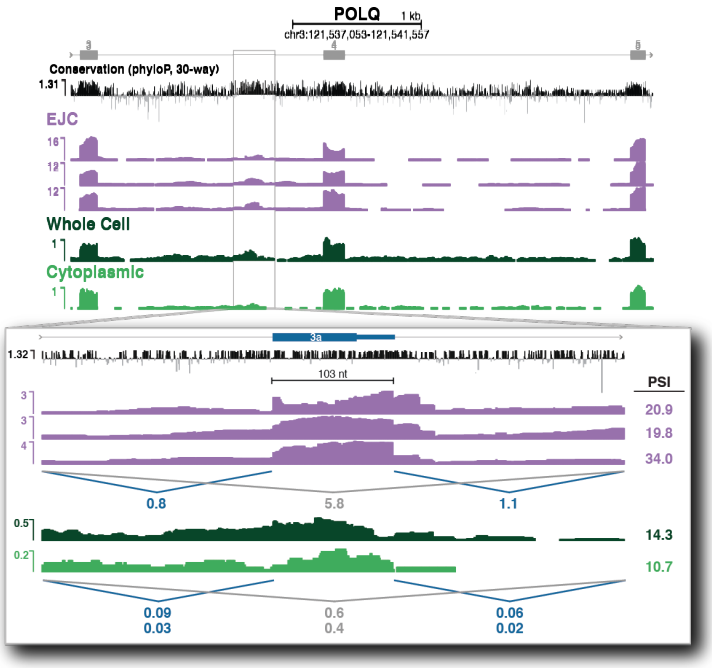

Figure 2.21: Genome browser tracks of library coverage across new poison cassette exons in POLQ. New cassette exons are shown in blue and numbered according to their placement in the major isoform observed in all libraries. Conservation tracks and annotations are as in Figure 2.3. Figure from Kovalak et al. 2020.

2.4 Discussion

Here we demonstrate that deep sequencing of transcripts in pre-translational RNPs provides a means to identify/quantify mRNA isoforms underrepresented in or absent from RNA-Seq libraries due to their rapid elimination by translation-dependent mRNA decay. We captured this pre-translational population by tandem immunoprecipitation (RIPiT) (Singh et al. 2014) of two core EJC proteins. EJCs are stably deposited upstream of exon junctions late in the pre-mRNA splicing process, and EJCs in 5′ UTRs and coding regions (~98% of all) are necessarily removed during the first or "pioneer" round of ribosome transit. Thus the EJC provides an excellent handle by which to enrich for fully-processed, but not-yet-translated mRNAs (Figure 2.1). Because they are specifically enriched for spliced transcripts, EJC RIPiT-Seq libraries also better capture low abundance splicing events than traditional RNA-Seq libraries. This enabled us to identify

thousands of new exon junctions not currently annotated in any of four major reference datasets based on RNA-Seq. Many of these new splicing events generate isoforms subject to NMD, with some being evolutionarily-conserved AS-NMD regulatory events. Thus EJC RIPiT-Seq constitutes a useful method to query the spliced transcriptome without the confounding effects of differential translation-dependent decay of individual mRNA isoforms.

2.4.1 Flux through AS-NMD pathways

Since its initial description (Morrison et al. 1997; Mitrovich & Anderson 2000), AS-NMD has increasingly emerged as a key post-transcriptional regulatory mechanism (Zheng et al. 2012; Hamid & Makeyev 2014; Yan et al. 2015). Due to their widely different decay rates, however, the flux through the alternative processing pathways resulting in protein-coding and NMD isoforms cannot be determined by traditional RNA-Seq methods. As shown in Figure 2.3 the vast majority of TRA2B (A) and U2AF2 (B) transcripts present in RNA-Seq libraries are the protein coding isoforms. Further, the lower poison exon PSI numbers in cytoplasmic than whole cell libraries are consistent with cytoplasmic decay of the NMD isoforms. The EJC RIPiT-Seq libraries, however, tell a very different story. For both TRA2B and U2AF2, the predominant pre-translational isoform is the poison-exon-included isoform, with poison exon PSIs averaging 94 and 73, respectively. Thus alternative splicing flux for both genes strongly favors poison exon inclusion. Similar results were observed for other RNA-binding protein genes known to maintain protein homeostasis by AS-NMD (Figures 2.3, 2.3, and 2.5). Indeed, enrichment of transcripts subject to translation-dependent decay (e.g., isoforms annotated as NMD and NSD) is a general feature of our EJC RIPiT libraries (Figures 2.7 and 2.8). We note, however, that to increase the abundance of pre-translational RNPs, we exposed our HEK293 cells to 2 mg/ml harringtonine for 60 minutes prior to cell

harvest and lysis (Metkar et al. 2018). At least in yeast growing under suboptimal conditions, inhibition of translation can induce rapid transcriptional upregulation of genes involved in ribosome biogenesis (Santos et al. 2019); the extent to which this is also true in mammalian cells growing under optimal conditions, and whether transcription and pre-mRNA processing of other gene classes are affected, has yet to be thoroughly explored. One recent study in HeLa cells, however, showed that, whereas a 15 minute exposure to 100 mg/ml cycloheximide had almost no effect on mRNA abundance in whole cell RNA-Seq, multiple mRNAs encoding ribosomal proteins decreased in abundance after a 24 hour cycloheximide treatment (i.e., the opposite of yeast) (Kearse et al. 2019). Because any transcriptional effects would confound the analysis, elimination of translation inhibitors would be advisable for any future EJC RIPiT-Seq study specifically aimed at quantifying flux through alternative RNA processing pathways in non-perturbed cells.

2.4.2 Identification of novel conserved splicing events

A major goal for this study was to assess the utility of EJC RIPiT-Seq libraries for identifying novel sites of exon ligation that are underrepresented in traditional RNA-Seq libraries. These could be splicing events resulting in either stable, low abundance isoforms or highly unstable transcripts such as NMD and NSD substrates. As illustrated in Figure 1.15, even the deepest analysis of RNA-Seq to date (CHESS) failed to capture all of the exon ligation events annotated in RefSeq, Ensembl or GENCODE. CHESS combined data from 9,795 GTEx RNA-Seq libraries covering dozens of tissues and comprising just under 900 billion reads. Yet EJC RIPiT-Seq libraries from a single cell type grown under a single condition encompassing only ~60 million reads enabled us to identify thousands of new exon junctions not currently annotated in RefSeq, Ensembl, GENCODE or CHESS (Figure 2.10B). Whereas the majority of these events occur at sites lack-

ing splice site conservation and so likely constitute splicing noise, hundreds exhibit high sequence conservation among mammals. Among this conserved set, the majority display features expected to generate an AS-NMD isoform (i.e., frameshift or in-frame stop codon).

2.4.3 New poison exons regulate genes linked to cancer

It has now been well established that changes to pre-mRNA splicing patterns can drive cancer initiation and progression (Sveen et al. 2016; Climente-González et al. 2017). Thus it is of particular note that three of the most conserved, high-abundance AS-NMD events discovered here are poison cassette exons in PTPRA, PHIP, and POLQ (Figure 2.20 and 2.21). All three genes have been linked to poor cancer prognosis when overexpressed (Tabiti et al. 1995; Ardini et al. 2000; Gu et al. 2017; De Semir et al. 2012; Wood & Doublié 2016; Goullet de Rugy et al. 2016). While protein overexpression in cancer often results from gene duplication or transcriptional dysregulation, decreased flux through a splicing pathway leading to poison exon inclusion would have the same effect. Previous studies examining the links between NMD and cancer have mainly focused on loss of tumor suppressor genes due to increased NMD (Lindeboom et al. 2016; Hu et al. 2017) or the advantageous effects of NMD in eliminating mRNA isoforms encoding neoepitopes that would otherwise be recognized by the immune system (Pastor et al. 2010). But our findings suggest that decreased poison exon inclusion should also be considered as a contributor to the mechanisms underlying cancer. An obvious means to alter splicing flux is a cis-acting mutation that disrupts splice site recognition and, thereby, poison exon inclusion. Although our examination of The Cancer Genome Atlas (Release 19) database (Weinstein et al. 2013) revealed no instances of splice site mutations associated with any of the new conserved poison cassette exons documented here, this possibility should certainly be considered in future

hunts for cancer-promoting mutations. Of note, current "exome" sequencing generally captures only DNA covering and surrounding annotated exons (Wang et al. 2018). Therefore, the unannotated cassette exons we identify here are likely absent from most DNA sequencing databases.

2.4.4 Conclusions

Sequencing of post-splicing, pre-translational mRNPs provides a powerful new approach to identify and quantify transient species that undergo rapid translation-dependent decay and are therefore under-represented in or completely absent from standard RNA-Seq libraries. The data here constitute just one snapshot of AS flux in HEK293 cells growing under optimal conditions. Future studies examining EJC RIPiT-Seq libraries from more diverse biological samples will undoubtedly lead to discovery of even more previously undocumented AS-NMD pathways. Examination of how flux through such pathways change in response to changing cellular conditions will increase our general understanding of how post-transcriptional mechanisms regulate protein abundance.

2.5 Availability of data and materials

The RIPiT datasets analyzed in this study were downloaded from NCBI GEO under accession number GSE115788 (specifically, samples GSM3189985, GSM3189986, and GSM3189987). RNA-Seq datasets were downloaded from the European Nucleotide Archive under accession number PRJEB4197 (specifically, runs ERR304485, ERR304486, ERR304487, and ERR304488).

2.6 Acknowledgments

We thank Alicia Bicknell, Athma Pai, Harleen Saini and Guramrit Singh for critical reading of the manuscript. We thank Weijun Chen for technical advice.

2.7 Methods

2.7.1 Deep sequencing libraries

All libraries were downloaded from the NCBI GEO GSE115788 (specifically, samples GSM3189985, GSM3189986, and GSM3189987) and the European Nucleotide Archive PRJEB4197 (specifically, runs ERR304485, ERR304486, ERR304487, and ERR304488).

EJC libraries were generated from 200-550 nt fragments by 3′ adaptor ligation and reverse transcription. Paired-end sequencing (150 nt reads) on the Illumina NextSeq platform resulted in 18-24 million mate pairs per replicate (Metkar et al. 2018). RNA-Seq datasets were obtained by paired-end sequencing (51 nt reads) on the Illumina HiSeq platform of Ribo-Zero-treated libraries generated with a modified Illumina TruSeq protocol (Sultan et al. 2014) containing 100 to 200 nt sized inserts. Each RNA-Seq replicate contained an average of 50 to 60 million mate pairs per library.

2.7.2 Library processing and alignment

Read counts for unprocessed libraries and for the individual processing steps detailed below are provided in Table 2.1. Prior to alignment, adaptor sequences and long stretches ($\geq$ 20 nt) of adenosines were trimmed from the 3′ end of se-

quencing reads. All libraries were filtered using STAR v2.5.3a (Dobin et al. 2013) for reads that aligned to repeat regions, as defined by RepeatMasker. Remaining reads were aligned with STAR on two-pass mode to the human genome, release 93 (Cunningham et al. 2019). This alignment allowed a maximum of 3 mismatches per pair and highly penalized deletions and insertions. Mapped reads were then filtered for low mapping quality (MAPQ < 5) and/or duplicated reads, identified with the MarkDuplicates tool (Picard v2.17.8).

2.7.3 RNA isoform quantification

RNA isoform abundances were determined using Kallisto (v0.44.0) (Bray et al. 2016), using only reads that passed the filtering and alignment steps described above. Transcript biotypes (i.e., "protein-coding", "nonsense-mediated decay", etc.) and intron counts used to categorize transcripts throughout Figure 2 are based on the transcriptome annotation from Ensembl (GRCh38.p12) (Cunningham et al. 2019).

2.7.4 Junction identification pipeline

The custom bioinformatics pipeline designed for our annotated and unannotated junction analysis is shown in detail in Figure 2.9. Transcriptome annotation files from RefSeq (hg38) (O'Leary et al. 2016), Ensembl (GRCh38.p12) (Cunningham et al. 2019), GENCODE (v29) (Frankish et al. 2019), and CHESS (v2.1) (Pertea et al. 2018) were combined to create a comprehensive reference file of all annotated introns. Any junction that appears in our libraries but is not annotated in one of the aforementioned transcriptomes is referred to as "unannotated."

To identify unannotated exon junctions, all reads with CIGAR strings containing

an "N" operation were isolated and then compared to the annotated intron reference file using Bedtools intersect (Quinlan & Hall 2010). Reads that did not match the length or location of a known intron were considered the result of potential unannotated splicing events. These junctions were further filtered based on the following criteria: (i) overlap with a known gene, (ii) reads must have ≥ 15 nt aligned on both sides of the potential junction, (iii) present in all replicates of any library type, (iv) major spliceosome dinucleotide consensus sequences at the $5'$ and $3'$ splice sites, and (v) mean read count ≥ 2 per library type.

2.7.5 Nearest annotated splice site analysis

For analysis of new splicing events near annotated exons (Figure 2.12), each unannotated $5'$ splice site was paired with its nearest annotated $5'$ splice site based on the $3'$ splice site used in both splicing events. Similarly, each unannotated $3'$ splice site was paired with its nearest annotated $3'$ splice site based on the $5'$ splice site used in both splicing events. The number of available GT and AG dinucleotides at nucleotide positions -30 to +30 surrounding each annotated splice site in this unannotated/annotated paired dataset.

2.7.6 Splice site strength and conservation

Splice site strength and mean conservation scores for annotated and unannotated splice sites were calculated using MaxEntScan (Yeo & Burge 2004) and phyloP 30-way basewise conservation scores (Pollard et al. 2010) (Figure 2.13). Random sequences of the appropriate length (9 nts for $5'$ splice sites and 23 nts for $3'$ splice sites) and internal to annotated genes were obtained from the hg38 annotation file (Cunningham et al. 2019) using the Bedtools random function (Quinlan & Hall 2010). Only those random sequences containing a GT at positions 4 and 5 or

an AG at positions 19 and 20 were used to calculate MaxENT and conservation scores for comparison to 5′ and 3′ splice sites, respectively.

2.7.7 Plotting and data visualization

Data visualization was performed in R using ggplot2, ggrepel, UpSetR, ggseqlogo, eulerr, and ggridges software packages. The UCSC Genome Browser was used to view sequencing library tracks and for transcript figures throughout the manuscript.

Chapter 3

Isolating the Activated Spliceosome

3.1 Preface

The contents of this Chapter have not been published previously and contain work completed during the first half of my doctoral research.

3.2 Introduction

Commitment of a pre-mRNA to a specific splicing pattern represents a critical step for regulation of alternatively spliced transcripts. Data suggests that commitment occurs in two phases for both constitutively and alternatively spliced substrates (Lim & Hertel 2004; Kotlajich et al. 2009). The first phase, commitment to splicing, is thought to occur upon E complex formation. *In vitro* chase experiments, in which stalled spliceosomes were first incubated with radiolabeled pre-mRNAs, and then chased with an excess of unlabeled pre-mRNAs, demonstrated that pre-

mRNAs bound by E complexes eventually spliced while naked transcripts did not (Legrain et al. 1988; Seraphin & Rosbash 1989). E complexes purified from mammalian cell extracts behaved similarly (Michaud & Reed 1991).

These early spliceosomes, however, have not yet committed to specific splice sites. *In vitro* kinetic trapping experiments, in which spliceosomes are stalled in the desired complex by altering buffer conditions, revealed that trapped E complexes could change splice sites when incubated with purified SR proteins (Lim & Hertel 2004; Kotlajich et al. 2009). Nevertheless, the addition of SR proteins did not affect the splicing pattern after U2 snRNP began assembly near the 3′ splice site. These results suggest that the formation of the first ATP-dependent complex commits the spliceosome to the chosen splice sites. It follows then that alternative splicing patterns must be decided early in the assembly pathway.

However, more recent single molecule experiments indicate that many spliceosome assembly steps are reversible (Hoskins et al. 2011). Colocalization single-molecule spectroscopy (CoSMoS) experiments use fluorescently tagged snRNPs and pre-mRNAs to follow spliceosome assembly in real time. These experiments revealed that spliceosome components could engage and dissociate from a pre-mRNA multiple times before splicing. In fact, further CoSMoS analysis demonstrated that early complex (i.e., prespliceosomes) formation occurs in a branched pathway of U1 or U2 snRNP recruitment (Figure 3.1; Shcherbakova et al. 2013). Though there does appear to be some level of commitment as pre-mRNAs associated with late stage spliceosome were more likely to complete the splicing cycle (Hoskins et al. 2011). Reversal of late stage spliceosomes suggests that regulation of splice site choice may occur much later in the assembly pathway than previously thought.

To observe alternative splicing choices on a transcriptome-wide scale, current analyses most frequently use RNA-Seq-based experimentation, but these datasets only

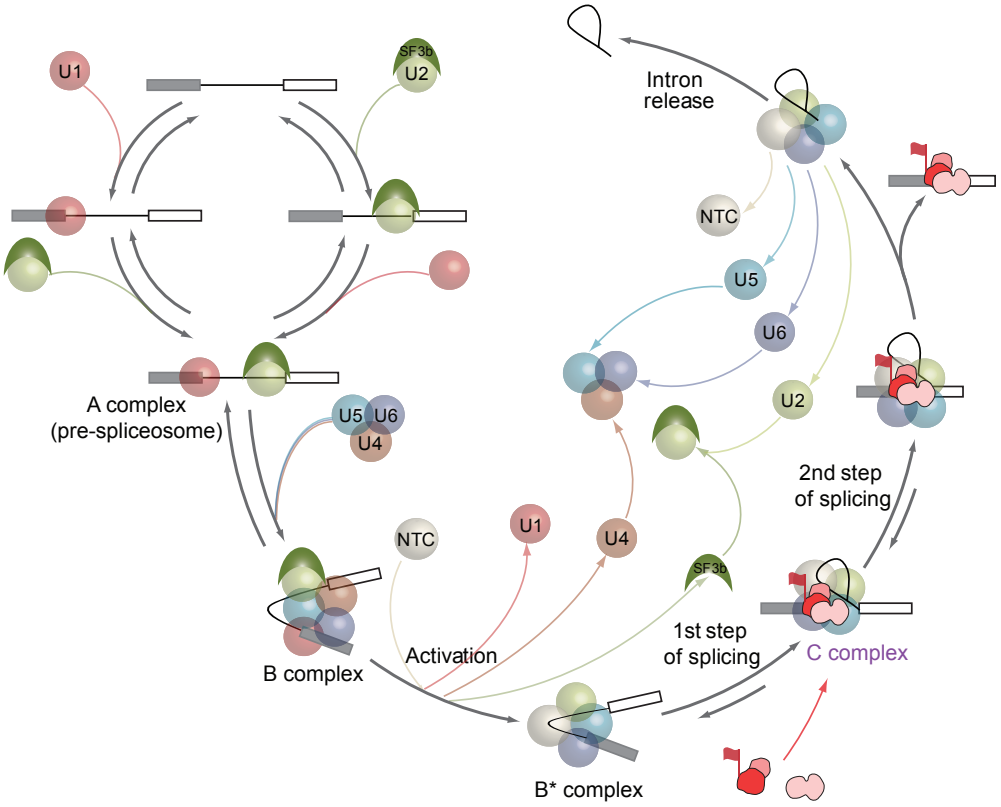

Figure 3.1: Schematic of the spliceosome assembly cycle showing the stepwise composition of snRNP(s) and RNA within each discrete spliceosomal complex. Updated version of Figure 1.3 reflecting assembly reversibility determined by CoSMoS experiments. Figure adapted from Shcherbakova et al. 2013. Copyright 2013 Elsevier.

contain evidence of final splicing choices. As such, evidence of reversed spliceosome assembly (and the resulting spliced intermediates) is lost. Recently, Chen *et al.* demonstrated that intron-lariat-containing complexes from the genetically tractable fission yeast are stable during tandem affinity purifications (Chen et al. 2014). By mapping the excised introns to their genomic locations, the authors were able to visualize spliceosome occupancy transcriptome-wide and identify more than a hundred previously unannotated yeast introns. Later dubbed "spliceosome profiling," this technique was modified to include a number of other sequencing methods (Figure 3.2; Chen et al. 2018). This allowed the authors to better globally map splice site and branch point locations (Figure 3.3), and led to the identification of hundreds of new introns.

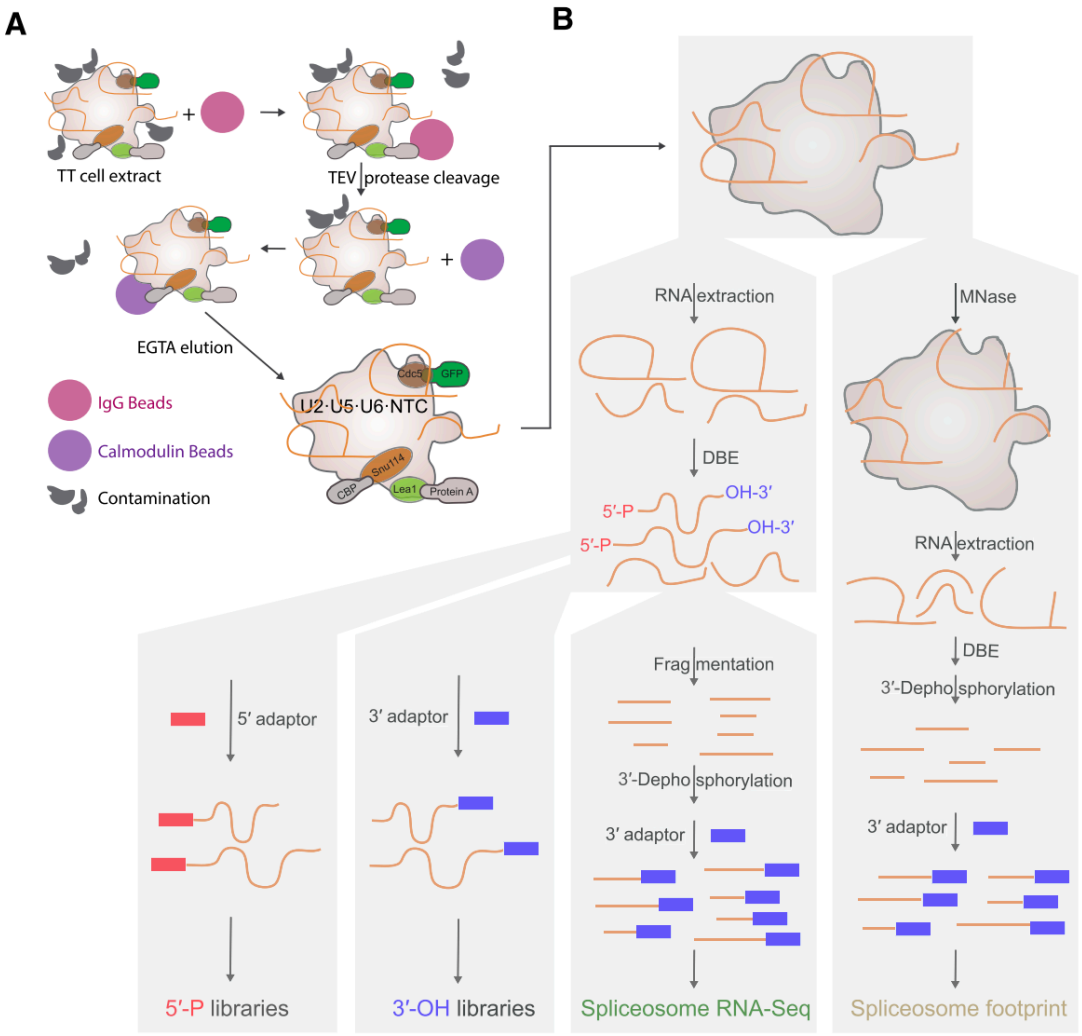

Figure 3.2: Schematic of 5'-P, 3'-OH, spliceosome RNA-seq, and spliceosome footprint libraries constructed to identify the footprint of intron lariat spliceosome complexes in S. pombe. 5'-P and 3'-OH libraries mark intronic boundaries; spliceosome RNA-seq libraries cover the entire intron whereas spliceosome footprinting protects only RNA within the isolated complex. Figure from Chen et al. 2018. Copyright 2018 Elsevier.

In this chapter, I present the initial findings from activated spliceosomes isolated from the nuclear fraction of HEK293 cells using RIPiT-Seq. These findings will be compared to both a control RIPiT-Seq targeting the *e*xon *j*unction *c*omplex (EJC; Section 1.1.3) and cytoplasmic RNA-Seq. Although this project was not completed, the data contained within this chapter illustrate the potential for footprinting mammalian spliceosomes transcriptome-wide using RIPiT-Seq.

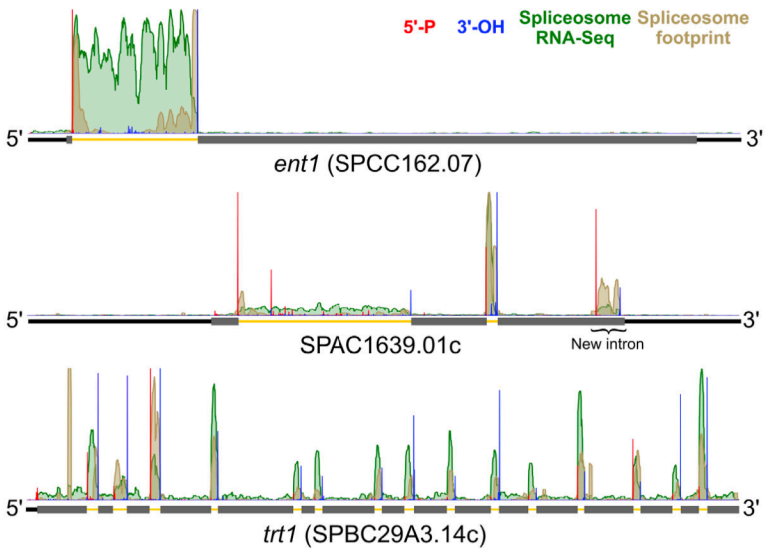

Figure 3.3: Genome browser tracks of library coverage across individual genes containing one (top), two (middle), or multiple (bottom) annotated introns. SPAC1638.01c also contains a previously unannotated intron. Figure from Chen et al. 2018. Copyright 2018 Elsevier.

3.3 Results

3.3.1 Isolating late-stage activated spliceosomes

In a previous study from the Moore lab that examined footprints of the EJC transcriptome-wide (Singh et al. 2012), components of the spliceosome were observed via mass spectrometry in the elution of single *i*mmuno*p*recipitations (IPs) targeting two different core EJC components (Figure 3.4, left). This result was somewhat expected as the EJC associates with the activated spliceosome complex during, and subsequent to, the first step of splicing (Section 1.3.1). Due to the dynamics of the spliceosome's composition, however, it was uncertain whether these interactions were sufficient to preserve assembly of the spliceosome during experimentation. Fortunately, the mass spec analysis suggested that the EJC RIPiT-Seq strategy could be adapted to isolate spliceosomes by targeting a spliceosomal protein in the second IP (Figure 3.4, right).

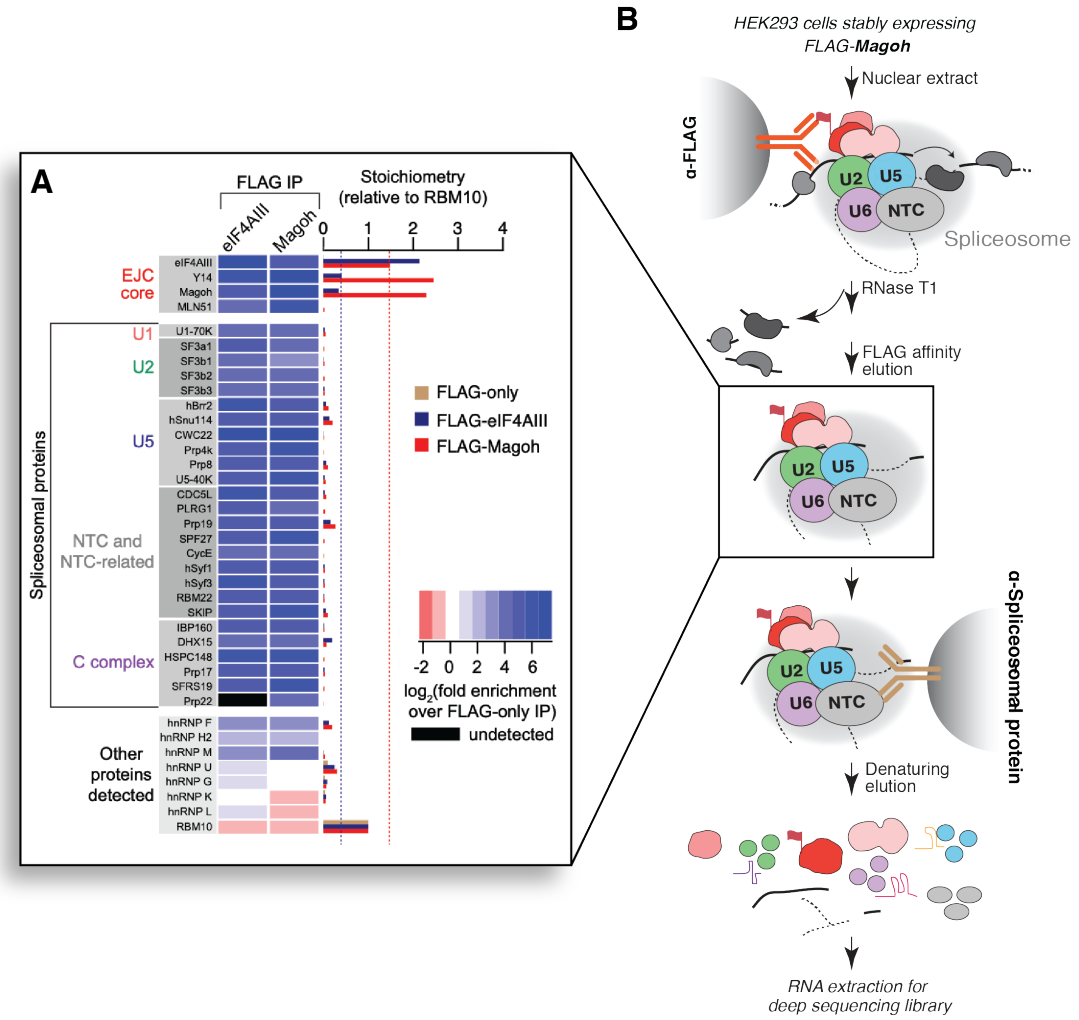

Figure 3.4: (A) Proteins (left) that co-purify with core EJC machinery after immunoprecipitation of FLAG-eIF4AII or FLAG-Magoh. Heat map (center) indicates fold enrichment of each protein compared to a control FLAG-only IP. Protein stoichiometries relative to RBM10 compared between each IP (right). Dashed line shows levels of copurified coe EJC proteins (blue, level of Y14 and Magoh in FLAG-eIF4AIII IP; red, level of eIF4AII in FLAG-Magoh IP.). (B) Schematic of RIPiT-Seq strategy to specifically purify activated spliceosome complexes. Spliceosome and core EJC proteins, colored; exonic RNA sequence, black line; intronic RNA sequence, dashed line. Figure adapted from Singh et al. 2012. Copyright 2012 Elsevier.

The modified RIPiT-Seq protocol (Figure 3.4, right) used for the experimentation described below can be found in detail in Section 3.5. In addition to the antibody change during the second IP, we also altered this protocol in two ways to further enrich for spliceosomes. First, we treated HEK293 cells with digitonin, a detergent that can permeabilize the plasma membrane but not the nuclear membrane. We then discarded the resulting cytoplasmic supernatant, including EJC-bound cytoplasmic mRNAs. Second, we treated the isolated nuclei with formaldehyde

to chemically crosslink RNAs and proteins. Crosslinking prevents protein dissociation and reassociation events during cell lysis and IP and had previously proven beneficial to mapping binding sites of other RNA-binding proteins (e.g., Staufen1) transcriptome-wide using RIPiT-Seq (Ricci et al. 2014; Singh et al. 2014). Whereas the EJC very stably binds to an mRNA and can maintain these interactions under native conditions throughout RIPiT-Seq, associations within the spliceosome and between the spliceosome and its pre-mRNA substrate are much more transient. Thus, we collected input material for spliceosome RIPiT-Seq experiments from nuclei treated both with or without formaldehyde.

Spliceosomes containing FLAG-tagged Magoh, a core EJC component, within this material were then immunoprecipitated using anti-FLAG beads. While on the beads, we fragmented associated RNA using a limited RNase T1 digestion. This deviates from the RNase 1 treatment in the RIPiT-Seq protocol described previously (Singh et al. 2014) as RNase T1 cleavage is restricted to guanine residues on single stranded RNA, thus leaving uridine-rich snRNAs mostly intact. Once eluted from the FLAG beads, we then mixed the immunoprecipitated material with beads conjugated to either anti-Prp19 or anti-IBP160. Both proteins are known components of late-stage activated spliceosomes (Figure 1.4). Prp19 is the namesake component of the nineteen complex (NTC), which joins the spliceosome prior to catalytic activation (reviewed in Wahl et al. 2009). IBP160 (KIAA0560) is a SF1 RNA helicase recruited to the spliceosomal C1 complex by components of the U2 snRNP. Material eluted from this second IP was then used for subsequent analysis.

3.3.2 Differences between EJC and spliceosome RIPiT-Seq

As the first IP in the spliceosome RIPiT-Seq strategy still targeted a core component of the EJC, it was uncertain whether changing the second IP would be sufficient to isolate a different macromolecular complex. To test whether this modified protocol successfully enriched for spliceosomes, we compared proteins in the final eluate to a single FLAG IP targeting FLAG-Magoh and an input control (Figure 3.5, right). The Western revealed that the second IP targeting the spliceosome enriched for components of the U2 and U5 snRNPs, the C complex, and the NTC in the eluted material, whereas most EJC components were partially lost. Moreover, denaturing PAGE gels showed distinct length differences between the RNA footprints of the EJC (Figure 3.6, A, lane 2) and the spliceosome (A, lane 3). The distribution of RNAs protected by the EJC contained both a shorter footprint attributable to a single EJC and a longer one from EJC multimers, corroborating previous data (Singh et al. 2012). However, the smear of RNAs in the spliceosome RIPiT-Seq sample demonstrated that the isolated complexes were different. Furthermore, this distribution remained the same across different spliceosome RIPiT-Seq samples (Figure 3.6, B). Collectively, this data confirmed that this modified version of RIPiT-Seq did in fact selectively target late-stage activated spliceosomes.

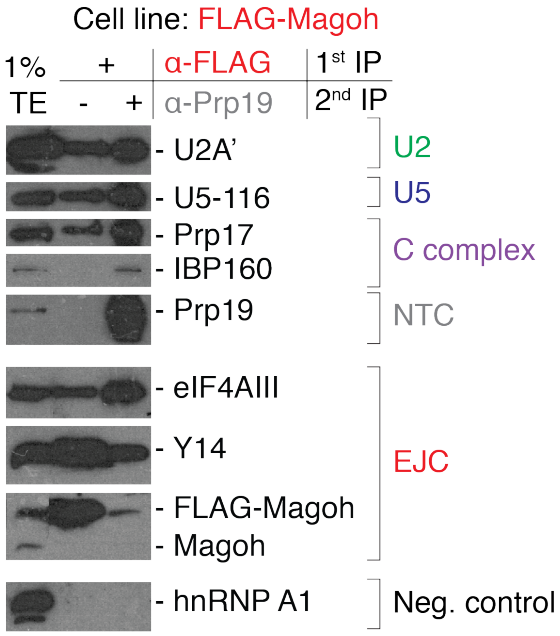

Figure 3.5: Purification of activated spliceosomal complexes using a RIPiT-Seq strategy targeting FLAG-Magoh followed by Prp19, a component of the Nineteen Complex (Section 1.1). Spliceosomal proteins are enriched in the tandem IP (right lane) compared to single FLAG-Magoh IP (center lane). Data courtesy of Guramrit Singh, PhD.

3.3.3 Composition of spliceosome RIPiT-Seq libraries

RNAs corresponding to spliceosome footprints were isolated from four different RIPiT-Seq experiments (Table 3.1), size selected between 30 and 80 nts, and sequenced after preparation using a library protocol developed by our lab (Heyer et al. 2015). The resulting sequenced libraries ranged in depth between 8 and 36 million reads. In the following analysis, we compared these libraries at times to previously published EJC RIPiT-Seq and cytoplasmic RNA-Seq libraries to identify footprints specific to the spliceosome (Singh et al. 2012; Ricci et al. 2014).

Prior to alignment to the genome, we first filtered out rRNA, snRNA and repeat mapping reads. Although ribosomal RNA made up approximately half of our uncrosslinked libraries, formaldehyde crosslinking and more stringent wash conditions reduced the amount of contaminating rRNA (Figure 3.7, left). As expected, snRNAs were much more abundant in the spliceosome RIPiT-Seq libraries than

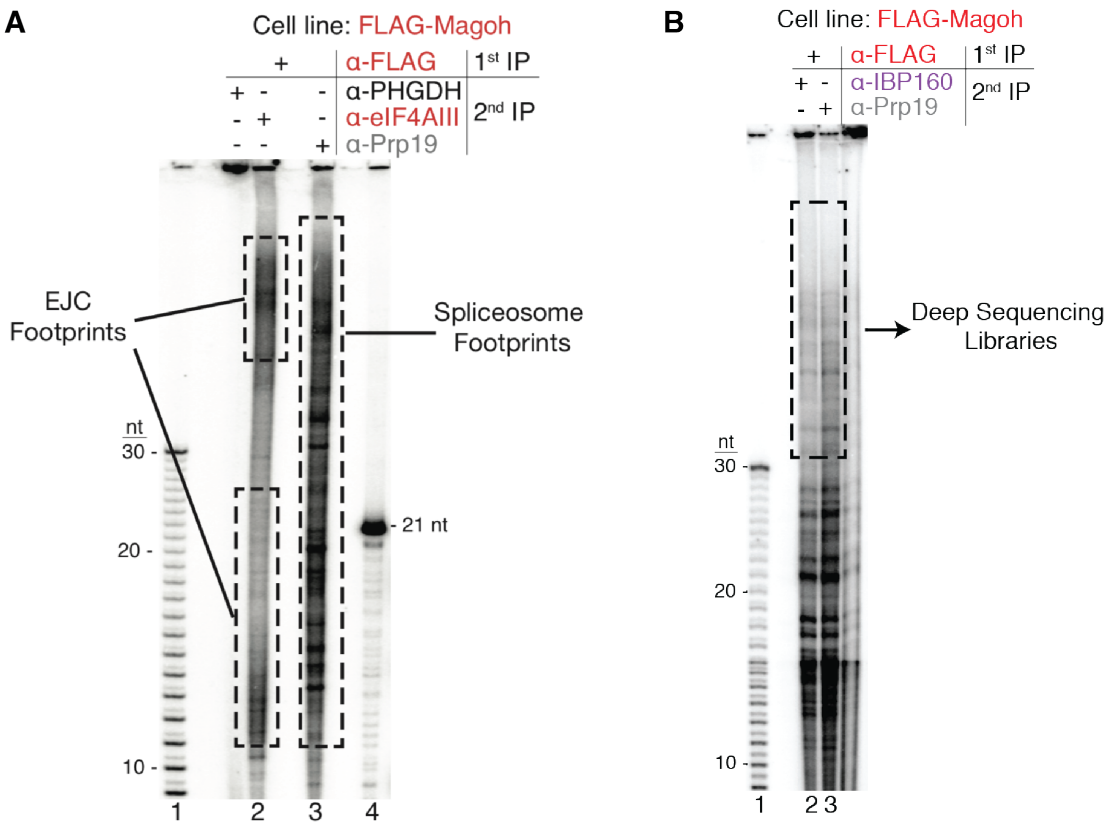

Figure 3.6: (A) Comparison of the length distribution of RNase T1-resistant footprints isolated from EJC (lane 2) or activated spliceosome (lane 3) RIPiTs. Included is a control RIPiT-Seq targeting a protein that does not associate with the EJC or spliceosome, PHGDH. Immunoprecipitated RNA fragments were 5' end labeled then separated by denaturing PAGE. (B) Comparison of the length distribution of RNase T1-resistant footprints isolated from activated spliceosome RIPiTs (lanes 2-3). The primary antibody used during the second IP step was varied between experiments to target different spliceosome complexes. Immunoprecipitated RNA fragments were 5' end labeled then separated by denaturing PAGE. Data courtesy of Guramrit Singh, PhD.

Library			**Sequencing & Alignment**		
Targeted Proteins					
First IP	*Second IP*	*Crosslinking*	*Insert Size*	*Sequenced Reads*	*Aligned Reads*
FLAG-Magoh	**IBP160**	-	30-50	10 Million	8 M
		+	50-80	28 M	26 M
FLAG-Magoh	**Prp19**	-	50-80	36 M	26 M
		+	50-80	35 M	30 M

Table 3.1: Sequencing and alignment information for activated spliceosome RIPiT-Seq libraries.

in cytoplasmic RNA-Seq regardless of crosslinking treatment. Moreover, we identified differences in the distribution of snRNA species between the two library types. U1 snRNA predominated in RNA-Seq libraries, likely due to its splicing-independent function in preventing premature mRNA cleavage and polyadenylation (Kaida et al. 2010). However, late stage spliceosomes contain U2, U5, and

U6 snRNPs after release of U1 and U4 snRNPs (Section 1.1.2); the snRNA read distribution in RIPiT-Seq libraries reflects this composition (Figure 3.7, right).

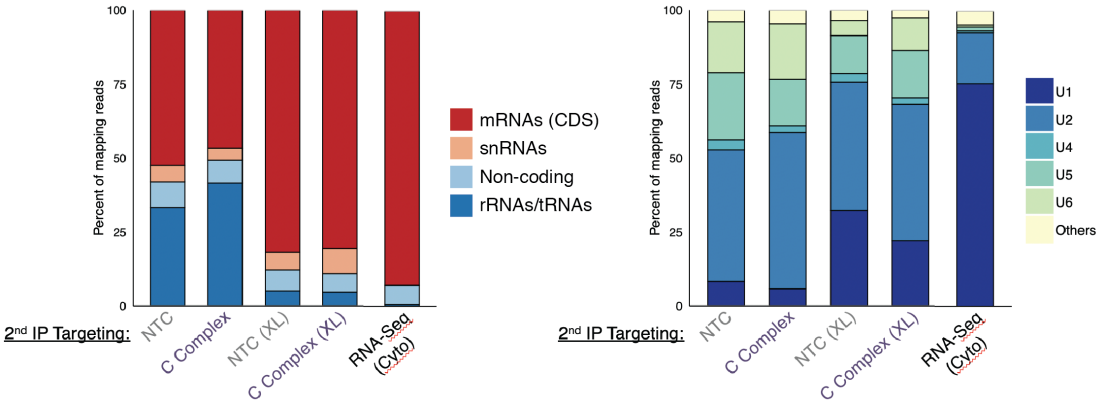

Figure 3.7: Classification of RNA footprints protected by spliceosome complexes. Bar graphs show fraction of mapped reads sorted by classes of RNA (left) or snRNAs (right).

The remaining reads were then mapped to hg19/GRCh37 using STAR with a low threshold for mismatched and gapped (i.e., insertions or deletions) alignments to increase mapping accuracy. We chose STAR as its two-pass alignment method, in which novel splice junctions are observed in the first pass and then used as annotation in the second, is ideal for discovering new exon ligation events (Dobin et al. 2013). Using this alignment tool, we aligned between 8 and 30 million reads in the spliceosome RIPiT-Seq libraries to the genome (Table 3.1).

Although we targeted different proteins in our second IP, both RIPiT-Seq strategies should isolate the same spliceosome complex and, therefore, should have (nearly) identical footprints across mRNA species. To investigate this, we first quantified the level of each mRNA species in the four libraries by calculating their reads per million (RPM) then normalizing to total mapped reads (Figure 3.8). Spliceosome-protected mRNAs were found in similar quantities when using the same antibody in the second IP with or without crosslinking (Figure 3.8A and B). However, coverage was even more equal in abundance when experiments were performed under crosslinked conditions regardless of antibody choice (Figure 3.8C). Overall, these results further confirmed that we had isolated the same late-stage

spliceosome complex using both RIPiT-Seq strategies.

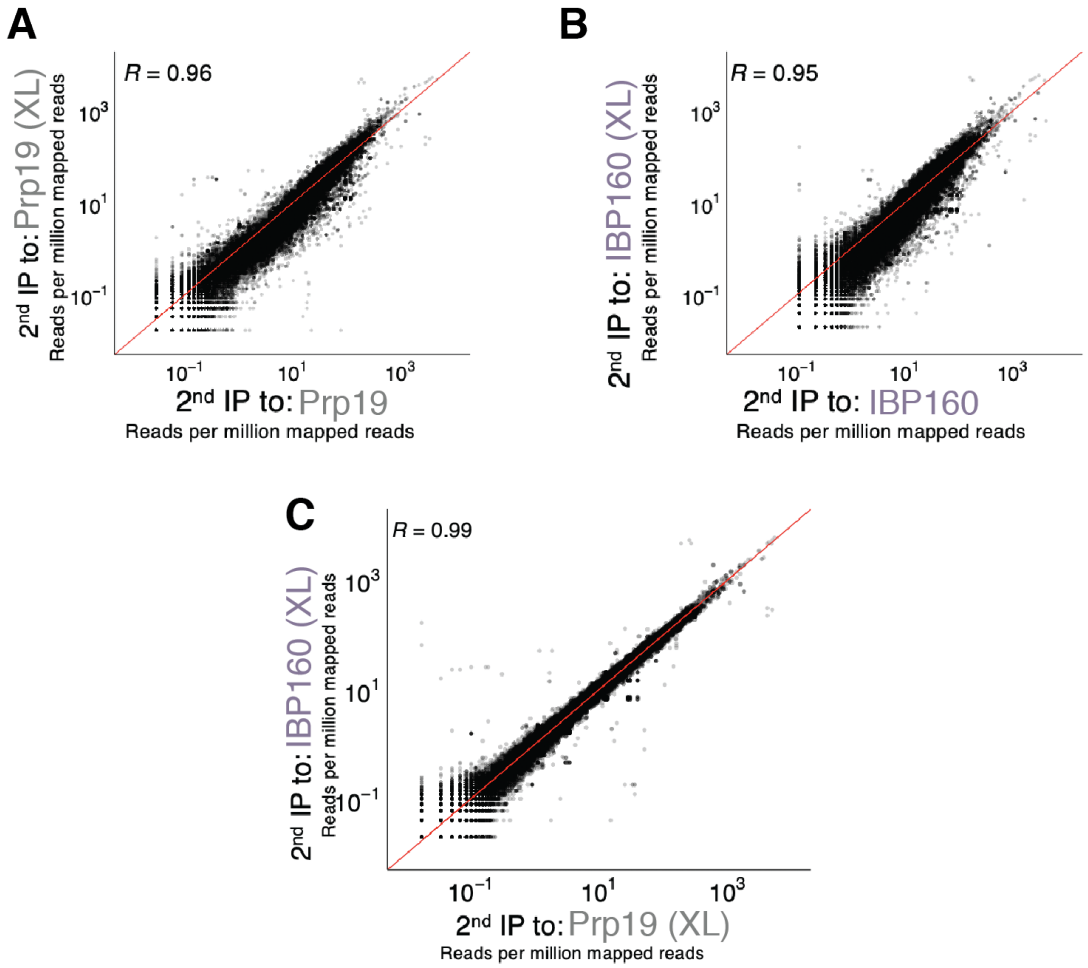

Figure 3.8: Scatterplots comparing reads per million (RPM) between indicated spliceosome RIPiT-Seq libaries. Comparisons are made between RIPiTs performed under native or crosslinked conditions (left) or between crosslinked RIPiTs selecting for different proteins in the second IP (right). R: Pearson's correlation.

3.3.4 Activated spliceosome footprints transcriptome-wide

Next, using the UCSC Genome Browser (Kent et al. 2002; Raney et al. 2014), we explored our data on a gene-by-gene basis across the transcriptome. To identify footprints specific to the activated spliceosome, we compared our libraries to both EJC RIPiT-Seq and cytoplasmic RNA-Seq. As expected, reads in the RNA-Seq library mapped almost exclusively to exons (Figure 3.9, black), and

the EJC (red) primarily bound a short region located ~24 nt upstream of the 5′ exon (Section 1.1.3; Singh et al. 2012). Based on earlier analysis of the *S. pombe* intron lariat complex (Figure 3.3), we anticipated that the footprint of the mammalian spliceosome would also cover intronic regions. However, this is not what we found. Instead, the activated spliceosome predominantly protected exons and splice sites (Figure 3.9, purple and grey). Notably, the intronic signal increased with crosslinking treatment, suggesting chemical crosslinks are necessary to maintain these interactions (Figure 3.9, bottom).

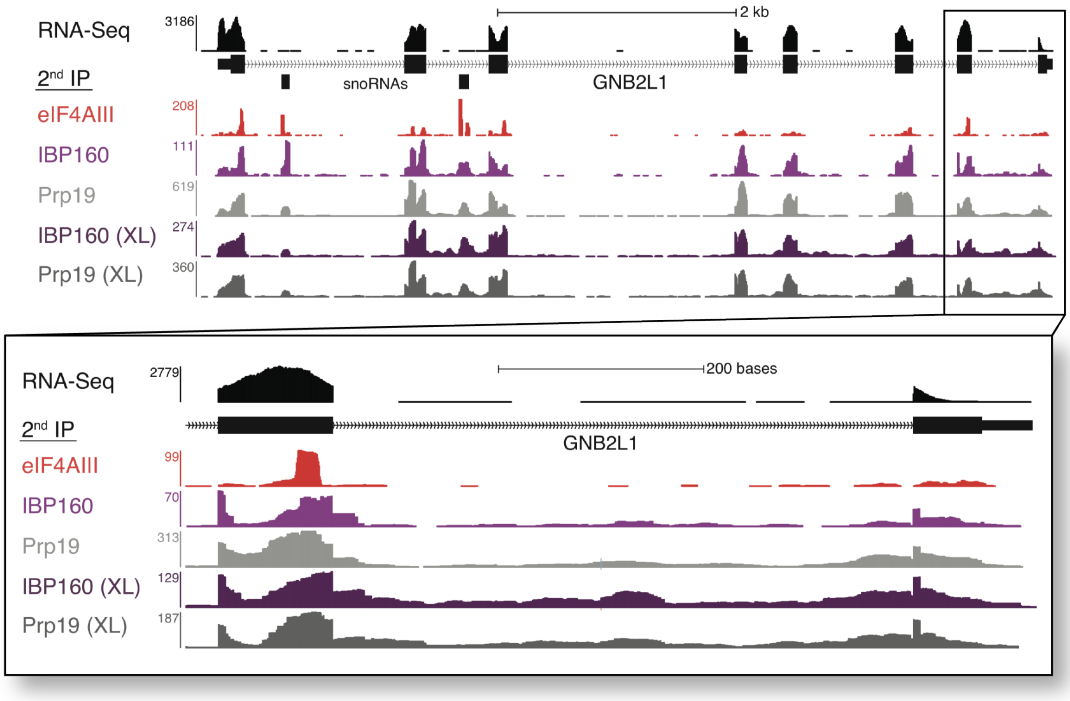

Figure 3.9: Distribution of mapped reads in EJC (red) and spliceosome RIPiT-Seq (grey, purple) libraries under native of crosslinked (XL) conditions. Crosslinking better maintains interactions between the spliceosome and intronic RNA, shown here across the GNB2L1 gene and its final intron (zoomed).

Further exploration across the transcriptome revealed variations in the splicing events recorded by spliceosome RIPiT-Seq and cytoplasmic RNA-Seq. One such example was found in the major and minor isoforms of PKM, which differ based on mutually exclusive exons 9 and 10 (Figure 3.10, bottom). Although we found little evidence of the minor isoform in RNA-Seq, this transcript was found in near equal abundance to the major isoform in crosslinked spliceosome RIPiT-Seq libraries

(Figure 3.10A). This suggested that this isoform was indeed spliced in HEK293 cells, but degraded before it could be captured by cytoplasmic RNA-Seq.

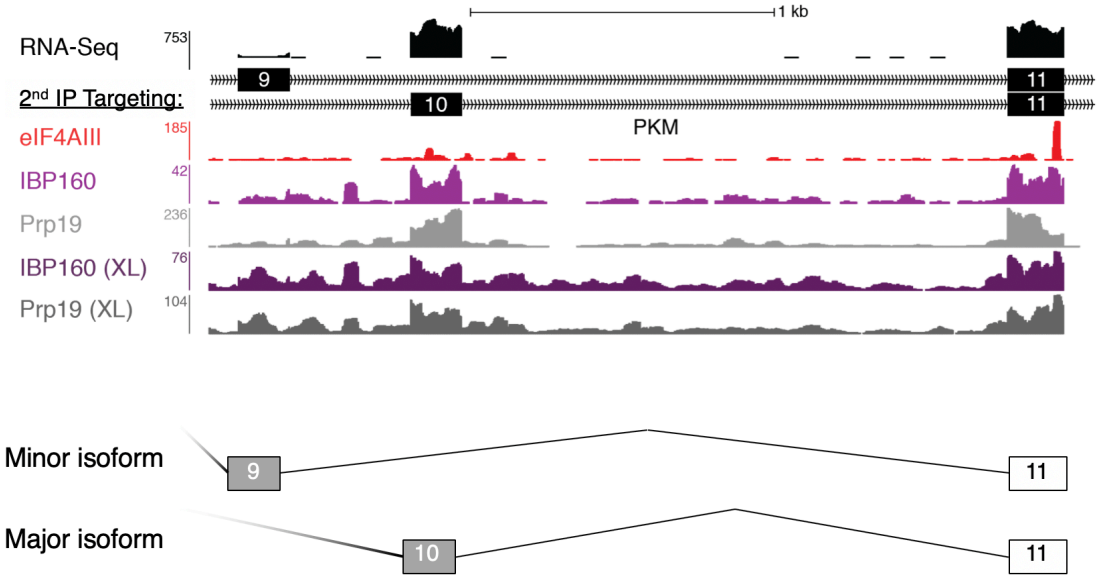

Figure 3.10: Comparison of coverage across the major and minor isoforms of PKM. Colors and labels as described for Figure 3.10.

3.3.5 Lack of evidence of "unannotated" splicing events

Based on this result, we sought out additional evidence of splicing events that eluded RNA-Seq. As transcriptome annotations are largely based on data from RNA-Seq experiments (Section 1.5), we suspected that these splicing events may even be missing from our annotation files. Fortuitously, the combination of EJC and spliceosome footprints highlighted exon junctions (Figure 3.9). Thus, we searched for potential introns by looking for areas with a reduced signal in spliceosome libraries that were preceded by an EJC footprint upstream in the 5′ exon. Based on this pattern, we believed we identified such a previously unannotated intron in the 3′ UTR of *hnRNPA2B1* (Figure 3.11). We confirmed this by examining sequences at the potential novel 5′ and 3′ splice sites and found they were similar to the consensus sequences (Figure 3.11, bottom). However, at this time,

our transcriptome was based entirely ons the RefSeq annotation (Section 1.5.1) and did not contain every previously annotated junction (Figure 1.15). In fact, further examination of annotated transcriptomes revealed this splicing event was previously documented in the ENSEMBL database as an NMD isoform (Section 1.5.2; Cunningham et al. 2019).

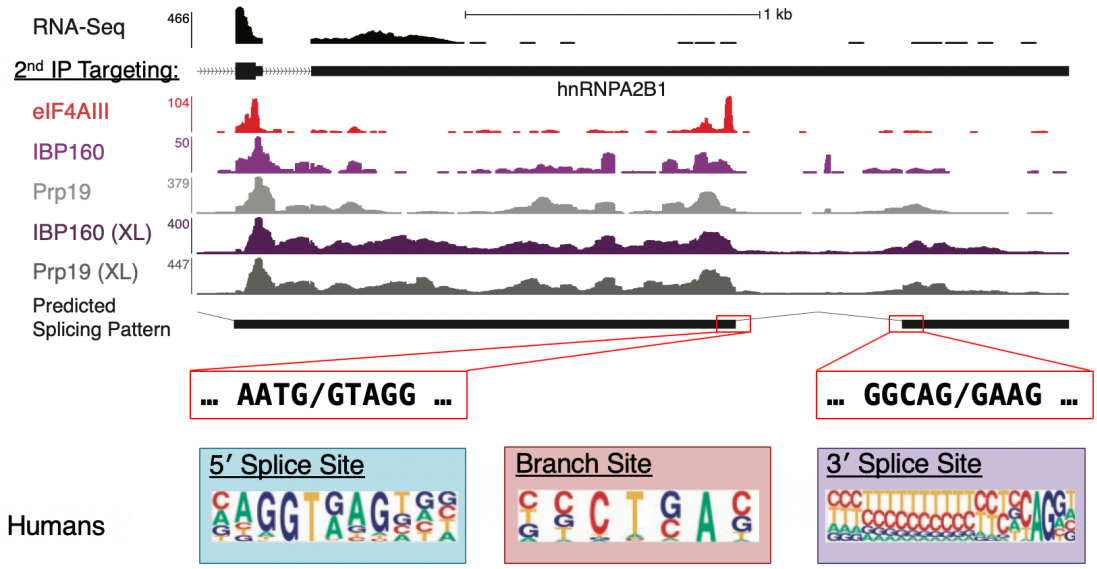

Figure 3.11: Evidence of potentially unannotated splicing events in EJC and spliceosome RIPiT-Seq libraries, shown here in the 3' UTR of hnRNPA2B1. Predicted transcript later identified as an NMD isoform unannotated in the RefSeq database. Colors and labels as described for Figure 3.10.

3.4 Discussion

It was at this time that we developed the pipeline to identify novel splicing events (Figure 2.9) and processed the spliceosome RIPiT-Seq libraries. Unfortunately, due to the shallow sequencing depth and lack of replicates, we were unable to find substantial evidence of unannotated splicing events in these libraries. Further attempts to construct similar libraries failed due to a number of issues (data not shown). Additional cell lines containing FLAG-tagged spliceosome proteins were later constructed to circumvent these problems (Appendix), but additional libraries could not be made before the Moore lab shut down. Thus, we switched to

analyzing these events in EJC RIPiT-Seq libraries that had since been generated in the lab (Chapter 2).

However, there was still much to be gained from this project. We demonstrated that it was possible to isolate late-stage spliceosomes from an established cell line by swapping antibodies used during RIPiT-Seq (Figure 3.4). With one rather inexpensive change, we could either isolate spliceosomes or the EJC (Figures 3.5 and 3.6). It also confirmed that interactions within the spliceosome or between the complex and its pre-mRNA substrate were maintained throughout the RIPiT-Seq protocol. Furthermore, we could compare these libraries with EJC RIPiT-Seq and cytoplasmic RNA-Seq libraries to identify true isoform expression prior to cytoplasmic changes (Figure 3.10 and 3.11). This significantly influenced the analysis we later performed when examining the pre-translational transcriptome (Chapter 2).

Unlike *S. pombe* (Figure 3.3), footprints of the mammalian spliceosome predominantly fell within exonic regions. This result was unexpected and further experimentation is necessary to adequately explain it. However, previous visualization of the spliceosome could provide some insight. In early electron micrograph images, large intron loops are visible during co-transcriptional splicing (Proudfoot 2000). These unbound introns are exposed to RNase T1 during the first IP of spliceosome RIPiT-Seq and likely lost during this step. More recently, Zhang *et al.* have reported the structure of human activated spliceosomes using cryo-electron microscopy (Zhang et al. 2017, 2019). These structures supported minimal interactions between the spliceosome and intronic sequences outside of a small portion of the lariat structure (Zhang et al. 2019). Neither fully explains why we see such a strong exonic signal, even across very long exonic regions (i.e., 3′ UTRs as in Figure 3.11). Intronic coverage did improve with formaldehyde crosslinking, suggesting this treatment during RIPiT-Seq is necessary for isolating intact

spliceosomes much like Staufen1 (Ricci et al. 2014).

Ideally, we would have further modified the RIPiT-Seq strategy to isolate other spliceosome complexes to identify whether earlier (i.e., A complex) or later (i.e., intron lariat complex) complexes had similar footprints or if they behaved more like *S. pombe* spliceosomes (Figure 3.3). Specifically, analysis of earlier stage spliceosome footprints would have allowed us to investigate how far spliceosome assembly progresses at unused splice sites transcriptome-wide. According to the commitment model, spliceosomes should only occupy sites used in mature transcripts. However, results from single molecule experiments suggest that early complexes frequently assemble without leading to splicing, thus the majority of splice sites should be occupied at some time by early stage spliceosomes. By tracking spliceosome occupation and release of pre-mRNAs at discrete stages throughout the assembly cycle, we could discriminate between these models.

3.5 Acknowledgments

I thank Guramrit Singh for preparing the RIPiT-Seq samples used in the bioinformatic analyses discussed in this chapter. I also thank Weijun Chen, Inna Shcherbakova, Makoto Ohira, Joerg Braun, Erin Heyer and Emiliano Ricci for their assistance throughout this project.

3.6 Methods

3.6.1 Spliceosome RIPiT-Seq from crosslinked HEK293 nuclei

All volumes are for one 150mm plate. Scale-up for more plates.

Before RIPiT-Seq experiment:

1. Grow Flp-In FLAG-tagged cells in a 150mm plate to 80-90% confluency in DMEM media.
2. Induce cells with Tet – 25 ng/mL for FLAG-Magoh for 16-20 hours. Combine 10 uL with 1 mL DMEM media, and then add 62.5 uL per plate.

Washing of HEK Cells:

3. Rinse cells twice with 15 mL ice-cold 1X PBS. For multiple plates: keeps plates with PBS on ice while washing others.
4. Use cell scraper to remove cells from plate. Resuspend cells in ice-cold 1X PBS. Collect cells in 50 mL tube.
5. Pellet cells at 500 x g for 5-10 min at 4°C. Remove supernatant.

Preparation and Crosslinking of Nuclei:

6. Resuspend cells in 3 mL ice-cold RSB-100 w/ Digitonin (40 ug/mL).
7. Incubate for 5 min on ice.
8. Pellet nuclei at 2000 x g for 8 min at 4°C. Collect supernatant (cytoplasmic fraction), as much as wanted for analysis.
9. Resuspend nuclei in 10 mL ice-cold 1X PBS per plate.

10. Add 0.1% formaldehyde (27 uL of 37% stock solution) to conical and nutate for 10 min at RT.
11. Add 1 mL quenching buffer and nutate for 5 min at RT.
12. Pellet crosslinked nuclei at 500 x g for 5 min at 4°C.

Preparation of Nuclear Lysate: 13. Resuspend cells in 3 mL ice-cold CLIP Lysis Buffer. Transfer to 5 mL Eppendorf tube. 14. Incubate extract for 10 min on ice. 15. Disrupt the extract by sonicating at 40% amplitude using a Microtip for 8 second pulses per plate (separated by 10 second breaks). 16. Transfer sonicated material to conical and bring up volume with CLIP Lysis Buffer to 3 mL per plate. Combine all material; use 50 mL conical if using more than 3 (if 150mm) plates. 17. Spin at 15,000 x g for 10 min at 4°C. (Supernatant = lysate, pellet = chromatin.) Transfer supernatant to new tube. Transfer 50 uL of supernatant to 1.5 mL Eppendorf tube (input) and freeze for analysis. Resuspend pellet in 100 uL ice-cold 1X PBS and freeze for analysis.

FLAG Immunoprecipitation:

15. Before beginning, wash anti-FLAG agarose beads (250 uL per plate) twice with 1 mL of Isotonic Wash Buffer (IsoWB). After the first wash, transfer to conical (same size as step 16).
16. Transfer lysate (supernatant) to conical with beads.
17. Nutate for 1 hour at 4°C.
18. Pellet beads at 400 x g for 1 min at 4°C. Transfer 50 uL of supernatant to 1.5 mL Eppendorf tube (anti-FLAG depletion).
19. Wash beads four times with ice-cold IsoWB using 3 mL per plate.
20. Resuspend beads in 1 mL ice-cold IsoWB and transfer to a 1.5 mL Eppendorf.
21. Pellet beads at 400 x g for 1 min at 4°C. Remove supernatant.

22. Resuspend beads in ice-cold IsoWB (125 uL per plate) with 10 U/uL RNase T1 (stock: 1000U/uL, use 10 uL per 1 mL IsoWB).
23. Incubate cells on Thermomixer with intermittent shaking (30 seconds shaking at 1000 rpm with 2 min breaks) for 10 min at 37°C. After RNase treatment, transfer to conical (same size as step 13) in 1 mL ice-cold IsoWB.
24. Wash beads four times with ice-cold IsoWB using 3 mL per plate.
25. Resuspend beads in 1 mL ice-cold IsoWB and transfer to a 1.5 mL Eppendorf.
26. Pellet beads at 400 x g for 1 min at 4°C. Remove supernatant.
27. Resuspend beads in ice-cold IsoWB (125 uL per plate) with 250 ug/mL FLAG peptide (stock: 5 mg/mL).
28. Elute by gentle shaking for 1.5 to 2 hours at 4°C. Transfer 40 uL (2 x 1/6th of plate) of eluate to 1.5 mL Eppendorf tube (FLAG elution).
29. To the beads, add 900 uL ice-cold IsoWB, mix with beads, then pellet beads at 400 x g for 1 min at 4°C. Combine this supernatant with previous elutions in 2 mL Eppendorf.

Specific Antibody Immunoprecipitation:

30. Before beginning, conjugate antibody to Protein A/G Dynabeads.
31. Add eluate to conjugated beads.
32. Nutate for 2 hours at 4°C.
33. Wash beads four times with 1 mL ice-cold IsoWB.
34. Resuspend beads in 500 uL ice-cold IsoWB and transfer to new 2 mL Eppendorf.
35. Wash beads four times with 1 mL ice-cold IsoWB. Remove supernatant.
36. Add 7 uL Clear Sample Buffer.
37. Incubate beads in buffer for 5 min at RT and then heat for 5 min at 95°C. Before moving to 95°C, pipette sample to mix.

38. Capture beads on magnet and transfer supernatant (final elution) to 1.5 mL Eppendorf. Keep at -20°C. Transfer 2 uL to separate 1.5 mL Eppendorf (final elution).

RNA Extraction:

39. To reverse crosslinks, heat sample for 45 min at 70°C. Bring up volume of eluate to 100 uL with H2O.
40. Extract twice with P/C/IAA (pH 4.5) and once with C/IAA.
41. Add 10 g glycogen, 0.1 X volume 3M sodium acetate pH 5.2, and three volumes of 100% ethanol.
42. Precipitate overnight at -20°C.
43. Pellet RNA at 13000 x g for 30 min at 4°C.
44. Wash once with 70% ethanol.
45. Resuspend RNA in 20 uL. Use 1 uL to quantify RNA, save remaining RNA for debranching and/or library prep.

Required Buffers

Keep the following buffers at 4°C:

1X PBST Buffer (4°C)	Final Concentration
PBS	1X
Tween20	0.02%

CLIP Lysis Buffer (4°C)	Final Concentration
PBS	1X
SDS	0.1%
Sodium deoxycholate	0.5%

CLIP Lysis Buffer (4°C)	Final Concentration
NP-40	0.5%

Isotonic Wash Buffer (4°C)	Final Concentration
Tris-HCl pH 7.5	20 mM
NaCl	150 mM
NP-40	0.5%

Keep the following buffers at room temperature:

Quenching Buffer (RT)	Final Concentration
Glycine	2M
Tris-HCl pH 7.0	25 mM

Clear Sample Buffer (RT)	Final Concentration
Tris-HCl pH 6.8	100 mM
SDS	4.0%
EDTA	10 mM
DTT (add fresh)	100 mM

3.6.2 Deep sequencing libraries

EJC RIPiT-Seq and cytoplasmic RNA-Seq libraries were downloaded from the NCBI GEO GSE41154 (specifically, samples GSM1009416, GSM1009417, GSM1009418, and GSM1009421).

Spliceosome RIPiT-Seq libraries were generated from 30-80 nt fragments by 3′ adaptor ligation and reverse transcription. Single-end sequencing (100 nt reads) on the Illumina ____ platform resulted in 10 to 36 million reads per library (Table 3.1).

3.6.3 Library processing and alignment

Read counts for unprocessed libraries and for the individual processing steps detailed below are provided in Figure 3.1. Prior to alignment, adaptor sequences and long stretches ($\geq$ 20 nt) of adenosines were trimmed from the 3′ end of sequencing reads. Spliceosome RIPiT-Seq and cytoplasmic RNA-Seq libraries were filtered using STAR v2.5.3a (Dobin et al. 2013) for reads that aligned to repeat regions, as defined by RepeatMasker. Remaining reads were then aligned with STAR on two-pass mode to the human genome, release hg19/GRCh37 (Dobin et al. 2013).

3.6.4 Plotting and data visualization

Data visualization was performed in R using the ggplot2 software package. The UCSC Genome Browser was used to view sequencing library tracks and for transcript figures in this chapter.

Chapter 4

Discussion

4.1 Analysis of the pre-translational transcriptome

The work presented in **Chapter 2** encompasses our detailed analysis of the pre-translational transcriptome in comparison to whole cell and cytoplasmic RNA-Seq. Though RNA-Seq is often used to evaluate differential gene expression, we found that EJC RIPiT-Seq better captured AS-NMD isoforms prior to translation-dependent decay. This provides a better picture of the flux through alternative processing pathways in mammalian cells. Furthermore, our examination of these libraries also revealed evolutionarily conserved splicing events absent from previous RNA-Seq-based annotations.

4.1.1 Flux observed in the pre-translation transcriptome

During the late 2000s, RNA-Seq emerged as the preferred method for evaluating differential gene expression on a global scale. This method provides a static

snapshot of the transcriptome, reflecting the intersection between the rate of transcription and decay for each mRNA. However, decay rates are not universal; an ever-increasing portion of the transcriptome has been annotated as substrates of rapid turnover through translation-dependent quality control pathways (Section 1.4). Over time, many of these transcripts have been shown to have important roles in post-transcriptional regulatory processes in higher eukaryotes.

Unfortunately, such isoforms are often lost in RNA-Seq samples of wild-type cells due to their increased degradation rate. To circumvent this issue, previous RNA-Seq-based studies have abrogated either translation or the responsible degradation pathway by treating cells with small molecule inhibitors, antibiotics, or siRNAs (Section 1.6). Unfortunately, these methods affect more than the intended target and cause confounding pleiotropic effects. For example, commonly used translational inhibitors activate a number of signaling pathways within an hour, sometimes less (Sidhu & Omiecinski 1998). Long-term exposure, as is required for siRNA-mediated knockdown of NMD factors, can be even more severe; within 48 hours of inhibiting NMD, cell death markedly increases (Wengrod et al. 2013). Thus, gene expression measured following any of these treatments is marred to an unknown degree by the TDD-independent changes.

Therefore in an effort to better observe TDD-regulated transcripts without these effects, we turned to previously published EJC RIPiT-Seq libraries (Metkar et al. 2018). This isolation technique captures the mRNA population that has completed splicing but not yet been translated (as EJCs are removed in the process, Section 1.1.3), effectively eliminating the impact of translation-dependent decay in wild type cells. A previous analysis of the EJC-bound mRNAs revealed increased occupancy on AS-NMD transcripts (Section 1.7.2; Singh et al. 2012), suggesting this method could be ideal for studying this specific fraction of the transcriptome. In fact, our comparison of these libraries to both whole cell and cytoplasmic RNA-Seq

(Figure 2.1) revealed EJC RIPiT-Seq improved the isolation of mRNAs destined for translation-dependent decay (Figures 2.3, Figure 2.6B, and 2.7). Furthermore, enrichment was not restricted to this class of transcripts; untranslated, yet spliced, RNAs (e.g., XIST) were also up-regulated in EJC RIPiT-Seq libraries (Figure 2.8).

However, the analysis performed thus far relied on available transcriptome annotations; as these are largely based on RNA-Seq data (Section 1.7), we questioned whether the EJC RIPiT-Seq data contained evidence of mRNAs that eluded this method of detection. To investigate this possibility, we created a bioinformatics pipeline to sequester junction-spanning reads that do not align to previously annotated splicing events (Figure 2.9). As there is some degree of variability between sources (Section 1.5), it was necessary for us to first consolidate annotations from four sources (RefSeq, Section 1.5.1; Ensembl and GENCODE, Section 1.5.2; and CHESS, Section 1.5.3) into a single database of almost 600,000 unique intronic locations. Reads that aligned elsewhere and passed our filtering steps (Figure 2.9) were then classified as originating from unannotated junctions. Though the vast majority of annotated junctions appeared in all three libraries (Figure 2.10B, left), we discovered more than five thousand previously unannotated junctions of which ~68% were present in only EJC RIPiT-Seq libraries (Figure 2.10B, right).

Nonetheless, not all mRNAs in the cell have functional relevance (Section 1.2.1.2) and some are simply the result of aberrant splicing events, or "splicing noise." As previous RNA-Seq-based analysis suggested that upwards of 0.7% of the transcriptome is attributable to such events (Pickrell et al. 2010), we recognized the need to further filter our newly identified junctions. Therefore, we evaluated the annotated and unannotated splice sites used in any of the three libraries based on conservation and MaxENT scores (Figure 2.13). We then focused on unannotated events that scored better than 95% of randomly selected sequences that otherwise mimicked splice sites (i.e., AG/GT dinucleotides at the expected positions; Figure

2.14A and 2.16A). Using this cut-off, we identified nearly 1000 evolutionarily-conserved unannotated 5′ and 3′ splice sites (Figure 2.14B). By examining these highly conserved events individually, we found a number of AS-NMD isoforms resulting from alternative 5′ and 3′ splice sites (Figure 2.15 and 2.17) or novel cassette exons (Figures 2.20 and 2.21).

4.1.2 Limitations of relying on previously published data

Although our analysis of the pre-translational transcriptome proved quite fruitful, the comparison to whole cell and cytoplasmic RNA-Seq libraries was not without flaws. Unfortunately, outside factors necessitated our reliance on the previously published datasets. Based on our initial findings in spliceosome RIPiT-Seq described in Chapter 3, we believed that the EJC RIPiT-Seq libraries prepared by Metkar *et al.* could provide additional insight about the mammalian transcriptome (Metkar et al. 2018). However, these datasets served as a control to new methodology in this publication, eliminating the need for traditional RNA-Seq. Thus, we searched through the literature for potential candidates. We chose the whole cell and cytoplasmic RNA-Seq libraries from the Yaspo lab based on their depth, reproducibility between replicates, and the similarities to the preparation of EJC RIPiT-Seq libraries.

In spite of that, differences between these libraries may affect the quality of this analysis and should be considered. One such disparity is introduced early in the EJC RIPiT-Seq protocol (Metkar et al. 2018). Prior to lysis, cells are treated for an hour with harringtonine, an antibiotic that blocks ribosome translocation. This elongation inhibitor prevents ribosomes from stripping EJCs from mRNAs during the pioneer round of translation (Section 1.1.3.2), leading to the accumu-

lation of transcripts bound by pre-translational RNPs. Though necessary for this methodology, this treatment muddies our analysis. As a consequence of inhibiting translation, translation-dependent decay pathways (Section 1.4) are also inhibited. Thus, we cannot discern whether the enrichment of NMD transcripts in the EJC RIPiT-Seq libraries is due to this or the technique.

Furthermore, a previous study in yeast demonstrated that cycloheximide, a similar inhibitor, caused a rapid increase in the transcription of ribosome biogenesis genes (Santos et al. 2019). However, cells used in this analysis were first starved of amino acids and may not reflect what occurs in mammalian cells under optimal conditions. In fact, Kearse *et al.* recently found that cycloheximide down-regulated transcription in HeLa cells after prolonged exposure (24 hours) but had no effect after a 15 minute treatment (Kearse et al. 2019). Based on these conflicting results, we are uncertain of what changes may be introduced by translation inhibition in our cells. To eliminate this potential source of contamination, further experimentation could eliminate this step during EJC RIPiT-Seq or expose cells to harringtonine before RNA-Seq.

Another major discrepancy between our analyzed libraries is the difference in sequenced read length (150 bp vs 51 bp). Though the reads were appropriately sized for the purposes of both original publications, the longer length of EJC RIPiT-Seq reads may improve their mapability and introduce bias in downstream analyses. A previous study found that computationally shortening 100 bp reads to 50 nt had no discernable influence on calculating differential gene expression (Chhangawala et al. 2015). However, the authors found that splice site identification worsened as reads were shortened. These results suggest that our ability to detect more unannotated splicing events in EJC RIPiT-Seq libraries may be an artifact of longer reads. We could evaluate the impact this had on our current libraries by shortening the EJC RIPiT-Seq reads, either to their first 50 nt or into 50 nt sub-

sections. As either approach would weaken our ability to detect splice sites rather than improve it, we would be better off sequencing longer RNA-Seq reads.

4.1.3 Future experimentation

Over the past decade, RNA-Seq-based analyses have greatly expanded our knowledge of the pervasiveness of gene regulation via AS-NMD in different species, tissues, and conditions (Section 1.4.2). However our data suggests that EJC RIPiT-Seq better captures rapidly degraded transcripts (Figure 2.7), including many resulting from infrequently used, yet conserved splice sites (Figures 2.10 and 2.11). How many of these events have been missed? How much of the regulatory transcriptome remains unknown?

Fortunately the setup for EJC RIPiT-Seq requires only that Magoh, or another core EJC protein (Section 1.1.3; Singh et al. 2012), be fused to a FLAG epitope tag. Given the recent advances in genetic modifications using CRISPR, we can now more easily tag these proteins in a variety of cell lines and organisms. This would allow us to study the pre-translational transcriptome in backgrounds that are not amenable to other methods of enriching TDD-regulated transcripts. For example, complete knockouts of NMD factors (i.e., UPF1 and UPF2) render mice inviable early in embryonic development (reviewed in Nasif et al. 2018). Although the use of conditional knockouts can overcome this issue in some tissues, others cannot withstand losing these essential proteins. On the other hand, FLAG-tagging EJC proteins has thus far shown no negative impact on cell viability or EJC functionality, suggesting we can apply this method without similar issues.

Furthermore, our success with this approach in finding hundreds of novel conserved splicing events suggests that it may be particularly beneficial to studying transcriptomes of cancerous cells and tissues. The link between cancer and

widespread changes to pre-mRNA splicing patterns has been well-established (reviewed in Climente-González et al. 2017), and the transcriptomes of many cancerous samples have already been catalogued in *T*he *C*ancer *G*enome *A*tlas (TCGA; Weinstein et al. 2013). Unfortunately, the vast majority of the data in TCGA has been obtained via exome sequencing methods that only target annotated exonic regions, omitting unannotated events similar to ones we observed. How can we treat the unknown? This was less of a concern when developing treatments that ubiquitously down-regulate NMD, such as small molecules that permit ribosomal read-through of stop codons (reviewed in Nasif et al. 2018; Nomakuchi et al. 2016). However such approaches can have innumerable off-target effects on the transcriptome. Thus, recent therapies have instead shifted to gene-specific NMD inhibition using antisense oligonucleotides (Nomakuchi et al. 2016). As these oligos function by blocking EJC binding events downstream of known PTCs, more thoroughly annotated transcriptomes could provide more potential therapeutic targets.

4.2 Late-stage spliceosome occupancy transcriptome-wide

The work contained within **Chapter 3** describes our success in modifying the previously published RIPiT-Seq protocol (Singh et al. 2014) to instead capture footprints of late-stage mammalian spliceosomes transcriptome-wide (Figures 3.4, 3.5, and 3.6). Though this project remains incomplete, our initial findings demonstrated that the spliceosome can withstand this isolation method even under native conditions (Figure 3.1). However, formaldehyde crosslinking did prove impactful. In fact, we observed better agreement between spliceosome RIPiT-Seq libraries when crosslinked than when using the same antibody during the second IP (Figures 3.7 and 3.8). This suggests that many RNA-protein interactions are not stable enough to survive both IPs, and comparing the two library types to each

other may allow us to identify sites of unstable interactions.

Still, even with limited data, this project was more than simply a proof-of-concept. We began our investigation as an attempt to map the sites of spliceosome action transcriptome-wide, a feat previously accomplished in *S. pombe* by our lab (Chen et al. 2014). Whereas the purified yeast spliceosome had an overwhelmingly intronic footprint (Figure 3.3), we found late-stage spliceosomes predominantly occupied exons (Figure 3.9). Protection of intronic regions improved upon crosslinking, suggesting interactions between the spliceosome and this portion of pre-mRNA substrates is unstable. Fortunately, the exonic bias retained evidence of flux through alternative splicing (Figure 3.10) and quality control pathways (Figure 3.11) before these transcripts were removed in the cytoplasm. Though we did not further investigate these datasets, these observations greatly influenced the analysis discussed in Chapter 1.

4.2.1 Observing sites of spliceosome action

One area in the field of mRNA splicing that remains largely unanswered is how and when splicing choices are decided in the cell. Early analyses suggested that spliceosomes commit to completing the splicing cycle by early complex formation, once the 5′ and 3′ splice sites have been identified (reviewed in Section 3.2). However, more recent single molecule experiments have challenged these conclusions by demonstrating that formation of each subcomplex throughout the assembly cycle is reversible (Hoskins et al. 2011). Unfortunately, the RNA species used during these analyses are typically quite short and cannot adequately reflect the more complex choices encountered in mammalian cells. Thus, our understanding of such cases would be greatly enhanced by a transcriptome-wide map of interactions between pre-mRNAs and the spliceosome.

Results from our initial attempt at creating such a map of mammalian activated spliceosomes can be split into two categories: the expected and unexpected. Based on a previous analysis of yeast spliceosomes, we anticipated a strong bias towards introns (Figure 3.3). This intronic disposition has previously aided the analysis of functional introns in yeast, allowing Chen *et al.* to both correct previous misannotations and identify many new ones (Chen et al. 2014, 2018). Though we did not find a bias to the same extent in the mammalian system, we observed extensive protection of intronic regions in all library types (Figures 3.9, 3.10, and 3.11).

Furthermore, the yeast spliceosome largely occupied sequences near the intronic boundaries, particularly in longer introns (Chen et al. 2014, 2018). In agreement with this finding, recent cryo-EM structures have demonstrated that late-stage yeast spliceosomes protect 38 nucleotides within the intron - 15 nt downstream of the 5′ splice site and 23 nt around the branch point (reviewed in Chen et al. 2018). Comparatively, the structure of activated mammalian spliceosomes indicates that these complexes bind to 59 nucleotides attributable to the intron lariat and 5′ exon (Zhang et al. 2017). However, these structures fail to account for why we often observed spliceosome occupation extending across the full length of the intron. It does appear that crosslinking and harsher wash conditions may enrich for protected fragments near the 5′ and 3′ splice sites as well as internally (Figures 3.9 and 3.10), but further experimentation is necessary to determine the full impact of this treatment.

Unexpectedly, we found spliceosome occupation of exonic regions to be particularly resistant to RNase T1 digestion compared to neighboring introns. In fact, exon coverage further increased under native conditions, suggesting these interactions between late-stage spliceosomes and exons are incredibly stable (Figures 3.9, 3.10, and 3.11). This result may be best explained by earlier *in vitro* experiments that found increased rates of splicing on multi-intron-containing transcripts (Crabb et

al. 2010). The authors determined this enhancement resulted from retention of the spliceosome, specifically the exon definition complex, after successful removal of nearby introns. Alternatively, it is also possible that the exonic signal could result from exonic DNA contamination due to continued association between pre-mRNAs and chromatin until transcripts are fully spliced (Bhatt et al. 2012; Pandya-Jones et al. 2013). However, both the spliceosome RIPiT-Seq (Section 3.6.1) and library preparation protocols (Heyer et al. 2015) should prevent us from sequencing such an impurity. Like the intronic signal, further experimentation and analysis is needed to understand this pattern of exonic coverage.

One feature apparent in the early analysis of this data is the enhanced coverage in spliceosome RIPiT-Seq libraries across alternatively spliced transcripts (Figures 3.10 and 3.11). Though regions specific to both of the shown isoforms are occupied by late-stage spliceosomes, we find little evidence of their existence in cytoplasmic RNA-Seq. In the case of PKM, the expression pattern of major and minor isoforms agrees with RNA-Seq datasets in the TCGA that were constructed from 16 different tumor samples (Desai et al. 2014). Previous studies have suggested the expression of the minor PKM isoform is downregulated by increased binding of three hnRNP proteins near this exon (David et al. 2010). Moreover, crosslinked IPs of hnRNPA2B1 found a similar binding pattern in its own 3′ UTR (Martinez et al. 2016), where we find increased late-stage spliceosome occupancy (Figure 3.11). If hnRNP proteins suppress spliceosome assembly (Section 1.2.1.3), why do we see EJC-containing spliceosomes in these regions? Spliceosome occupation may signify these regulators significantly slow rather than outright prevent assembly or promote the reversal of late-stage assembly before completion of the cycle. Additional analysis of earlier and later complexes, and comparison to known binding profiles of other RNA-binding proteins, could help elucidate this method of regulation.

4.2.2 Limitations of current spliceosome RIPiT-Seq strategy

Although we successfully modified the EJC RIPiT-Seq protocol to isolate late-stage spliceosomes, further experimentation using this exact setup is somewhat confined in its scope. This is a consequence of relying on a cell line that contains a FLAG-tagged EJC protein rather than targeting a spliceosomal protein for the analysis performed in Chapter 3. However, this strategy was the most practical option at first because it allowed us to begin investigating mammalian spliceosome footprints without first establishing a new cell line. Unfortunately, by using these HEK293 cells, only Magoh-containing spliceosomes can be pulled down in the first IP, which limits the stage at which spliceosomes can be isolated during the second. Thus, this setup in its current state precludes the option of isolating earlier or later complexes.

After this project began, other labs since reported the crystal structures of various human spliceosomal complexes throughout the assembly cycle. These structures both confirmed the absence of the EJC in the pre-activation B complex (Bertram et al. 2017) and its release during the formation of the intron lariat structure (Zhang et al. 2019). Moreover, structures of activated spliceosomes showed that Magoh and other EJC components bind to the periphery of the spliceosome rather than near its core (Figure 4.1). In this confirmation, eIF4AII interacts extensively with the spliceosome, particularly with splicing factors Cwc22 and Snu114, whereas other core EJC proteins contact only each other and eIF4III. The location of Magoh in this complex may explain why this particular EJC protein was amenable to isolating spliceosomes during the RIPiT-Seq protocol.

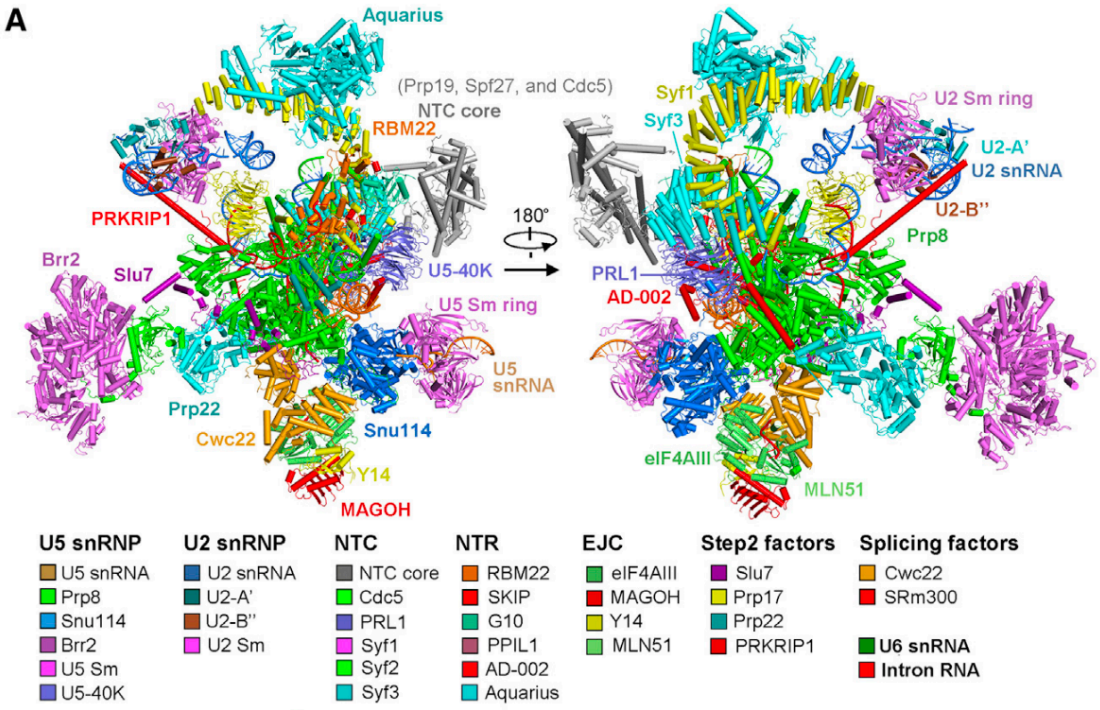

Figure 4.1: Crystal structure of the human C complex spliceosome including three snRNAs (U2, U5, and U6), the free 5' exon, the intron lariat, and nearly 50 proteins. RNAs and proteins are color coded and categorized below by their primary complex association. Figure from Zhang et al. 2017. Copyright 2017 Elsevier.

Bearing this in mind, these structures could guide us in choosing spliceosomal proteins that are potentially the most accommodating to epitope tagging. Though the FLAG tag is a relatively short addition, the fusion of any epitope tag to a protein of interest could interfere with both binding and/or functionality (reviewed in Singh et al. 2014). This is particularly important to consider when tagging spliceosomal proteins as countless protein-protein and protein-RNA interactions mediate progression through the spliceosome assembly cycle (Section 1.1.2). An epitope tag that interferes with these required interactions would inhibit splicing transcriptome-wide and cause a lethal phenotype. Furthermore, even if the tagged protein behaved normally, spliceosomal components located within the core of the macromolecular complex are likely inaccessible during IPs. In fact, our success in spliceosome RIPiT-Seq experiments may be due to the peripheral location of both Prp19 and IBP160 (a.k.a Aquarius), the two proteins targeted in the second IP, in Magoh-containing spliceosomes (Figure 4.1). Thus, the ideal candidates when

creating new FLAG-tagged cell lines are spliceosomal proteins that are similarly located, that can function in spite of a new tag, and that associate with multiple complexes throughout assembly. Initial efforts in establishing such cell lines is included in Appendix 1.

4.2.3 Potential applications of mammalian spliceosome footprinting

4.2.3.1 Evidence of splicing catalysis

Due to the premature termination of this project, there remains much to be explored using this methodology concerning spliceosome occupancy and action in mammalian cells. Splicing of human introns faces a number of complex challenges that cannot be investigated in the simpler yeast system.

For example, some mammalian introns can be quite long - some even longer than 100,000 nt and littered with cryptic splice sites! Though these sequences may resemble 5′ and 3′ splice sites, the final mRNA record only reflects a single splicing event between the extreme ends of the long intron. Is the intron truly removed in a single event? Or does it occur in piecemeal though the excision of shorter, nested fragments until the full intron has been spliced? The latter, known as recursive or nested splicing, was first discovered in *Drosophila* (Hatton et al. 1998), then later confirmed to occur during the processing of an 110 kb intron in the human dystrophin pre-mRNA (Suzuki et al. 2013). As evidence of such events is erased after the final exon ligation event, how can we identify sites of spliceosome action transcriptome-wide?

Fortunately, records of spliceosome catalysis are also maintained in the byproduct of the splicing cycle, the intron lariat. Though the lariat corresponds to the full

length of an excised intron, regions downstream of the 5′ exon and surrounding the branch point can both be identified by sequencing reads that transverse the branched structure. However, this Y-shaped structure impedes reverse transcriptase processivity during RNA-Seq library preparation, causing these reads to be sequenced very infrequently (Taggart et al. 2012). In previous studies, extensive deep sequencing libraries and methodology tailored to capturing this specific structure have been used to map branch points transcriptome-wide (Taggart et al. 2012; Mercer et al. 2015). Yet these methods identified less than 20% of previously annotated introns and showed a propensity towards shorter introns. Thus, by further modifying spliceosome RIPiT-Seq to isolate lariat-containing complexes like the work performed by Chen *et al.* in yeast (Chen et al. 2018), we could both annotate additional branch point locations and seek evidence of recursive splicing reactions.

4.2.3.2 Spliceosome assembly at unused splice sites

Although this project was performed in an established HEK293 cell line for reasons outlined in Section 4.2.2, we initially intended to use this methodology to study *a*lternative *s*plicing (AS) in primary cells or tissues. In these backgrounds, the expression of many mRNA isoforms is known to occur in a stimulus-dependent fashion (Eric T. Wang et al. 2008). Specifically, widespread AS events have been implicated in fine-tuning the innate immune response when either dendritic cells or macrophages interact with a pathogen (Rodrigues et al. 2013; Wells et al. 2006). When treated with the Gram-negative bacterial membrane protein lipopolysaccharide, macrophages activate the NF- B signal transduction pathway (Ma et al. 2003), triggering the expression of dozens of alternatively spliced isoforms (Wells et al. 2006).

This provides us with a means of controlling the appearance of specific isoforms, and therefore the use of specific splice sites. In cases of isoform switching, do late stage spliceosomes occupy unused sites in unstimulated cells? Pre-assembly of the early spliceosome complexes would allow faster splice site recognition, and hence faster immune response. However, when analyzing AS events triggered by the acute inflammatory response, the Smale and Black labs determined that constitutive introns splice faster than alternatively spliced cassette exons (Bhatt et al. 2012). This methodology could not pinpoint which step in assembly causes this lag. By footprinting spliceosomal complexes throughout the assembly cycle before and after activation, we can determine whether assembly begins *de novo* at previously unused splice sites or if early complexes are poised on pre-mRNAs for immediate use.

Appendix: FLAG-tagging components of the mammalian spliceosome

4.3 Introduction

Although using the previously established EJC-tagged HEK293 cell line for spliceosome RIPiT-Seq allowed us to begin the project immediately, we recognized the limitations imposed by this setup (Section 4.2.2). Thus, while we performed these experiments, we also created new cell lines by FLAG-tagging components of the spliceosomal. When we first selected which proteins to tag, the yeast and human spliceosome cryo-EM structures had not yet been released. As such, we evaluated candidates based on two criteria. First, we picked proteins that had been previously immunoprecipitated successfully when fused to FLAG or another epitope tag. Earlier studies revealed that tags can alter the normal interactions of an RNA binding protein (Chan et al. 2004; reviewed in Singh et al. 2014); we wanted to avoid wasting time on such an outcome. Furthermore, we focused on proteins that allowed us to isolate a variety of complexes. Having demonstrated that modifying the second IP in RIPiT-Seq was sufficient to isolate either the EJC or spliceosome from one cell line, we wanted the same flexibility in future experiments. With this

in mind, we used the available mass spectrometry analysis of discrete spliceosome complexes (Figure 1.4) to inform our choices.

4.4 Results

Based on these criteria, we selected four ideal candidates for spliceosome RIPiT-Seq: Prp19, Prp17, TFIP11, and U2A′. In order to express FLAG-tagged versions of these proteins in HEK293 cells, we used the FLP-In system (Methods). This transfection method stably integrates an epitope-tagged protein of interest into the genome under the control of a tetracycline-regulated promoter. An inducible promoter provides a means of fine-tuning the expression of FLAG-tagged proteins to match the levels of endogenous copies (data not shown). This is done to ensure that our experiments more accurately reflect endogenous RNPs.

In total, we created seven new cell lines with one of four tagged spliceosomal proteins to complement our EJC-tagged cells (Table 4.1).

Table 4.1: Available FLAG-tagged HEK293 cell lines for spliceosome RIPiT-Seq experiments.

Protein	Spliceosome Complex	Terminus	Published in
eIF4AIII	B to EJC	N	Singh et al. 2012
Magoh	B to EJC	N	Singh et al. 2012
Prp19	B* to ILS	N	
Prp17	C to C*	N and C	
TFIP11	B to ILS	N and C	
U2A’	A to C*	N and C	

4.5 Acknowledgments

I thank Guramrit Singh, Joerg Braun, Makoto Ohira, and Mihir Metkar for their guidance on cell culture experiments and transfection methods.

4.6 Methods

4.6.1 FLP-In transfection in HEK293 cells

FLAG-tagged proteins of interest were stably integrated into HEK293 cells using the FLP-In System available from ThermoFisher Scientific (#K6500-01). Once established, cells were stored in liquid nitrogen and catalogued.

References

Aken BL et al. 2016. The Ensembl gene annotation system. Database. 1–19. doi: 10.1093/database/baw093.

Andersen CBF et al. 2006. Structure of the exon junction core complex with a trapped DEAD-Box ATPase bound to RNA. Science. 313:1968–1972. doi: 10.1126/science.1131981.

Anderson JS, Parker R. 1998. The 3' to 5' degradation of yeast mRNAs is a general mechanism for mRNA turnover that requires the SK12 DEVH box protein and 3' to 5' exonucleases of the exosome complex. EMBO J. 17:1497–1506. doi: 10.1093/emboj/17.5.1497.

Ardini E, Agresti R, Tagliabue E, Greco M, Aiello P, Yang L-T, Ménard S, Sap J. 2000. Expression of protein tyrosine phosphatase alpha (RPTPα) in human breast cancer correlates with low tumor grade, and inhibits tumor cell growth in vitro and in vivo. Oncogene. 19:4979–4987. doi: 10.1038/sj.onc.1203869.

Barbosa-Morais NL et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. Science. 338:1587–1593. doi: 10.1126/science.1230612.

Beilharz TH, Preiss T. 2007. Widespread use of poly(A) tail length control to accentuate expression of the yeast transcriptome. RNA. 13:982–997. doi: 10.1261/rna.569407.

Belew AT, Hepler NL, Jacobs JL, Dinman JD. 2008. PRFdb: a database of computationally predicted eukaryotic programmed -1 ribosomal frameshift signals. BMC Genomics. 9:339. doi: 10.1186/1471-2164-9-339.

Belew AT, Meskauskas A, Musalgaonkar S, Advani VM, Sulima SO, Kasprzak WK, Shapiro BA, Dinman JD. 2014. Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway. Nature. 512:265–269. doi: 10.1038/nature13429.

Berget SM. 1995. Exon recognition in vertebrate splicing. J of Biol Chem. 270:2411–2414. doi: 10.1074/jbc.270.6.2411.

Berget SM, Moore C, Sharp PA. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. PNAS. 74:3171–3175. doi: 10.1073/pnas.74.8.3171.

Bertram K et al. 2017. Cryo-EM structure of a pre-catalytic human spliceosome primed for activation. Cell. 170:701–713.e11. doi: 10.1016/j.cell.2017.07.011.

Bhatt DM, Pandya-Jones A, Tong AJ, Barozzi I, Lissner MM, Natoli G, Black DL, Smale ST. 2012. Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell. 150:279–290. doi: 10.1016/j.cell.2012.05.043.

Bicknell AA, Cenik C, Chua HN, Roth FP, Moore MJ. 2012. Introns in UTRs: Why we should stop ignoring them. BioEssays. 34:1025–1034. doi: 10.1002/bies.201200073.

Bousquet-Antonelli C, Presutti C, Tollervey D. 2000. Identification of a regulated pathway for nuclear pre-mRNA turnover. Cell. 102:765–775. doi: 10.1016/S0092-8674(00)00065-9.

Braun JE, Friedman LJ, Gelles J, Moore MJ. 2018. Synergistic assembly of human pre-spliceosomes across introns and exons. eLife. 7:1–18. doi: 10.7554/eLife.37751.

Brawerman G. 1987. Determinants of messenger RNA stability. Cell. 48:5–6. doi: 10.1016/0092-8674(87)90346-1.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. Nature Biotech. 34:525–527. doi: 10.1038/nbt.3519.

Breathnach R, Benoist C, O'Hare K, Gannon F, Chambon P. 1978. Ovalbumin gene: Evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. PNAS. 75:4853–4857. doi: 10.1073/pnas.75.10.4853.

Brow DA. 2002. Allosteric cascade of spliceosome activation. Ann Rev of Genetics. 36:333–360. doi: 10.1146/annurev.genet.36.043002.091635.

Busch A, Hertel KJ. 2012. Evolution of SR protein and hnRNP splicing regulatory factors. WIREs RNA. 3:1–12. doi: 10.1002/wrna.100.

Bühler M, Steiner S, Mohn F, Paillusson A, Mühlemann O. 2006. EJC-independent degradation of nonsense immunoglobulin-μ mRNA depends on 3' UTR length. NSMB. 13:462–464. doi:

10.1038/nsmb1081.

Carter MS, Doskow J, Morris P, Li S, Nhim RP, Sandstedt S, Wilkinson MF. 1995. A regulatory nechanism that detects premature nonsense codons in T-cell receptor transcripts. Biol Chem. 270:28995–29003.

Carter MS, Li S, Wilkinson MF. 1996. A splicing-dependent regulatory mechanism that detects translation signals. EMBO J. 15:5965–5975. doi: 10.1002/j.1460-2075.1996.tb00983.x.

Chan CC, Dostie J, Diem MD, Feng W, Mann M, Rappsilber J, Dreyfuss G. 2004. eIF4A3 is a novel component of the exon junction complex. RNA. 10:200–209. doi: 10.1261/rna.5230104.

Chen G et al. 2013. Incorporating the human gene annotations in different databases significantly improved transcriptomic and genetic analyses. RNA. 19:479–489. doi: 10.1261/rna.037473.112.

Chen W, Ashar-Patel A, Yan J, Moore MJ, Shulha HP, Rhind N, Weng Z, Green KM, Query CC. 2014. Endogenous U2 • U5 • U6 snRNA complexes in S. pombe are intron lariat spliceosomes. RNA. 20:308–320. doi: 10.1261/rna.040980.113.

Chen W, Moore J, Ozadam H, Shulha HP, Rhind N, Weng Z, Moore MJ. 2018. Transcriptome-wide interrogation of the functional intronome by spliceosome profiling. Cell. 173:1031–1044.e13. doi: 10.1016/j.cell.2018.03.062.

Chhangawala S, Rudy G, Mason CE, Rosenfeld JA. 2015. The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome Biology. 16:1–10. doi: 10.1186/s13059-015-0697-y.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell. 12:1–8. doi: 10.1016/0092-8674(77)90180-5.

Climente-González H, Porta-Pardo E, Godzik A, Eyras E. 2017. The functional impact of alternative splicing in cancer. Cell Reports. 20:2215–2226. doi: 10.1016/j.celrep.2017.08.012.

Coller J, Parker R. 2004. Eukaryotic mRNA decapping. Ann Rev of Biochem. 73:861–890. doi: 10.1146/annurev.biochem.73.011303.074032.

Colombo M, Karousis ED, Bourquin J, Bruggmann R, Mühlemann O. 2017. Transcriptome-wide identification of NMD-targeted human mRNAs reveals extensive redundancy between SMG6-

and SMG7-mediated degradation pathways. RNA. 23:189–201. doi: 10.1261/rna.059055.116.

Crabb TL, Lam BJ, Hertel KJ. 2010. Retention of spliceosomal components along ligated exons ensures efficient removal of multiple introns. RNA. 16:1786–1796. doi: 10.1261/rna.2186510.

Cremer KJ, Silengo L, Schlessinger D. 1974. Polypeptide formation and polyribosomes in Escherichia coli treated with chloramphenicol. J of Bacteriology. 118:582–589.

Cunningham F et al. 2019. Ensembl 2019. Nucleic Acids Research. 47:D745–D751. doi: 10.1093/nar/gky1113.

Curwen V et al. 2004. The Ensembl automatic gene annotation system. Genome Research. 1–9. doi: 10.1101/gr.1858004.942.

David CJ, Chen M, Assanah M, Canoll P, Manley JL. 2010. HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. Nature. 463:364–368. doi: 10.1038/nature08697.

Decker CJ, Parker R. 1993. A turnover pathway for both stable and unstable mRNAs in yeast: Evidence for a requirement for deadenylation. Genes & Development. 7:1632–1643. doi: 10.1101/gad.7.8.1632.

Desai S et al. 2014. Tissue-specific isoform switch and DNA hypomethylation of the pyruvate kinase PKM gene in human cancers. Oncotarget. 5:8202–8210. doi: 10.18632/oncotarget.1159.

De Semir D et al. 2012. Pleckstrin homology domain-interacting protein (PHIP) as a marker and mediator of melanoma metastasis. PNAS. 109:7067–7072. doi: 10.1073/pnas.1119949109.

Desmet FO, Hamroun D, Lalande M, Collod-Bëroud G, Claustres M, Béroud C. 2009. Human Splicing Finder: An online bioinformatics tool to predict splicing signals. Nucleic Acids Research. 37:1–14. doi: 10.1093/nar/gkp215.

Deveson IW et al. 2018. Universal alternative splicing of noncoding exons. Cell Systems. 6:245–255.e5. doi: 10.1016/j.cels.2017.12.005.

Dinman JD. 2012. Mechanisms and implications of programmed translational frameshifting. WIREs RNA. 3:661–673. doi: 10.1002/wrna.1126.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29:15–21. doi:

10.1093/bioinformatics/bts635.

Doma MK, Parker R. 2006. Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. Nature. 440:561–564. doi: 10.1038/nature04530.

Dostie J, Dreyfuss G. 2002. Translation is required to remove Y14 from mRNAs in the cytoplasm. Current Biology. 12:1060–1067. doi: 10.1016/S0960-9822(02)00902-8.

Dou Y, Fox-Walsh KL, Baldi PF, Hertel KJ. 2006. Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. RNA. 12:2047–2056. doi: 10.1261/rna.151106.

Down T et al. 2002. The Ensembl genome database project. Nucleic Acids Research. 30:38–41. doi: 10.1093/nar/30.1.38.

Drechsel G, Kahles A, Kesarwani AK, Stauffer E, Behr J, Drewe P, Ratsch G, Wachter A. 2013. Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome. The Plant Cell. 25:3726–3742. doi: 10.1105/tpc.113.115485.

Fox-Walsh KL, Hertel KJ. 2009. Splice-site pairing is an intrinsically high fidelity process. PNAS. 106:1766–1771. doi: 10.1073/pnas.0813128106.

Frankish A et al. 2019. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research. 47:D766–D773. doi: 10.1093/nar/gky955.

Frankish A et al. 2015. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genomics. 16:S2. doi: 10.1186/1471-2164-16-s8-s2.

Franks TM, Lykke-Andersen J. 2008. The control of mRNA decapping and P-body formation. Molecular Cell. 32:605–615. doi: 10.1016/j.molcel.2008.11.001.

Frischmeyer PA, Van Hoof A, O'Donnell K, Guerrerio AL, Parker R, Dietz HC. 2002. An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. Science. 295:2258–2261. doi: 10.1126/science.1067338.

Garneau NL, Wilusz J, Wilusz CJ. 2007. The highways and byways of mRNA decay. Nature Rev MCB. 8:113–126. doi: 10.1038/nrm2104.

Gehring NH, Neu-Yilik G, Schell T, Hentze MW, Kulozik AE. 2003. Y14 and hUpf3b form an NMD-activating complex. Molecular Cell. 11:939–949. doi: 10.1016/S1097-2765(03)00142-4.

Goullet de Rugy T, Bashkurov M, Datti A, Betous R, Guitton-Sert L, Cazaux C, Durocher D, Hoffmann JS. 2016. Excess Polθ functions in response to replicative stress in homologous recombination-proficient cancer cells. Biology Open. 5:1485–1492. doi: 10.1242/bio.018028.

Gu Z, Fang X, Li C, Chen C, Liang G, Zheng X, Fan Q. 2017. Increased PTPRA expression leads to poor prognosis through c-Src activation and G1 phase progression in squamous cell lung cancer. Intl J of Oncology. 51:489–497. doi: 10.3892/ijo.2017.4055.

Guthrie C. 1991. Messenger RNA splicing in yeast: clues to why the spliceosome is a ribonucleoprotein. Science. 253:157–163. doi: 10.1126/science.1853200.

Guydosh NR, Green R. 2014. Dom34 rescues ribosomes in 3' untranslated regions. Cell. 156:950–962. doi: 10.1016/j.cell.2014.02.006.

Guydosh NR, Kimmig P, Walter P, Green R. 2017. Regulated Ire1-dependent mRNA decay requires no-go mRNA degradation to maintain endoplasmic reticulum homeostasis in S. Pombe. eLife. 6. doi: 10.7554/eLife.29216.

Hamer DH, Leder P. 1979. Splicing and the formation of stable RNA. Cell. 18:1299–1302. doi: 10.1016/0092-8674(79)90240-X.

Hamid FM, Makeyev EV. 2014. Regulation of mRNA abundance by polypyrimidine tract-binding protein-controlled alternate 5' splice site choice. PLoS Genetics. 10. doi: 10.1371/journal.pgen.1004771.

Hannigan MM, Zagore LL, Licatalosi DD. 2018. Mapping transcriptome-wide protein-RNA interactions to elucidate RNA regulatory programs. Quant Biology. 6:289–313. doi: 110.1016/j.bbi.2017.04.008.

Harigaya Y, Parker R. 2012. Global analysis of mRNA decay intermediates in Saccharomyces cerevisiae. PNAS. 109:11764–11769. doi: 10.1073/pnas.1119741109.

Harrow J et al. 2012. GENCODE: The reference human genome annotation for the ENCODE project. Genome Research. 22:1760–1774. doi: 10.1101/gr.135350.111.

Hatton AR, Subramaniam V, Lopez AJ. 1998. Generation of alternative Ultrabithorax isoforms

and stepwise removal of a large intron by resplicing at exon-exon junctions. Molecular Cell. 2:787–796. doi: 10.1016/S1097-2765(00)80293-2.

He F, Celik A, Wu C, Jacobson A. 2018. General decapping activators target different subsets of inefficiently translated mRNAs. eLife. 7:1–30. doi: 10.7554/eLife.34409.

Heyer EE, Ozadam H, Ricci EP, Cenik C, Moore MJ. 2015. An optimized kit-free method for making strand-specific deep sequencing libraries from RNA fragments. Nucleic Acids Research. 43:e2–e2. doi: 10.1093/nar/gku1235.

Hodgkin J, Papp A, Pulak R, Ambros V, Anderson P. 1989. A new kind of informational suppression in the nematode Caenorhabditis elegans. Genetics. 123:301–313.

Hoskins A a et al. 2011. Ordered and Dynamic Assembly of Single Spliceosomes. Science. 331:1289–1295. doi: 10.1126/science.1198830.

Hu W, Petzold C, Coller J, Baker KE. 2010. Nonsense-mediated mRNA decapping occurs on polyribosomes in Saccharomyces cerevisiae. NSMB. 17:244–247. doi: 10.1038/nsmb.1734.

Hu W, Sweet TJ, Chamnongpol S, Baker KE, Coller J. 2009. Co-translational mRNA decay in Saccharomyces cerevisiae. Nature. 461:225–229. doi: 10.1038/nature08265.

Hu Z, Yau C, Ahmed AA. 2017. A pan-cancer genome-wide analysis reveals tumour dependencies by induction of nonsense-mediated decay. Nature Comm. 8:1–9. doi: 10.1038/ncomms15943.

Hug N, Longman D, Cáceres JF. 2015. Mechanism and regulation of the nonsense-mediated decay pathway. Nucleic Acids Research. 44:1483–1495. doi: 10.1093/nar/gkw010.

Hurt JA, Robertson AD, Burge CB. 2013. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. Genome Research. 23:1636–1650. doi: 10.1101/gr.157354.113.

Imamachi N, Salam KA, Suzuki Y, Akimitsu N. 2017. A GC-rich sequence feature in the 3' UTR directs UPF1-dependent mRNA decay in mammalian cells. Genome Research. 407–418. doi: 10.1101/gr.206060.116.27.

Jacobs JL, Belew AT, Rakauskaite R, Dinman JD. 2007. Identification of functional, endogenous programmed -1 ribosomal frameshift signals in the genome of Saccharomyces cerevisiae. Nucleic Acids Research. 35:165–174. doi: 10.1093/nar/gkl1033.

Johnson JK, Waddell N, Chenevix-Trench G. 2012. The application of nonsense-mediated mRNA decay inhibition to the identification of breast cancer susceptibility genes. BMC Cancer. 12. doi: 10.1186/1471-2407-12-246.

Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, Dreyfuss G. 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature. 468:664–668. doi: 10.1038/nature09479.

Kapp LD, Lorsch JR. 2004. The molecular mechanics of eukaryotic translation. Ann Rev of Biochem. 73:657–704. doi: 10.1146/annurev.biochem.73.030403.080419.

Kataoka N, Diem MD, Kim VN, Yong J, Dreyfuss G. 2001. Magoh, a human homolog of Drosophila mago nashi protein, is a component of the splicing-dependent exon-exon junction complex. EMBO J. 20:6424–6433. doi: 10.1093/emboj/20.22.6424.

Kawashima T, Douglass S, Gabunilas J, Pellegrini M, Chanfreau GF. 2014. Widespread use of non-productive alternative splice sites in Saccharomyces cerevisiae. PLoS Genetics. 10. doi: 10.1371/journal.pgen.1004249.

Kearse MG et al. 2019. Ribosome queuing enables non-AUG translation to be resistant to multiple protein synthesis inhibitors. Genes & Development. doi: 10.1101/gad.324715.119.

Keeling KM et al. 2013. Attenuation of nonsense-mediated mRNA decay enhances in vivo nonsense suppression. PLoS ONE. 8. doi: 10.1371/journal.pone.0060478.

Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S. 2013. Function of alternative splicing. Gene. 514:1–30. doi: 10.1016/j.gene.2012.07.083.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler aD. 2002. The Human Genome Browser at UCSC. Genome Research. 12:996–1006. doi: 10.1101/gr.229102.

Klauer AA, Hoof A van. 2012. Degradation of mRNAs. RNA. 3:649–660. doi: 10.1002/wrna.1124.Degradation.

Klerk E de, 't Hoen PAC. 2015. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. Trends in Genetics. 31:128–139. doi: 10.1016/j.tig.2015.01.001.

Klessig DF. 1977. Two adenovirus mRNAs have a common 5' terminal leader sequence encoded at least 10 kb upstream from their main coding regions. Cell. 12:9–21. doi: 10.1016/0092-

8674(77)90181-7.

Kotlajich MV, Crabb TL, Hertel KJ. 2009. Spliceosome assembly pathways for different types of alternative splicing converge during commitment to splice site pairing in the A complex. Molecular and Cellular Biology. 29:1072–1082. doi: 10.1128/mcb.01071-08.

Kurihara Y et al. 2009. Genome-wide suppression of aberrant mRNA-like noncoding RNAs by NMD in Arabidopsis. PNAS. 106:2453–2458. doi: 10.1073/pnas.0808902106.

Kurosaki T, Popp MW, Maquat LE. 2019. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. Nature Rev MCB. 20:406–420. doi: 10.1038/s41580-019-0126-2.

Lander ESL et al. 2001. Initial sequencing and analysis of the human genome. Nature. 409:860–921.

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature. 446:926–929. doi: 10.1038/nature05676.

Leeds P, Peltz SW, Jacobson A, Culbertson MR. 1991. The product of the yeast UPF1 gene is required for rapid turnover of mRNAs containing a premature translational termination codon. Genes & Development. 5:2303–2314. doi: 10.1101/gad.5.12a.2303.

Leeds P, Wood JM, Lee BS, Culbertson MR. 1992. Gene products that promote mRNA turnover in Saccharomyces cerevisiae. Molecular and Cellular Biology. 12:2165–2177. doi: 10.1128/mcb.12.5.2165.

Legrain P, Seraphin B, Rosbash M. 1988. Early commitment of yeast pre-mRNA to the spliceosome pathway. Molecular and Cellular Biology. 8:3755–3760. doi: 10.1128/MCB.8.9.3755.

Le Hir H, Izaurralde E, Maquat LE, Moore MJ. 2000. The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. EMBO J. 19:6860–6869. doi: 10.1093/emboj/19.24.6860.

Le Hir H, Moore MJ, Maquat LE. 2000. Pre-mRNA splicing alters mRNP composition: Evidence for stable association of proteins at exon-exon junctions. Genes & Development. 14:1098–1108. doi: 10.1101/gad.14.9.1098.

Lewis BP, Green RE, Brenner SE. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. PNAS. 100:189–192. doi: 10.1073/pnas.0136770100.

Lim LP, Burge CB. 2001. A computational analysis of sequence features involved in recognition of short introns. PNAS. 98:11193–11198. doi: 10.1073/pnas.201407298.

Lim SR, Hertel KJ. 2004. Commitment to splice site pairing coincides with a complex formation. Molecular Cell. 15:477–483. doi: 10.1016/j.molcel.2004.06.025.

Lindeboom RGH, Supek F, Lehner B. 2016. The rules and impact of nonsense-mediated mRNA decay in human cancers. Nature Genetics. 48:1112–1118. doi: 10.1038/ng.3664.

Liu H, Rodgers ND, Jiao X, Kiledjian M. 2002. The scavenger mRNA decapping enzyme DcpS is a member of the HIT family of pyrophosphatases. EMBO J. 21:4699–4708. doi: 10.1093/emboj/cdf448.

Lobel JH, Tibble RW, Gross JD. 2019. Pat1 activates late steps in mRNA decay by multiple mechanisms. PNAS. 116:23512–23517. doi: 10.1073/pnas.1905455116.

Longman D, Hug N, Keith M, Anastasaki C, Patton EE, Grimes G, Cáceres JF. 2013. DHX34 and NBAS form part of an autoregulatory NMD circuit that regulates endogenous RNA targets in human cells, zebrafish and Caenorhabditis elegans. Nucleic Acids Research. 41:8319–8331. doi: 10.1093/nar/gkt585.

Losson R, Lacroute F. 1979. Interference of nonsense mutations with eukaryotic messenger RNA stability. PNAS. 76:5134–7.

Luo MJ, Reed R. 1999. Splicing is required for rapid and efficient mRNA export in metazoans. PNAS. 96:14937–14942. doi: 10.1073/pnas.96.26.14937.

Lykke-Andersen S, Chen Y, Ardal BR, Lilje B, Waage J, Sandelin A, Jensen TH. 2014. Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes. Genes & Development. 28:2498–2517. doi: 10.1101/gad.246538.114.

Ma J, Chen T, Mandelin J, Ceponis A, Miller NE, Hukkanen M, Ma GF, Konttinen YT. 2003. Regulation of macrophage activation. CMLS. 60:2334–2346. doi: 10.1007/s00018-003-3020-0.

Mabin JW, Woodward LA, Patton RD, Yi Z, Jia M, Wysocki VH, Bundschuh R, Singh

G. 2018. The exon junction complex undergoes a compositional switch that alters mRNP structure and nonsense-mediated mRNA decay activity. Cell Reports. 25:2431–2446.e7. doi: 10.1016/j.celrep.2018.11.046.

Maquat LE. 2004. Nonsense-mediated mRNA decay: Splicing, translation and mRNP dynamics. doi: 10.1038/nrm1310.

Maquat LE, Tarn WY, Isken O. 2010. The pioneer round of translation: Features and functions. Cell. 142:368–374. doi: 10.1016/j.cell.2010.07.022.

Martinez FJ et al. 2016. Protein-RNA networks regulated by normal and ALS-associated mutant hnRNPA2B1 in the nervous system. Neuron. 92:780–795. doi: 10.1016/j.neuron.2016.09.050.

Mendell JT, Sharifi NA, Meyers JL, Martinez-Murillo F, Dietz HC. 2004. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. Nature Genetics. 36:1073–1078. doi: 10.1038/ng1429.

Mercer TR et al. 2015. Genome-wide discovery of human splicing branchpoints. Genome Research. 25:290–303. doi: 10.1101/gr.182899.114.

Merz C, Urlaub H, Will CL, Lührmann R. 2007. Protein composition of human mRNPs spliced in vitro and differential requirements for mRNP protein recruitment. RNA. 13:116–128. doi: 10.1261/rna.336807.

Metkar M, Ozadam H, Lajoie BR, Imakaev M, Mirny LA, Dekker J, Moore MJ. 2018. Higher-order organization principles of pre-translational mRNPs. Molecular Cell. 72:715–726.e3. doi: 10.1016/j.molcel.2018.09.012.

Michaud S, Reed R. 1991. An ATP-independent complex commits pre-mRNA to the mammalian spliceosome assembly pathway. Genes & Development. 5:2534–2546. doi: 10.1101/gad.5.12b.2534.

Mili S, Steitz JA. 2004. Evidence for reassociation of RNA-binding proteins after cell lysis: Implications for the interpretation of immunoprecipitation analyses. RNA. 10:1692–1694. doi: 10.1261/rna.7151404.

Mitrovich QM, Anderson P. 2000. Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans. Genes & Development. 14:2173–2184. doi: 10.1101/gad.819900.veillance.

Moore MJ. 2002. Nuclear RNA turnover. Cell. 108:431–434. doi: 10.1016/S0092-8674(02)00645-1.

Moore MJ, Sharp PA. 1993. Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing. Nature. 365:364–268.

Moriarty PM, Reddy CC, Maquat LE. 1998. Selenium deficiency reduces the abundance of mRNA for Se-dependent glutathione peroxidase 1 by a UGA-dependent mechanism likely to be nonsense codon-mediated decay of cytoplasmic mRNA. Molecular and Cellular Biology. 18:2932–2939. doi: 10.1128/mcb.18.5.2932.

Morillon A, Gautheret D. 2019. Bridging the gap between reference and real transcriptomes. Genome Biology. 20:112. doi: 10.1186/s13059-019-1710-7.

Morrison M, Harris KS, Roth MB. 1997. smg mutants affect the expression of alternatively spliced SR protein mRNAs in Caenorhabditis elegans. PNAS. 94:9782–9785. doi: 10.1073/pnas.94.18.9782.

Muhlrad D, Decker CJ, Parker R. 1994. Deadenylation of the unstable mRNA encoded by the yeast MFA2 gene leads to decapping followed by 5' → 3' digestion of the transcript. Genes & Development. 8:855–866. doi: 10.1101/gad.8.7.855.

Muhlrad D, Decker CJ, Parker R. 1995. Turnover mechanisms of the stable yeast PGK1 mRNA. Molecular and Cellular Biology. 15:2145–2156. doi: 10.1128/mcb.15.4.2145.

Muhlrad D, Parker R. 1999. Aberrant mRNAs with extended 3' UTRs are substrates for rapid degradation by mRNA surveillance. RNA. 5:1299–1307.

Mühlemann O. 2008. Recognition of nonsense mRNA: towards a unified model. Biochem Soc Trans. 36:497–501. doi: 10.1042/BST0360497.

Mühlemann O, Eberle AB, Stalder L, Zamudio Orozco R. 2008. Recognition and elimination of nonsense mRNA. Biochim Biophys Acta. 1779:538–549. doi: 10.1016/j.bbagrm.2008.06.012.

Nasif S, Contu L, Mühlemann O. 2018. Beyond quality control: The role of nonsense-mediated mRNA decay (NMD) in regulating gene expression. Sem in Cell and Dev Biol. 75:78–87. doi: 10.1016/j.semcdb.2017.08.053.

Nellore A et al. 2016. Human splicing diversity and the extent of unannotated splice junctions

across human RNA-seq samples on the Sequence Read Archive. Genome Biology. 17:1–14. doi: 10.1186/s13059-016-1118-6.

Ni JZ et al. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes & Development. 21:708–718. doi: 10.1101/gad.1525507.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. Nature. 463:457–463. doi: 10.1038/nature08909.

Noensie EN, Dietz HC. 2001. A strategy for disease gene identification though nonsense-mediated mRNA decay inhibition. Nature Biotech. doi: 10.1038/88099.

Nomakuchi TT, Rigo F, Aznarez I, Krainer AR. 2016. Antisense oligonucleotide–directed inhibition of nonsense-mediated mRNA decay. Nature Biotech. 34:164–166. doi: 10.1038/nbt.3427.

Novoa I, Gallego J, Ferreira PG, Mendez R. 2010. Mitotic cell-cycle progression is regulated by CPEB1 and CPEB4-dependent translational control. Nature Cell Biology. 12:447–456. doi: 10.1038/ncb2046.

O'Leary NA et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 44:D733–D745. doi: 10.1093/nar/gkv1189.

Palacios IM, Gatfield D, St. Johnston D, Izaurralde E. 2004. An eIF4AIII-containing complex required for mRNA localization and nonsense-mediated mRNA decay. Nature. doi: 10.1038/nature02351.

Pan Q, Saltzman AL, Yoon KK, Misquitta C, Shai O, Maquat LE, Frey BJ, Blencowe BJ. 2006. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes & Development. 20:153–158. doi: 10.1101/gad.1382806.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics. 40:1413–1415. doi: 10.1038/ng.259.

Pandya-Jones A, Bhatt DM, Lin CH, Tong AJ, Smale ST, Black DL. 2013. Splicing kinetics and transcript release from the chromatin compartment limit the rate of Lipid A-induced gene

expression. RNA. 19:811–827. doi: 10.1261/rna.039081.113.

Passos DO, Doma MK, Shoemaker CJ, Muhlrad D, Green R, Weissman J, Hollien J, Parker R. 2009. Analysis of Dom34 and its function in no-go decay. Mol Biol of the Cell. 20:3025–3032. doi: 10.1091/mbc.E09 – 01– 0028.

Pastor F, Kolonias D, Giangrande PH, Gilboa E. 2010. Induction of tumour immunity by targeted inhibition of nonsense-mediated mRNA decay. Nature. 465:227–230. doi: 10.1038/nature08999.

Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser F, Chang Y-C, Madugundu A, Pandey A, Salzberg S. 2018. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biology. 19:208. doi: 10.1186/s13059-018-1590-2.

Pickrell JK, Pai AA, Gilad Y, Pritchard JK. 2010. Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genetics. 6:1–11. doi: 10.1371/journal.pgen.1001236.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research. 20:110–121. doi: 10.1101/gr.097857.109.

Presnyak V et al. 2015. Codon optimality is a major determinant of mRNA stability. Cell. 160:1111–1124. doi: 10.1016/j.cell.2015.02.029.

Proudfoot NJ. 2000. Connecting transcription to messenger RNA processing. Trends in Biochem Sciences. 25:290–293. doi: 10.1016/S0968-0004(00)01591-7.

Pruitt KD et al. 2014. RefSeq: An update on mammalian reference sequences. Nucleic Acids Research. 42:756–763. doi: 10.1093/nar/gkt1114.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research. 35:501–504. doi: 10.1093/nar/gkl842.

Pulak R, Anderson P. 1993. mRNA Surveillance by the Caenorhabditis elegans smg genes. Genes & Development. 7:1885–1897. doi: 10.1101/gad.7.10.1885.

Pyrkosz AB, Cheng H, Brown CT. 2013. RNA-Seq mapping errors when using incomplete reference transcriptomes of vertebrates. arXiv Preprint. 1–17. http://arxiv.org/abs/1303.2411.

Query CC, Moore MJ, Sharp PA. 1994. Branch nucleophile selection in pre-mRNA splicing: Evidence for the bulged duplex model. Genes & Development. 8:587–597. doi: 10.1101/gad.8.5.587.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842. doi: 10.1093/bioinformatics/btq033.

Radhakrishnan A, Chen YH, Martin S, Alhusaini N, Green R, Coller J. 2016. The DEAD-Box protein Dhh1p couples mRNA decay and translation by monitoring codon optimality. Cell. 167:122–132.e9. doi: 10.1016/j.cell.2016.08.053.

Rajput B, Pruitt KD, Murphy TD. 2019. RefSeq curation and annotation of stop codon recoding in vertebrates. Nucleic Acids Research. 47:594–606. doi: 10.1093/nar/gky1234.

Raney BJ et al. 2014. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics. 30:1003–1005. doi: 10.1093/bioinformatics/btt637.

Reichert VL, Hir HL, Jurica MS, Moore MJ. 2002. 5' exon interactions within the human spliceosome establish a framework for exon junction complex structure and assembly. Genes & Development. 16:2778–2791. doi: 10.1101/gad.1030602.

Ricci EP, Kucukural A, Cenik C, Mercier BC, Singh G, Heyer EE, Ashar-Patel A, Peng L, Moore MJ. 2014. Staufen1 senses overall transcript secondary structure to regulate translation. NSMB. 21:26–35. doi: 10.1038/nsmb.2739.

Rodrigues R, Grosso AR, Moita L. 2013. Genome-wide analysis of alternative splicing during dendritic cell response to a bacterial challenge Chamaillard, M, editor. PLoS ONE. 8:e61975. doi: 10.1371/journal.pone.0061975.

Rosenfeld MG, Lin CR, Amara SG, Stolarsky L, Roos BA, Ong ES, Evans RM. 1982. Calcitonin mRNA polymorphism: Peptide switching associated with alternative RNA splicing events. PNAS. 79:1717–1721. doi: 10.1073/pnas.79.6.1717.

Rosenfeld MG, Mermod J-J, Amara SG, Swanson LW, Sawchenko PE, Rivier J, Vale WW, Evans RM. 1983. Production of a novel neuropeptide encoded by the calcitonin gene via tissue-specific RNA processing. Nature. 304:129–135. doi: 10.1038/304129a0.

Sachs AB, Davis RW, Kornberg RD. 1987. A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability. Molecular and Cellular Biology.

7:3268–3276. doi: 10.1128/mcb.7.9.3268.

Saito S, Hosoda N, Hoshino SI. 2013. The Hbs1-Dom34 protein complex functions in non-stop mRNA decay in mammalian cells. J of Comp Biol. 288:17832–17843. doi: 10.1074/jbc.M112.448977.

Santos DA, Shi L, Tu BP, Weissman JS. 2019. Cycloheximide can distort measurements of mRNA levels and translation efficiency. Nucleic Acids Research. 47:4974–4985. doi: 10.1093/nar/gkz205.

Saudemont B, Popa A, Parmley JL, Rocher V, Blugeon C, Necsulea A, Meyer E, Duret L. 2017. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. Genome Biology. 18:1–15. doi: 10.1186/s13059-017-1344-6.

Saulière J et al. 2012. CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. NSMB. 19:1124–1131. doi: 10.1038/nsmb.2420.

Schneider E, Blundell M, Kennell D. 1978. Translation and mRNA decay. MGG Molecular & General Genetics. 160:121–129. doi: 10.1007/BF00267473.

Schwartz DC, Parker R. 2000. mRNA decapping in yeast requires dissociation of the cap binding protein, eukaryotic translation initiation factor 4E. Molecular and Cellular Biology. 20:7933–7942. doi: 10.1128/mcb.20.21.7933-7942.2000.

Schwartz DC, Parker R. 1999. Mutations in translation initiation factors lead to increased rates of deadenylation and decapping of mRNAs in Saccharomyces cerevisiae. Molecular and Cellular Biology. 19:5247–5256. doi: 10.1128/MCB.19.8.5247.

Seraphin B, Rosbash M. 1989. Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. Cell. 59:349–358. doi: 10.1016/0092-8674(89)90296-1.

Shcherbakova I, Hoskins AA, Friedman LJ, Serebrov V, Corrêa IR, Xu MQ, Gelles J, Moore MJ. 2013. Alternative spliceosome assembly pathways revealed by single-molecule fluorescence microscopy. Cell Reports. 5:151–165. doi: 10.1016/j.celrep.2013.08.026.

Sheth U, Parker R. 2003. Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. Science. 300:805–808. doi: 10.1126/science.1082320.

Shibuya T, Tange T, Sonenberg N, Moore MJ. 2004. eIF4AIII binds spliced mRNA in the exon junction complex and is essential for nonsense-mediated decay. NSMB. 11:346–351. doi: 10.1038/nsmb750.

Shoemaker CJ, Eyler DE, Green R. 2010. Dom34:Hbs1 promotes subunit dissociation and peptidyl-tRNA drop-off to initiate no-go decay. Science. 330:369–372. doi: 10.1126/science.1192430.

Sidhu JS, Omiecinski CJ. 1998. Protein synthesis inhibitors exhibit a nonspecific effect on phenobarbital-inducible cytochome P450 gene expression in primary rat hepatocytes. J of Comp Biol. 273:4769–4775. doi: 10.1074/jbc.273.8.4769.

Simms CL, Yan LL, Zaher HS. 2017. Ribosome collision is critical for quality control during no-go decay. Molecular Cell. 68:361–373.e5. doi: 10.1016/j.molcel.2017.08.019.

Singh G, Kucukural A, Cenik C, Leszyk JD, Shaffer SA, Weng Z, Moore MJ. 2012. The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. Cell. 151:750–764. doi: 10.1016/j.cell.2012.10.007.

Singh G, Ricci EP, Moore MJ. 2014. RIPiT-Seq: A high-throughput approach for footprinting RNA: Protein complexes. Methods. 65:320–332. doi: 10.1016/j.ymeth.2013.09.013.

Smith JE, Alvarez-Dominguez JR, Kline N, Huynh NJ, Geisler S, Hu W, Coller J, Baker KE. 2014. Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. Cell Reports. 7:1858–1866. doi: 10.1016/j.celrep.2014.05.023.

Solnick D. 1985. Alternative splicing caused by RNA secondary structure. Cell. 43:667–676. doi: 10.1016/0092-8674(85)90239-9.

Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H. 2005. Function of alternative splicing. Gene. 344:1–20. doi: 10.1016/j.gene.2004.10.022.

Sultan M, Amstislavskiy V, Risch T, Schuette M, Dökel S, Ralser M, Balzereit D, Lehrach H, Yaspo ML. 2014. Influence of RNA extraction methods and library selection schemes on RNA-seq data. BMC Genomics. 15. doi: 10.1186/1471-2164-15-675.

Sun X, Li X, Moriarty PM, Henics T, LaDuca JP, Maquat LE. 2001. Nonsense-mediated decay of mRNA for the selenoprotein phospholipid hydroperoxide glutathione peroxidase is detectable in cultured cells but masked or inhibited in rat tissues. Mol Biol of the Cell. 12:1009–1017. doi:

10.1091/mbc.12.4.1009.

Sun X, Moriarty PM, Maquat LE. 2000. Nonsense-mediated decay of glutathione peroxidase 1 mRNA in the cytoplasm depends on intron position. EMBO J. 19:4734–4744. doi: 10.1093/emboj/19.17.4734.

Suzuki H, Kameyama T, Ohe K, Tsukahara T, Mayeda A. 2013. Nested introns in an intron: Evidence of multi-step splicing in a large intron of the human dystrophin pre-mRNA. FEBS Letters. 587:555–561. doi: 10.1016/j.febslet.2013.01.057.

Sveen A, Kilpinen S, Ruusulehto A, Lothe RA, Skotheim RI. 2016. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene. 35:2413–2427. doi: 10.1038/onc.2015.318.

Sweet T, Kovalak C, Coller J. 2012. The DEAD-box protein Dhh1 promotes decapping by slowing ribosome movement Maquat, L, editor. PLoS Biology. 10:e1001342. doi: 10.1371/journal.pbio.1001342.

Tabiti K, Smith DR, Goh H-S, Pallen CJ. 1995. Increased mRNA expression of the receptor-like protein tyrosine phosphatase α in late stage colon carcinomas. Cancer Letters. 93:239–248. doi: 10.1016/0304-3835(95)03816-F.

Taggart AJ, Desimone AM, Shih JS, Filloux ME, Fairbrother WG. 2012. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. NSMB. 19:719–721. doi: 10.1038/nsmb.2327.

Tarun SZ, Sachs AB. 1996. Association of the yeast poly(A) tail binding protein with translation initiation factor eIF-4G. EMBO J. 15:7168–7177. doi: 10.1002/j.1460-2075.1996.tb01108.x.

Tenenbaum SA, Carson CC, Lager PJ, Keene JD. 2000. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. PNAS. 97:14085–14090. doi: 10.1073/pnas.97.26.14085.

Tharun S, Parker R. 2001. Targeting an mRNA for decapping: Displacement of translation factors and association of the Lsm1p-7p complex on deadenylated yeast mRNAs. Molecular Cell. 8:1075–1083. doi: 10.1016/S1097-2765(01)00395-1.

The GTEx Consortium et al. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 348:648–660. doi: 10.1126/science.1262110.

Treisman R, Orkin SH, Maniatis T. 1983. Specific transcription and RNA splicing defects in five cloned β-thalassaemia genes. Nature. 302:591–596. doi: 10.1038/302591a0.

Tucker M, Staples RR, Valencia-Sanchez MA, Muhlrad D, Parker R. 2002. Ccr4p is the catalytic subunit of a Ccr4p/Pop2p/Notp mRNA deadenylase complex in Saccharomyces cerevisiae. EMBO J. 21:1427–1436. doi: 10.1093/emboj/21.6.1427.

Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. 2003. CLIP identifies Nova-regutated RNA networks in the brain. Advancement Of Science. 302:1212–1215.

Van Hoof A, Frischmeyer PA, Dietz HC. 2002. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. Science. 295:2262–2264. doi: 10.1126/science.1067272.

Wachtel C, Li B, Sperling J, Sperling R. 2004. Stop codon-mediated suppression of splicing is a novel nuclear scanning mechanism not affected by elements of protein synthesis and NMD. RNA. 10:1740–1750. doi: 10.1261/rna.7480804.

Wahl MC, Will CL, Lührmann R. 2009. The spliceosome: Design principles of a dynamic RNP machine. Cell. 136:701–718. doi: 10.1016/j.cell.2009.02.009.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. Nature. 456:470–6. doi: 10.1038/nature07509.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. Nature. 456:470–476. doi: 10.1038/nature07509.

Wang VG, Kim H, Chuang JH. 2018. Whole-exome sequencing capture kit biases yield false negative mutation calls in TCGA cohorts. PLoS ONE. 13:1–14. doi: 10.1371/journal.pone.0204912.

Webster MW, Chen YH, Stowell JAW, Alhusaini N, Sweet T, Graveley BR, Coller J, Passmore LA. 2018. mRNA deadenylation is coupled to translation rates by the differential activities of Ccr4-Not nucleases. Molecular Cell. 70:1089–1100.e8. doi: 10.1016/j.molcel.2018.05.033.

Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics. 45:1113–1120. doi: 10.1038/ng.2764.

Wells CA et al. 2006. Alternate transcription of the Toll-like receptor signaling cascade. Genome Biology. 7. doi: 10.1186/gb-2006-7-2-r10.

Wengrod J, Martin L, Wang D, Frischmeyer-Guerrerio P, Dietz HC, Gardner LB. 2013. Inhibition of nonsense-mediated RNA decay activates autophagy. Molecular and Cellular Biology. 33:2128–2135. doi: 10.1128/mcb.00174-13.

Wiederhold K, Passmore LA. 2010. Cytoplasmic deadenylation: Regulation of mRNA fate. Biochemical Society Transactions. 38:1531–1536. doi: 10.1042/BST0381531.

Will CL, Luhrmann R. 2011. Spliceosome structure and function. CSH Perspectives in Biol. 3:a003707–a003707. doi: 10.1101/cshperspect.a003707.

Wittmann J, Hol EM, Jack H-M. 2006. hUPF2 silencing identifies physiologic substrates of mammalian nonsense-mediated mRNA decay. Molecular and Cellular Biology. 26:1272–1287. doi: 10.1128/mcb.26.4.1272-1287.2006.

Wood RD, Doublié S. 2016. DNA polymerase θ (POLQ), double-strand break repair, and cancer. DNA Repair. 44:22–32. doi: 10.1016/j.dnarep.2016.05.003.

Woodward LA, Mabin JW, Gangras P, Singh G. 2017. The exon junction complex: a lifelong guardian of mRNA fate. WIREs RNA. 8. doi: 10.1002/wrna.1411.

Wu P-Y, Phan JH, Wang MD. 2013. Assessing the impact of human genome annotation choice on RNA-seq expression estimates. BMC Bioinformatics. 14:S8. doi: 10.1186/1471-2105-14-S11-S8.

Yan Q, Weyn-Vanhentenryck SM, Wu J, Sloan SA, Zhang Y, Chen K, Wu JQ, Barres BA, Zhang C. 2015. Systematic discovery of regulated and conserved alternative exons in the mammalian brain reveals NMD modulating chromatin regulators. PNAS. 112:3445–3450. doi: 10.1073/pnas.1502849112.

Yang YW, Flynn RA, Chen Y, Qu K, Wan B, Wang KC, Lei M, Chang HY. 2014. Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. eLife. 3:1–19. doi: 10.7554/elife.02046.

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J of Comp Biol. 11:377–394. doi: 10.1089/1066527041410418.

Yeo GW, Van Nostrand E, Holste D, Poggio T, Burge CB. 2005. Identification and analysis of alternative splicing events conserved in human and mouse. PNAS. 102:2850–2855. doi: 10.1073/pnas.0409742102.

Yepiskoposyan H, Aeschimann F, Nilsson D, Okoniewski M, Mu O. 2011. Autoregulation of the nonsense-mediated mRNA decay pathway in human cells. RNA. 2108–2118. doi: 10.1261/rna.030247.111.strate.

Zhang X, Yan C, Hang J, Finci LI, Lei J, Shi Y. 2017. An atomic structure of the human spliceosome. Cell. 169:918–929.e14. doi: 10.1016/j.cell.2017.04.033.

Zhang X, Zhan X, Yan C, Zhang W, Liu D, Lei J, Shi Y. 2019. Structures of the human spliceosomes before and after release of the ligated exon. Cell Research. 29:274–285. doi: 10.1038/s41422-019-0143-x.

Zhao J et al. 2010. Genome-wide identification of Polycomb-associated RNAs by RIP-seq. Molecular Cell. 40:939–953. doi: 10.1016/j.molcel.2010.12.011.

Zhao S. 2014. Assessment of the impact of using a reference transcriptome in mapping short RNA-Seq reads. PLoS ONE. 9. doi: 10.1371/journal.pone.0101374.

Zhao S, Zhang B. 2015. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics. 16:1–14. doi: 10.1186/s12864-015-1308-8.

Zheng S, Gray EE, Chawla G, Porse BT, O'Dell TJ, Black DL. 2012. PSD-95 is post-transcriptionally repressed during early neural development by PTBP1 and PTBP2. Nature Neurosci. 15:381–388. doi: 10.1038/nn.3026.