Cryptic splice sites are agents of order and chaos

during pre-mRNA splicing

by

Brian Joseph

A Dissertation

Presented to the Faculty of the Louis V. Gerstner, Jr

Graduate School of Biomedical Sciences,

Memorial Sloan Kettering Cancer Center

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

New York, NY

October, 2020

_____ _____

Eric C. Lai, PhD Date

Dissertation Mentor

For my parents

William and Victoria

ABSTRACT

High throughput sequencing and informatic pipelines have greatly impacted study of the genome and the transcriptome. These approaches have provided unparalleled views of RNA processing and expanded the annotation of conserved, non-canonical mechanisms that promise new discoveries in the field of RNA metabolism.

Cryptic splice sites (CSS) are the archetypal non-canonical element. The eukaryotic pre-mRNA landscape contains thousands of CSS that match consensus splice motifs, but do not show direct evidence of activation on mature RNAs. Several lines of observation suggest these can facilitate and antagonize fruitful RNA processing, but their functions remain poorly explored. In this work, I apply a combination of molecular biology, genetics and bioinformatics to explore CSS in the fruit fly. I define and characterize the cryptic splicing landscape using multiple genomic strategies, analyze the mechanism of CSS activity during pre-mRNA maturation and examine *in vivo* requirements during host gene expression.

In the first two sections, I study intron removal within the context of recursive splicing (RS). RS has been proposed as a strategy to process unusually long introns (average length ~ 50000 nt) as multiple smaller fragments. This phenomenon requires distinctive cryptic intronic substrates composed of directly adjacent splice acceptor (SA) and splice donor (SD) sequences, referred to as ratchet points (RPs). RPs represent an interesting class since they are deeply conserved, but clearly suboptimal splice substrates. In part one, I investigate how intronic RPs are recognized by the spliceosome. Partial mutagenesis of RPs *in vivo* causes characteristic loss-of-function phenotypes through defects in host gene RNA processing. Disrupting RP SD do not abolish recursive splicing, but instead activate cryptic exons (RS-exons) consisting of sequence immediately downstream of the RP. I show that RS-exons are required for RP definition and generalize my findings by discovering conserved, cryptic SD shortly downstream of RPs, transcriptome-wide. I propose a two-step intronic recursive splicing model: First, activation of RP SA occurs through cryptic RS-exon definition, resulting in removal of the upstream intronic segment. Next, the regenerated SD outcompetes the cryptic RS-exon SD to remove the remaining intronic sequence.

In part two, I conduct molecular and genetic analyses to explore the biology of RPs. Experimental dissection using a spectrum of RS-exon sequences indicates that splice site strength and exonic splicing enhancer sequences can control RS-exon alternative splicing. Conversely, minigene manipulation suggests that the Exon Junction

Complex (EJC) may suppress recursive splicing and promote RS-exon inclusion. Finally, I delete nine RPs in the animal to examine *in vivo* requirements of recursive splicing. Unlike partial disruptions, which yield processing defects and loss-of-function phenotypes, intronic RP deletion mutants are overtly normal and produce correctly spliced mRNA. Curiously, deletion of the *Ubx* recursive microexons yield RNA processing defects, as well as animal phenotypes.

In the final section, I explore a class of poisonous CSS that is suppressed by the EJC. Using publicly available EJC loss-of-function RNA sequencing datasets, I demonstrate that the *Drosophila* EJC suppresses hundreds of functional CSS, even though the most of these bear weak splicing motifs and might appear incompetent. Characterization indicates a majority of these sites map to exons and are concentrated within exon junction sequences. Consequently, their activation results in spurious loss of mRNA sequences. Mechanistically, I show that the EJC directly conceals critical splicing elements by virtue of its position-specific recruitment, preventing cryptic SS definition. Unexpectedly, I also discover the EJC inhibits scores of regenerated 5' and 3' recursive splice sites on segments that have already undergone splicing, and that loss of EJC regulation triggers faulty resplicing of mRNA. An important corollary is that certain intronless cDNA expression constructs yield high levels of unanticipated, truncated transcripts generated by resplicing. These findings highlight an ancestral function of the EJC and emphasize conserved roles to defend transcriptome fidelity.

Altogether my work advances our conception of intron removal and underscores the function of cryptic splice sites in this process. The discovery of cryptic RS-exons underscores the paradigm of splice site regulation after exon definition. Hence, I propose a general role for CSS in facilitating definition of suboptimal exons. Conversely, CSS can also cause undesirable RNA processing and require special mechanisms to silence their activity.

BIOGRAPHICAL SKETCH

Brian Joseph was born in Karachi, Pakistan. Following graduation from St. Patrick's High School (Karachi) in 2009, he moved to the United States of America to attend college at the Massachusetts Institute of Technology (Cambridge, MA). There, he majored in Biological Engineering (Course 20) and engaged in scientific research in the lab of Dr. Robert Langer through the Undergraduate Research Opportunities Program (UROP). Following college graduation in 2013, Brian moved to the Upper East Side of New York City, to begin the first year of his PhD at the Gerstner Sloan Kettering Graduate School of Biomedical Sciences. In 2014, he joined the laboratory of Dr. Eric Lai at the Sloan-Kettering Institute to undertake his dissertation research.

ACKNOWLEDGEMENTS

This dissertation is written using a first-person, singular voice. But even in the most isolating circumstances, experiments, data, results – this document – represents the efforts of several individuals, institutional support and endless good fortune.

I thank my advisor, Dr. Eric Lai, for his pivotal role in my training and professional accomplishments. For continuously pushing me to translate my intellectual curiosities into research pursuits and for nurturing a sense of fearlessness in my scientific endeavors. I am grateful for his tireless mentorship during the beginning of my graduate work, when I struggled with planning, executing, analyzing and communicating. And for his virtuous patience during this period as I amassed essential skillsets through continuous failure. More than anything else, I am deeply grateful for the immense independence and encouragement he has given me over the years.

I also extend sincere gratitude to members of the Lai lab, past and present who have been dear friends, scientific mentors and remarkable colleagues. I thank Piero Sanfilippo for his mentorship and for many productive conversations on RNA biology and post-transcriptional regulation. Similarly, I thank Daniel Garaulet, Luis De Navas and Binglong Zhang for their guidance and advice with fly pushing, genetics and IHC. And Jeffrey Vedanayagam and Sol Shenker for their support with bioinformatics and computation. Importantly, I thank our lab technicians, managers and administrators who made my job infinitely easier. Especially Nan Pang, Sayani Sen and Chaz Scala who assisted with numerous experiments.

I also want to acknowledge and thank Lijuan Kan and Eun Sil Park, with whom I had a long and productive collaboration on the m^6A pathway. It has been a joy to work in their company.

I am deeply indebted to Dr. Shu Kondo and his amazing staff of technicians. They engineered the first recursive splice donor mutants – these eventful reagents completely shifted my research focus and have led to this entire body of work.

I thank members of my thesis committee, Drs. Stewart Shuman and Mary Baylies for being wonderful advocates. I have benefited a lot from their useful feedback and regular encouragement. I thank Dr. Dirk Remus, who graciously offered to chair my defense, and Dr. Jernej Ule, who agreed to serve as external examiner. I also thank members of the faculty at MSK and beyond who have mentored me formally and informally, throughout my life. Special mentions to Dr. Chris Lima, who was my first-year counsellor and Dr. Daniel Heller, with whom I began my scientific career.

I am truly grateful to the graduate program for gambling on me, and for supporting me continuously. I thank the former and current deans, Drs. Ken Marians and Mike Overholtzer for their advocacy and support. I also thank associate dean Linda Burnley for her resourcefulness and help with all matters related to my personal and professional life. Moreover, I thank past and present members of the office who have always prioritized student needs. Beyond those mentioned, I extend thanks to all members of our community.

On a personal note, I wish to thank my friends who have greatly improved my life. I thank my high school, college and grad school friends, who have always been willing to lend me their ears and offer pearls of wisdom. I also thank my chess and running friends, who have provided an outlet for physical and mental release. I am especially indebted to Erman Karasu, my roommate, my first friend in NYC and my cheerleader.

I thank my family for their unparalleled love, support and guidance. Starting with my parents, William and Victoria, who have imparted to me their suicidal work ethic and honesty. I thank my siblings Chrysann, Collin and Edward, and my brother-in-law

Jonathan for being a source of joy and pride. Despite being separated I have never felt alone, and I am grateful to my family for being there for me, always.

Finally, to my partner Ruth Isserman: thank you, I love you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SS – Splice site

RP – Ratchet point

RSS – Recursive splice site

BP – Branch point

SRE – Splicing regulatory elements

RNAPII – RNA Polymerase II

ESE – Exonic splicing enhancer

ESS – Exonic splicing suppressor

EJC – Exon Junction Complex

hnRNA – heterogeneous nuclear RNA

Ad2 – human adenovirus 2

EM – electron microscopy

RNP – ribonucleoproteins

snRNA – small nuclear RNA

snRNP – small nuclear RNP

Sm – Smith

NTC – nineteen complex

NTR – NTC related

ISL – internal-stem-loop

ILS – intron lariat spliceosome

ASAP/PSAP – ACIN1/PNN-RNPS1-SAP1 complex

AS – alternative splicing

RBP – RNA binding protein

rt-PCR – reverse transcription-polymerase chain reaction

CNS – central nervous system

kb – kilo base or kilo base pair

cDNA – complementary DNA

Chapter 1

Introduction

In eukaryotes, genes can occur in noncontiguous pieces, thus transcription must be succeeded by a specialized processing step to join together messenger sequences and remove intervening RNA, a process called splicing. Sequences that are removed during this stage are called introns, whereas those preserved, exons. A wealth of genetic, biochemical, molecular and informatic studies over the last 4 decades has revealed fundamental principles of intron removal. Furthermore, these studies have provided a keen sense of how cells regulate intron removal to diversify mRNA and protein output, a mechanism called alternative splicing. That splicing and regulation of splicing is essential to eukaryotes is abundantly evident by the deep conservation of the splicing machinery, as well as human disorders and diseases that arise due to disruptions of these processes. Moreover, our deep appreciation of the splicing pathway has proven a useful resource for therapeutic intervention and biotechnological innovation. Nevertheless, there are still fundamental, unexplored curiosities regarding the splicing pathway. Cryptic splice sites, sequences that match consensus splice motifs, can be found throughout the transcriptome but their exact functions during pre-mRNA splicing remains mysterious. These sites are assumed to silent because their selection cannot be inferred by assessing mRNA sequences. However, it is also possible to generate mRNA via cryptic splice site activation. The peculiar phenomenon of recursive splicing is one such example and could provide novel insight into the process of intron removal. In recursive splicing (RS), a single intron is proposed to be removed as two or more smaller fragments. This is clearly a deviation from canonical splicing, which removes introns as single fragments. RS has only been observed in unusually long introns, so while instances of this process have been recognized since

the early aughts, and in both vertebrate and invertebrate organisms, mechanistic and functional dissection has remained a challenge due to the technical limitations of manipulating long introns. My dissertation research explores the landscape, mechanism and function of recursive splicing, and through these efforts, I discover that cryptic splice sites can have useful and harmful impacts on fruitful gene expression. Due to the multifaceted nature of my interests, I used the genetically tractable fruit fly model organism to examine this process and aimed to study RS in the context of animal development. In this introduction, I first provide a historical overview of the discovery of splicing, followed by a review of the splicing reaction and pathway. I also introduce concepts related to the regulation of splicing, and elaborate on noncanonical modes of splicing, including cryptic splice sites and recursive splicing. Lastly, I summarize my thesis objectives and main findings.

The discovery of RNA splicing and genes in pieces

The relationship between hnRNA and mRNA

The pioneering discovery of mRNA as an unstable intermediate carrying information from genes to ribosomes in bacteria (BRENNER et al., 1961; Cobb, 2015; GROS et al., 1961) led to a search for the same in eukaryotes. While discovery of eukaryotic mRNA soon followed, pulse-chase radiolabeling of newly synthesized RNA in animal cells revealed that a large fraction of nuclear RNA was rapidly degraded after synthesis, leaving a small fraction that was exported to the cytoplasm (**Figure 1.1**) (HARRIS, 1959; HARRIS & WATTS, 1962; SCHERRER et al., 1963). This somewhat confusing observation was difficult to resolve using contemporaneous tools, but it was clear that the short-lived RNA species was much longer in length, up to tens of kilobases

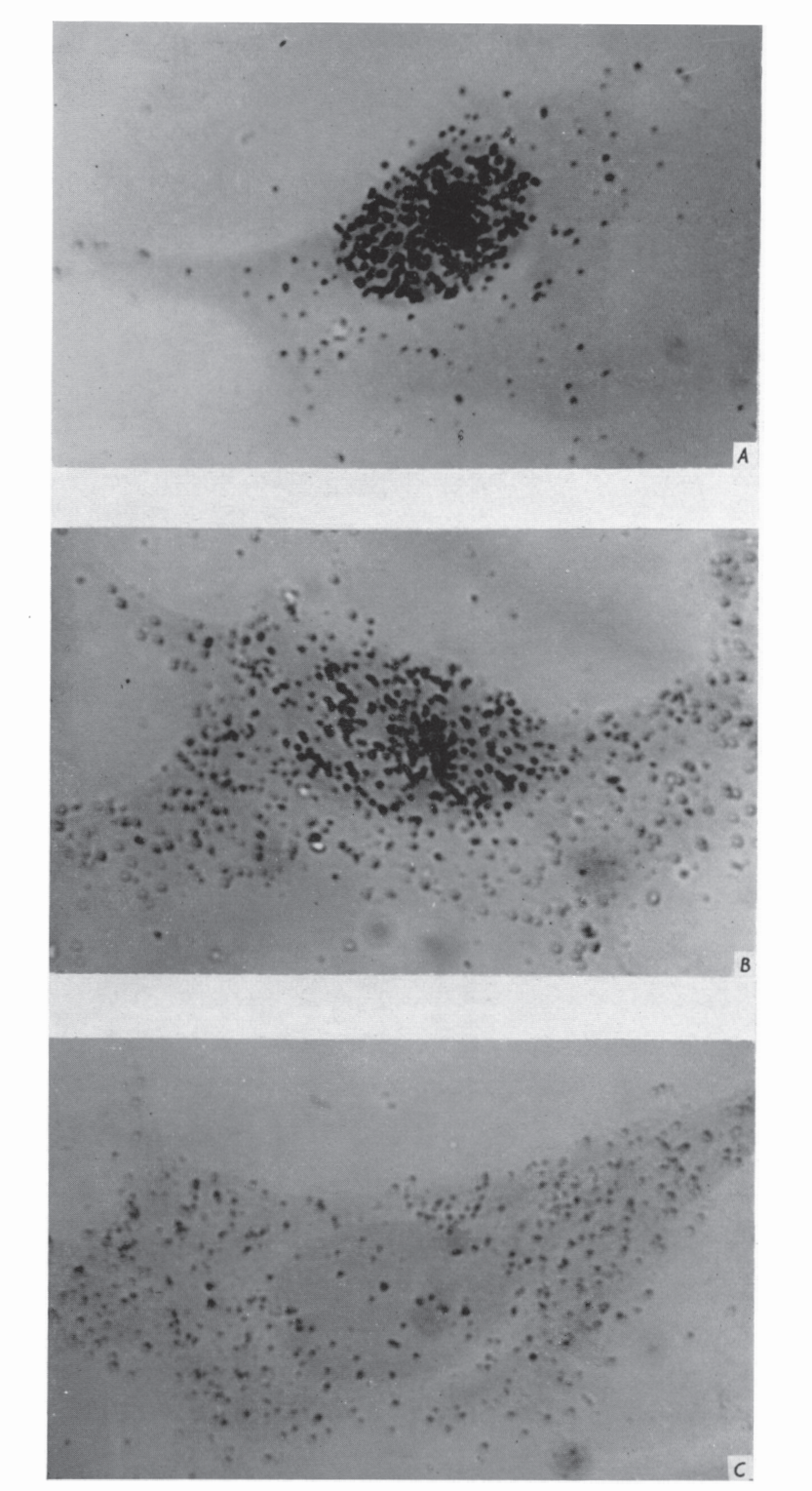

A
B
C

Figure 1.1 Evidence for a short-lived nuclear RNA species. A. Radioautograph of a connective-tissue cell fixed after incubation for 20 min. in medium containing tritium-labelled adenosine. The cell was not in the phase of DNA synthesis. The nucleus is more heavily labelled than the cytoplasm and the heaviest labelling is over the nucleolus. B. Radioautograph of a connective-tissue cell fixed after incubation for 20 min. in medium containing tritium-labelled adenosine and 3 hr. in non-radioactive medium containing 2 mM-adenosine and 2 mM-guanosine. The cell was not in the phase of DNA synthesis at the time of exposure to the tritium-labelled precursor. The nucleus is less heavily labelled than the nucleus in A, but the cytoplasm is much more heavily labelled. The nucleolus is still the most heavily labelled part of the cell. C. Radioautograph of a connective-tissue cell fixed after incubation for 20 min. in medium containing tritium-labelled adenosine and 12 hr. in non-radioactive medium containing 2 mM-adenosine and 2 mM-guanosine. The nucleus is now much less heavily labelled than in A and B, but the cytoplasm shows about the same degree of labelling as in B. Used with permission (HARRIS, 1959)

– significantly longer than cytoplasmic mRNAs. Consequently, the short-lived species was referred to as heterogenous nuclear RNA (hnRNA) (Soeiro et al., 1966, 1968). Two important findings suggested that hnRNA function as a precursor to mRNA. First, at the 5' end, both hnRNA and mRNA were shown to have a 5' cap structure (Rottman et al., 1974). Second, at the 3' end, both hnRNA and mRNA were also shown to be polyadenylated (Darnell et al., 1971). However, that only 5-10% of synthesized RNA was exported to the cytoplasm challenged the view that hnRNA was precursor to mRNA. Therefore, the exact nature of hnRNA and its relationship to mRNA remained a major conundrum in molecular biology.

Mapping of the abundantly expressed split Adenovirus *hexon* mRNA

Prior to the development of molecular cloning techniques, mRNA metabolism was typically explored using DNA viruses to infect cells. This was, in part, because studying viral mRNAs was a more tractable system than endogenous animal genes. 1. viruses had a much smaller number of genes compared to animal cells, 2. viral DNA could be isolated from virion particles and 3. Large scale infection resulted in strong viral mRNA synthesis that could be coupled with downstream experimentation. The human adenovirus 2 (Ad2) was one choice of model systems to study mRNA synthesis. Importantly, it was shown to produce the short-lived hnRNA species in cells, making it an attractive system to examine mRNA synthesis in animals (Berk, 2016; Wall et al., 1972).

Thus, the discovery of splicing is intertwined with efforts to map the genomic location encoding the abundantly expressed Ad2 *hexon* mRNA, a precursor to the structural protein hexon. The technology used to define genomic location was a newly discovered method that stabilized R-loop formation *in vitro,* and visualized RNA-DNA hybrids using electron microscopy (EM) (**Figure 1.2**). Since formation of RNA-DNA

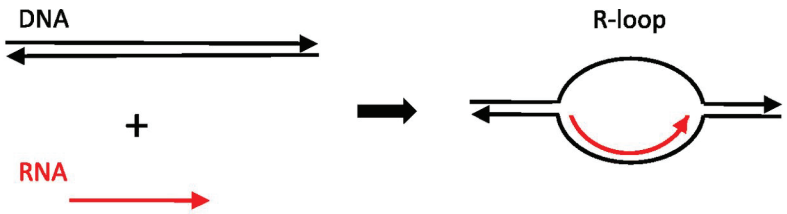


Figure 1.2. R-loop formation as a method to map mRNA on DNA sequences
Used with permission (Berk, 2016)

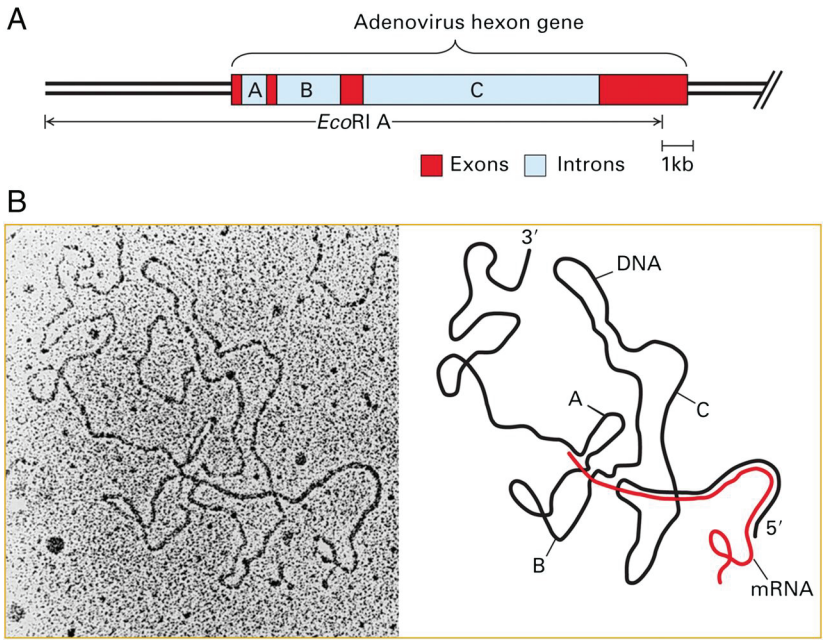


Figure 1.3. Discovery of the noncontinuous hexon mRNA.
EM of hybrid between purified hexon mRNA and the transcribed strand of the Ad2 EcoRI A DNA fragment. (A) Diagram of the positions of hexon mRNA exons (red) and the introns between them (A, B, and C in light blue) in the left ~25 kb of the Ad2 genome. (B) EM of hybrid between hexon mRNA and the EcoRI A coding strand. In the interpretation shown on the right, the mRNA is shown in red and DNA in black. Regions where the red RNA is parallel to the black DNA strand represent base paired regions of RNA-DNA hybrid. Used with permission (Berk, 2016 and Berget et al., 1977).

hybrid results in branching out of single stranded DNA, EM imaging of R-loops could in theory indicate the region of mRNA-DNA complementarity, which was inferred as genomic coordinates. Importantly, this technology had been employed to map the location of *Drosophila* rDNA (White & Hogness, 1977). Using this method, the Sharp and Roberts labs discovered that the *hexon* mRNA formed three distinct R-loops when incubated with Ad2 DNA. These findings clearly indicated that three noncontiguous regions of the Ad2 genome were joined into the *hexon* mRNA. Thus, the concept of a split gene materialized. And the knowledge of a longer hnRNA that spanned the entire locus led to the proposal of RNA processing as a means to remove intervening sequences and splice together what was subsequently called exons to produce the *hexon* mRNA (**Figure 1.3**) (Berget et al., 1977; Chow et al., 1977).

Split genes are a common feature of eukaryotic genes

Shortly after the discovery of the split *hexon* mRNA, split genes were also identified in endogenous eukaryotic genes, such as mouse *β-globin* and chicken *ovalbumin* (Lai et al., 1978; Tilghman et al., 1978). Subsequently, through the invention and deployment of molecular cloning, high throughput sequencing methods and informatic pipelines, it is now well appreciated that split genes are a deeply conserved and common feature of eukaryotic genes.

Discovery of the splicing machinery

Only three years following the discovery of the split *hexon* gene, in a landmark article titled “Are snRNPs involved in splicing?” the Steitz lab proposed that the abundant snRNAs U1, U2, U4, U5 and U6, found within ribonucleoproteins (RNPs) are involved in

RNA processing. This hypothesis was built on three clues. First, that the machinery must be conserved across eukaryotes. Second, involvement in RNA biogenesis predicts that expression of these snRNAs must be most abundant in metabolically active cells. And lastly, the 5' end of the U1 snRNA showed high complementarity to known conserved splice junction sequences. Supporting these clues, in the same article it was demonstrated that the U1, U2, U4, U5 and U6 snRNAs interact with hnRNA, but a degraded form of U1 lacking the 5' complementary sequences no longer sediments with hnRNA (Lerner et al., 1980). Development of an *in vitro* system for splicing studies proved invaluable to test this hypothesis and elucidate the mechanism of splicing (Hernandez & Keller, 1983; Padgett, Hardy, et al., 1983). With the generation of monoclonal antibodies that inhibited the snRNPs, it was possible to functionally demonstrate their requirement during splicing (Padgett, Mount, et al., 1983). Shortly thereafter, it was determined that snRNPs formed the spliceosome that executed intron removal (Padgett et al., 1986). Mass spectrometry has been employed on purified complexes to detect other members of the spliceosome (Rappsilber et al., 2002).

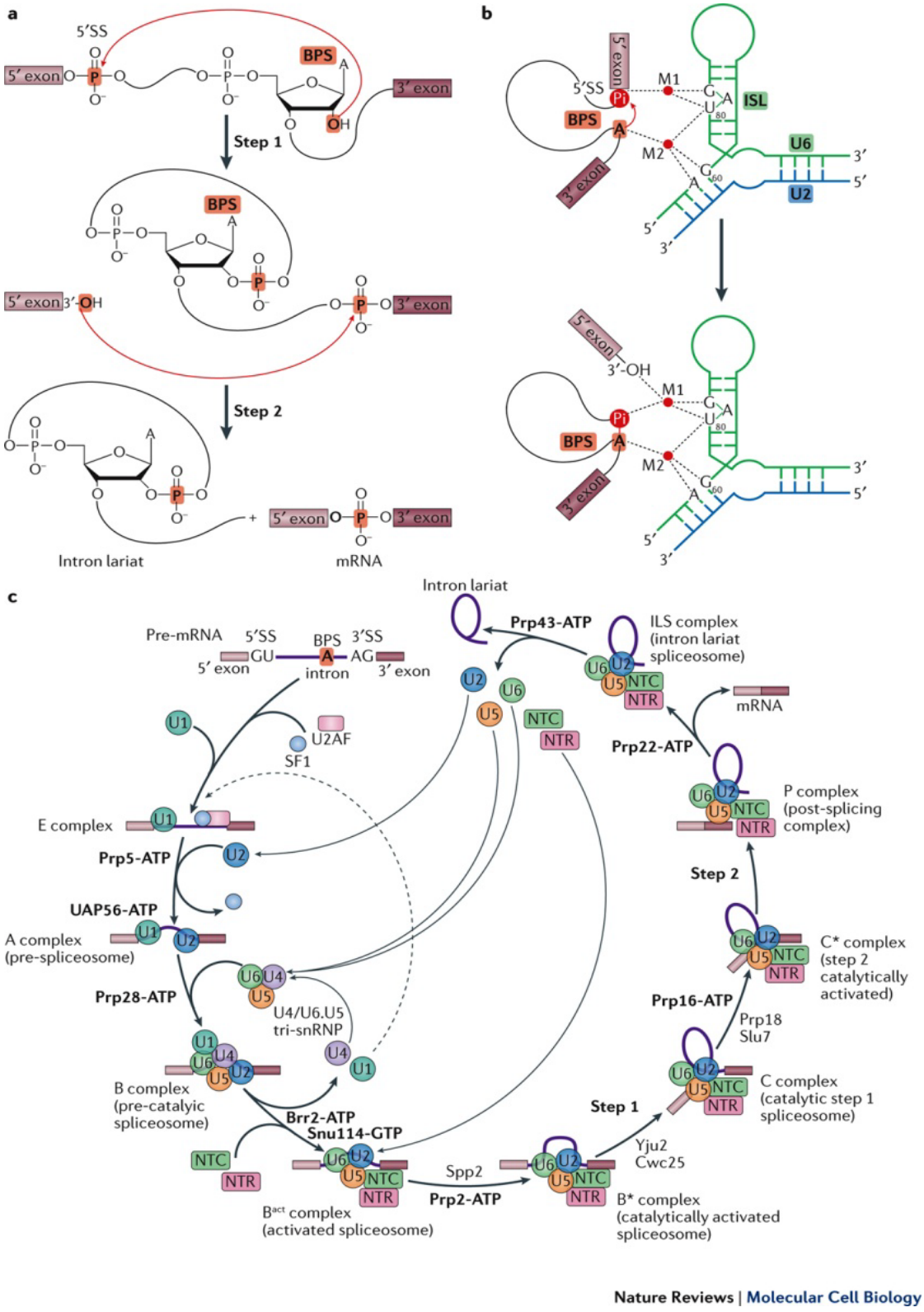
The Splicing Reaction

Biochemical characterization of *in vitro* splicing assays (Hernandez & Keller, 1983; Padgett, Hardy, et al., 1983) in the 1980s has enabled a thorough delineation of the splicing reaction (Domdey et al., 1984; Padgett et al., 1984; Rodriguez et al., 1984; Ruskin et al., 1984). All intron substrates contain three essential elements: 5' splice site (5'SS), branch point (BP) adenosine and 3' splice site (3'SS). While the 5'SS and 3'SS mark the beginning and ends of introns, BP adenosines are typically located between 18-40 nt upstream of the 3'SS (Taggart et al., 2017). These sequences are used in two S_N2-type transesterification reactions that result in intron removal and exon ligation. In

the first step (branching), the 2' hydroxyl of the BP adenosine attacks the phosphodiester linkage connecting the 5' exon and the 5'SS. This reaction releases the 5' exon and results in the formation of an intron-lariat-3' exon intermediate (**Figure 1.4A**). Importantly, the cleaved 5' exon is released with a 3' hydroxyl group, and this moiety attacks the 3'SS in the second transesterification reaction to ligate the 5' and 3' exons, as well as produce an intron lariat (**Figure 1.4A**).

The two steps of intron removal are catalyzed by a highly dynamic molecular machinery called the spliceosome (Brody & Abelson, 1985) (see The Splicing Pathway) along with ATP and magnesium (Mg^{2+}) (Hardy et al., 1984). While the ATP is required for the vast molecular gymnastics involved in splicing, two Mg^{2+} cations are carefully positioned in the active site (Fica et al., 2013), from where they activate nucleophiles and stabilize intermediates (Sontheimer et al., 1997; Steitz & Steitz, 1993) (**Figure 1.4B**).

The remarkable specificity and precision of splicing can – in part – be attributed to the sequence content contained in the 5'SS and 3'SS. This is because base pairing interactions between splice sequences and the macromolecular splicing machinery, the spliceosome, are required for the dynamic steps involved in intron processing (see The Splicing Pathway). Survey of transcriptome-wide splice sites highlights this relationship, as both 5'SS and 3'SS contain invariant nucleotide signatures and show complementarity to spliceosome components (**Figure 1.5**). While the mechanism of splicing is deeply conserved (Fica et al., 2013), there are species-specific differences in splice sequences. For example, in *S. cerevisiae*, 5'SS and BPs occur as stringent sequences with motifs GUAUGU and UACUA**A**C (BP adenosine in bold), but in insects, such as *D. melanogaster*, and mammals, such as *M. musculus* and others, these elements appear more degenerate. Similarly, in yeast, the 3'SS has a consensus of



Nature Reviews | Molecular Cell Biology

Figure 1.4. The pre-mRNA splicing reactions and the splicing cycle
A. Two steps of transesterification take place during pre-mRNA splicing. In step 1 (branching), the 2′-OH of the branch point sequence (BPS) adenine nucleotide attacks the phosphate of the guanine nucleotide at the 5′ end of the 5′ splice site (5′SS). In step 2 (ligation), the 3′-OH of the 3′ end nucleotide of the 5′ exon attacks the phosphate of the 5′ end nucleotide of the 3′ exon. B.. Coordination of the catalytic metal ions before and after the first step of transesterification. The upper and lower panels represent the B^{act} and C complexes, respectively. The two metals, designated as M1 and M2, are bound mainly by phosphates from U6 small nuclear RNA. In the first step of transesterification, M2 activates the nucleophile, whereas M1 stabilizes the leaving group. In the second step of transesterification, M1 activates the nucleophile, whereas M2 stabilizes the leaving group. C. Assembly and activation of the yeast spliceosome and the complete splicing-reaction cycle. The 5′SS, BPS and 3′SS are first recognized by the U1 small nuclear ribonucleoprotein (snRNP), splicing factor 1 (SF1; also known as branchpoint-bridging protein) and U2AF, respectively, forming an early spliceosome (known as the E complex). SF1 is displaced by the U2 snRNP to form the pre-spliceosome (A complex), which associates with the U4/U6.U5 tri-snRNP to assemble into the pre-catalytic spliceosome (B complex). The B complex represents the first fully assembled spliceosome. There are at least six additional distinct spliceosome complexes: B^{act}, B*, C, C*, P and the intron lariat spliceosome (ILS). Each complex has a unique composition, and conversion between complexes is driven by highly conserved RNA-dependent ATPase/helicases (in bold). Notably, a spliceosomal complex can have distinct conformational states, which may also differ in composition. For example, the B and ILS complexes each have at least two distinct conformations. Reprinted with permission (Shi, 2017)

A. 5'SS consensus motif

B. 3'SS consensus motif

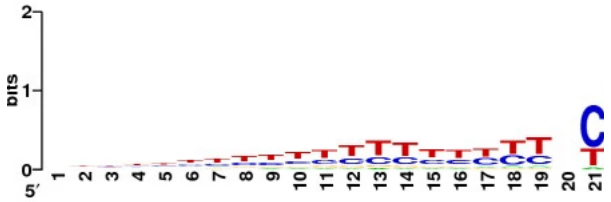


Figure 1.5. Splice site consensus motifs.
Used with permission (Rogozin et al., 2012)

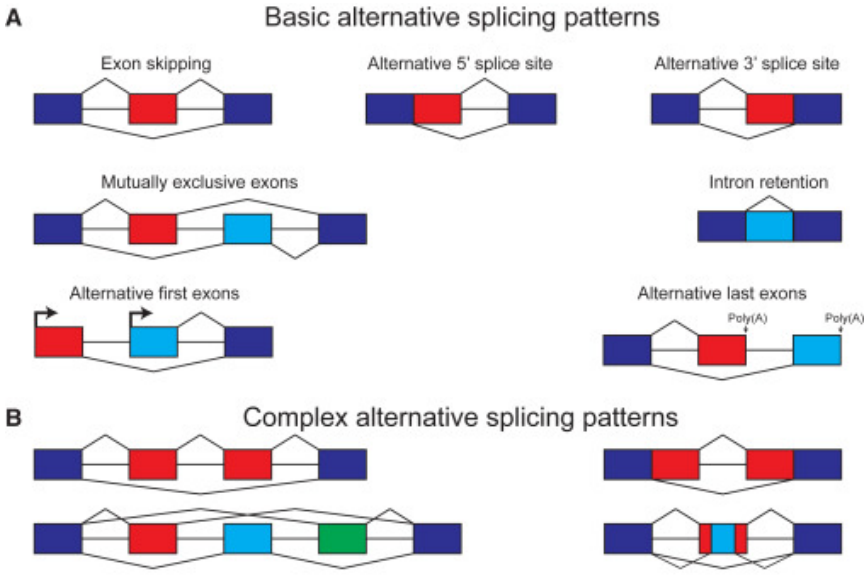


Figure 1.6. Alternative splicing patterns
(A and B) Basic (A) and complex (B) patterns of alternative splicing. Dark-blue boxes represent constitutively spliced exons. Red, light-blue, and green boxes represent alternatively spliced exons. Used with permission (Park et al., 2018).

YAG, but in other eukaryotes, the 3'SS has an additional polypyrimidine tract preceding YAG.

Components of the Spliceosome

The spliceosome is the enzyme that catalyzes the splicing reaction. The core of the spliceosome contains five small nuclear RNAs (snRNAs) and approximately hundred proteins, thus it is a large, dynamic, ribonucleoprotein (RNP) (Kastner et al., 2019). The five core RNAs within this machinery are the U1, U2, U4, U5 and U6 snRNAs. These RNAs are keenly involved in indispensable tasks including the recognition of splice sites, as well as organization of the active site during splicing. Each snRNA assembles as a small nuclear RNP (snRNP) with seven homologous Smith (Sm) proteins (seven LSm proteins for U6), as well as snRNP-specific proteins (Achsel et al., 1999; Bringmann & Lührmann, 1986; Lerner & Steitz, 1979; Séraphin, 1995). Additionally, the snRNPs engage with a host of splicing factors and ATP-dependent RNA helicases that are involved in the process of splicing.

The formation and execution of splicing also involves scores of other proteins. The most well-known of these are found within the nineteen complex (NTC, also known as the PRP19-CDC5L complex) and the NTC-related (NTR) complex (Chan et al., 2003; Tarn et al., 1994). The functional importance of these proteins is further elaborated in the following section.

The Splicing Pathway

The intricate dynamics of intron removal are illustrated in **Figure 1.4C**. The process begins when components of the spliceosome recognize critical splicing elements (the 5'SS, 3'SS and the BP), a stage that is referred to as E complex

formation. This is accomplished by binding of U1 snRNP to the 5'SS through base pairing interactions between the 5' end of the U1 snRNA and the 5'SS (Lerner et al., 1980; Zhuang & Weiner, 1986). Similarly, at the 3' end of the intron, the BP and the 3'SS are bound by the SF1 and the U2AF complexes (Berglund et al., 1998). The formation of the E complex essentially determines intron boundaries and is an important step towards splicing commitment. This step is extensively regulated through several mechanisms (summarized in Alternative Splicing) and results in diversification of mRNA sequences through alternative splice site choice (Ule & Blencowe, 2019).

In the next step, the SF1 and U2AF complexes are displaced through the activity of ATP-dependent helicases, and U2 snRNP binds the BP sequence. This stage is also referred to as the A complex or pre-spliceosome. Here base pairing interactions between the U2 snRNA and the BP on the pre-mRNA result in the formation of the branch helix (Parker et al., 1987; J. Wu & Manley, 1989). Subsequently, the preassembled U4/U6.U5 tri-snRNP joins the pre-spliceosome to form the pre-B complex. Although the tri-snRNP joins U1 and U2 snRNP, no major rearrangements occur at the pre-B stage. The recruitment of the tri-snRNP is facilitated by interactions between the 5' end of the U2 snRNA and the 3' end of the U6 snRNA (Hausner et al., 1990). While the U6 and U5 snRNPs are important components of the spliceosome active site, at this stage, the active site is yet to be formed. Critically, base pairing between U4/U6 maintain the U6 snRNA in a pre-catalytic conformation (Nguyen et al., 2015; Wan et al., 2016). At this state, all the RNA elements necessary for a catalytically active spliceosome have connected but require major conformational changes. These contortions occur in stepwise fashion, initiated by activation of the ATP-dependent DEAD-box helicase PRP28. PRP28 unwinds the 5'SS from the U1 snRNP, which allows the free 5'SS to anneal to the U6 snRNP via the ACAGAGA loop (Charenton et al., 2019). This

interaction induces further conformational changes, including the loading of the helicase Brr2 onto the U4 snRNP.

The transition from B to B^{act} complex involves widespread remodeling and begins with the unwinding of the extensive base pairing between the U4 and U6 snRNAs by Brr2. This frees up U6 to form additional interactions with the U2 snRNP (Yan et al., 2016). An important consequence of this new U2/U6 interface is the formation of two short RNA helices adjacent to the internal-stem-loop (ISL) in U6 snRNA (Madhani & Guthrie, 1992). These rearrangements serve to correctly position Mg^{2+} ion-coordinating phosphate groups from the helix and the ISL within the active site (Fica et al., 2013; Steitz & Steitz, 1993). In addition to organizing the active site, the extended U2/U6 pairing and the interactions between the U5 snRNA loop1 and the sequence immediately upstream of the 5'SS juxtaposes the 5'SS and the 3'SS (Sontheimer & Steitz, 1993). The NTC and NTR protein complexes are required for the formation of the B^{act} complex and chiefly function by constraining and stabilizing the RNA catalytic core (Chan et al., 2003; Fabrizio et al., 2009). While in close proximity within a catalytically competent active site, the 5'SS and BP adenosine are obstructed by CWC24 (5'SS) and proteins of the SF3a (5'SS) and the SF3b (BP) complexes respectively (Haselbach et al., 2018; N.-Y. Wu et al., 2017; X. Zhang et al., 2018).

Disruption of these inhibitory contacts requires the activity of four DEAH-box ATPase enzymes, the first of which is Prp2 (Cordin et al., 2012). Prp2 can bind single-stranded RNAs at 3' ends and translocate in a 3' to 5' direction, disrupting dsRNA or RNA-protein interactions (Pyle, 2008). This activity is thought to be responsible for disrupting the inhibitory binding of the SF3a and SF3b within the active site. Once released, the BP adenosine, bulged out of the branch helix, is available to attack the 5'SS. At this stage, the spliceosome is catalytically competent and is referred to as the B* complex. When the first step of splicing has occurred, the massive complex is known

as the C complex. Several auxiliary proteins promote the chemistry of this branching step. These include Yju2, Cwc25 and NTC component Isy1 (Wan et al., 2019).

The active site undergoes further remodeling to accommodate the exon ligation step of splicing. These movements are orchestrated by the DEAH-box ATPase Prp16, which destabilizes key existing contacts, and facilitates new interactions (B Schwer & Guthrie, 1992). Prp16 activity results in the rotation and conversion of the branch helix to a canonical A-form helix. These changes accommodate space in the active site for correct positioning of the 3'SS (Bertram et al., 2017; Fica et al., 2017; Yan et al., 2017) and the remodeled conformation is referred to as the C* complex. At this stage, factors including Prp18 enable critical base pairing interactions between the exons to the U5 snRNA loop 1 (Horowitz, 2012; James et al., 2002). The 3'SS, unlike the 5'SS and the BP is not recognized through interactions with snRNPs. Instead, the 3'SS is generally selected as the first YAG sequences more than 10 nt downstream of the BP (Horowitz, 2012). More specifically, the 3'SS forms non-Watson-Crick interactions with the 5'SS and the BP adenosine. The unique structure of the branched lariat consists of a covalent linkage between the 5' end of the intron (G nucleotide) and the BP adenosine. These two nucleotides interact with A and G nucleotide of the 3'SS at the 3' end of the intron (S. Liu et al., 2017; Parker & Siliciano, 1993; Wilkinson et al., 2017).

Exon ligation results in the formation of the postcatalytic (P) complex. Transition from this state to the next requires the release of the mRNA. The conformational changes occurring during this phase are poorly understood. However, it is well appreciated that activation of the helicase Prp22 releases the ligated exons through 3' to 5' translocation along the 3' exon of the mRNA (Company et al., 1991; Beate Schwer, 2008). With the mRNA released, the remaining structure is referred to as the intron lariat spliceosome (ILS). This must be disassembled to allow reuse of the spliceosomal components as well as decay of the intron lariat. Activity is set in motion through the

helicase Prp43, which also has important roles in ribosome biogenesis (Combs et al., 2006; Leeds et al., 2006). Prp43 is engaged by the Ntr1 complex, resulting in the release of core active site components, including the U6, U2 and U5 snRNP, as well as the NTC proteins (Fourmann et al., 2013). Disassembly is under strict regulation to ensure that only postcatalytic spliceosomes or those harboring weak substrates are allowed to initiate this procedure (Koodathingal et al., 2010).

The Exon Junction Complex (EJC) is deposited on RNAs during splicing

In metazoans, a multi-protein complex is deposited onto mRNAs at exon junction during the process of splicing. This complex consists of three core members, eIF4AIII, Y14 and MAGOH. Assembly of the core complex initiates during splicing when spliceosomal factor Cwc22 binds the DEAD-box protein eIF4AIII using its MIF4G domain (Barbosa et al., 2012; Steckelberg et al., 2012). Within the spliceosome, Cwc22 can function as a molecular ruler and deposits eIF4AIII 20-24 nt upstream of the exon junction. The precise moment of eIF4AIII/RNA binding is unknown, but has been observed as early as the C complex in human cryo-EM structures (Bertram et al., 2017; Galej et al., 2016). Subsequently, a MAGOH-Y14 heterodimer binds eIF4A3 to create the stable core-EJC, but details of this assembly are still elusive.

The core EJC has strong interactions with a host of factors involved in broad gene regulatory mechanisms (reviewed in (Schlautmann & Gehring, 2020)). In the context of intron removal, the EJC associates with splicing factor RNPS1, typically found in the ASAP/PSAP (ACIN1/PNN-RNPS1-SAP1) complexes, and also associates with higher order SR-protein containing mRNPs (Singh et al., 2012). Overall, EJC activity has been characterized in several RNA metabolism pathways. These include nuclear roles, such as splicing (Ashton-Beaucage et al., 2010; Blazquez et al., 2018; Boehm et al.,

2018; Fukumura et al., 2016; Hayashi et al., 2014; Lence et al., 2016; Malone et al., 2014; Z. Wang et al., 2014), RNA Pol II promoter-proximal pause release (Akhtar et al., 2019), and nuclear RNP export (reviewed in Heath et al., 2016), as well as cytoplasmic functions, such as enhancement of translation (Ma et al., 2008; Palmiter et al., 1991) and mRNA quality control (nonsense-mediated decay) (reviewed in Kishor et al., 2019). These roles highlight the central role of EJC in facilitating and surveying accurate gene expression.

Alternative Splicing

The discovery of the spliceosome/splicing and its deep conservation within eukaryotes also spawned fields concerned with the functional importance of splicing to biological systems. The question has been approached using *in vivo* as well as *in silico* methodology, leading to two crucial principles. First, that splicing allows diversification of the proteome. This was first predicted by Walter Gilbert in a 1978 article titled “Why genes in pieces?” (the terms introns and exons were first described here) (Gilbert, 1978b). Gilbert proposed that exon shuffling or duplication through recombination could serve as a means to assort and expand useful peptide functions. The discovery of nonrandom distribution of intron phase (higher proportion of phase 0 introns) (Long et al., 1995) provides critical support for this hypothesis.

Second, that splicing may serve as a means of gene regulation. In the same article, Gilbert argued that modification of splicing efficiency could allow genes control over what sequences are included into mRNAs, and hence provide regulatory capacity for multiple functional outputs (Gilbert, 1978a). Such a phenomenon was soon discovered for the immunoglobulin μ gene (Alt et al., 1980; Early et al., 1980) and quickly expanded to other genes thereafter. This ability to alternate mRNA sequence through

splice site choice is now commonly referred to as alternative splicing (AS) and varies classes of AS have been schematized in **Figure 1.6**. More recently, the application of high-throughput sequencing methods to assemble genomes and analyze transcriptomes has greatly expanded our appreciation of alternative splicing. While broadly observable among eukaryotes, the level and class of AS mapping to multiexon genes can vary substantially between species (Grau-Bové et al., 2018; Kim et al., 2007).

AS is one of the main sources of proteomic diversity in multicellular eukaryotes. A striking example is the *Drosophila melanogaster* gene *Down syndrome cell adhesion molecule (Dscam),* for which 38016 distinct AS products are possible (Schmucker et al., 2000), in comparison to 15500 genes expressed in the organism. In humans, ~ 95% of multiexon genes undergo AS (Pan et al., 2008; E. T. Wang et al., 2012), but while AS is inherently expansive, regulation of this process brings coherence to gene expression by establishing cell type-specific patterns. These regulated patterns are now understood to be important for the development of cell types and cellular behaviors (Baralle & Giudice, 2017a). These include development of tissues and organs such as the brain, heart, muscles as well as dynamic states such as the epithelial-mesenchymal transition (Baralle & Giudice, 2017b; Ule & Blencowe, 2019). As this dissertation explores mechanisms of cryptic splice site activation and avoidance, it is critical to appreciate what is already known about mechanisms of alternative SS choice.

Mechanisms of alternative splicing

Since splicing is co-transcriptional, efforts have been directed to understand how every aspect of the co-transcriptional nuclear environment may influence splicing. This includes abstract concepts like gene architecture, as well as tangible ideas like the interaction between RNA and splicing factors. Gene expression is highly coordinated,

and this remarkable feature can be thoroughly appreciated through the lens of alternative splicing.

Intron-Exon architecture dictates splice site definition

Early experiments aimed at investigating the effects of exon and intron length on splicing revealed important principles regarding the impact of gene architecture on pre-mRNA processing. Chief among these is the notion that SS are not recognized independently but are instead defined in concert. Typically, this is taken to mean that the 5'SS and 3'SS within an intron are recognized simultaneously, a process called intron definition (**Figure 1.7**). This may be the predominant mechanism in lower eukaryotes where introns are usually small and exons, considerably longer. But in higher eukaryotes, such as vertebrates, intron-containing genes have exons of average length ~145 nt that sandwich considerably longer introns. For such cases, definition and juxtaposition of SS via intron definition presents challenges due to the intron length and alternate strategies are required to pair SS. Splicing minigene reporters with long introns demonstrated that for long intron/short exon combinations, spliceosomal components assemble across exons, a strategy called exon definition (**Figure 1.7**). In exon definition, exons are first recognized by the splicing machinery, and only later the intervening sequence marked as introns (Berget, 1995; De Conti et al., 2013).

It is important to note that choice of exon versus intron definition strategy leads to distinct products in the presence of SS regulation. For instance, the suppression of a SS on an internal exon will lead to intron retention if processed by intron definition and internal exon skipping if processed by exon definition (**Figure 1.7**). In the latter case, it has also been observed that exon length may influence splicing efficiency. A striking

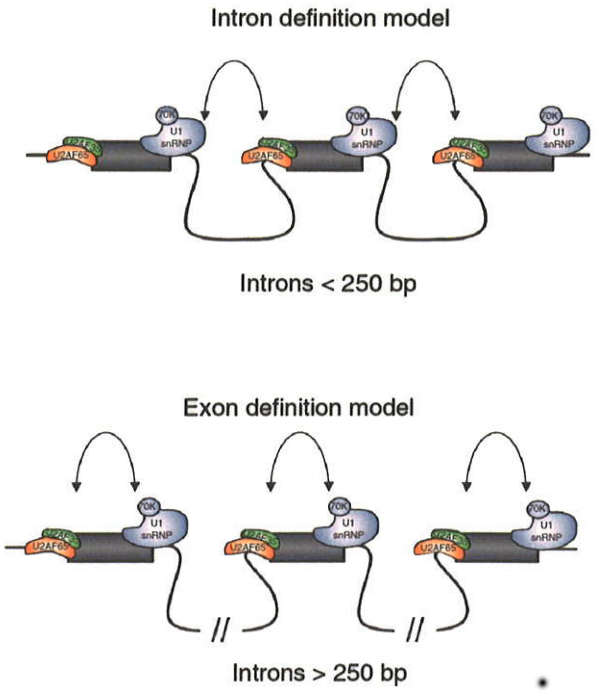


Figure 1.7. Exon and Intron definition models.
The top panel depicts the Intron definition model according to which pairing between the splice sites takes place across an intron when long exons are separated by short (< 250 bp) introns. On the other hand the bottom panel shows the Exon definition model, where the splice site communication occurs across exons when they are separated by long (> 250 bp) introns. Used with permission (De Conti et al., 2013)

example is that of the mouse c-*src* gene, where extending the alternatively spliced N1 exon from 18 to 109 nt results in constitutive inclusion (Black, 1991).

Splicing Regulatory Elements (SRE) and *trans*-acting factors

Hundreds of proteins are involved in the dynamic assembly of the catalytically active spliceosome. Some of these, such as the SF1 and U2AF complexes, make direct contact with RNA; interactions that are fundamental for progress of the splicing reaction. However, there is a whole other set of RNA binding proteins (RBPs) that dock onto pre-mRNA during transcription and regulate SS choice. Such factors are commonly referred to as splicing factors and their sites of interactions on RNA, splicing regulatory elements (SREs). SREs can be found on exonic or intronic sequences, and splicing factor engagement at such sites can be both enhancing or inhibitory for nearby SS, depending on the context. Direct contacts are established through the activity of RNA binding domains (RBD), such as RNA recognition motifs (RRMs), zinc fingers, KH domains and double-stranded RNA binding motifs (dsRBMs), and give RBPs affinity for a range of substrates, from highly sequence/structure specific to independent (Antoine Cléry and Frédéric H.-T. Allain, 2011). The SRE/*trans*-acting factor mechanism of SS regulation is explained below.

SR proteins are a well-studied class of splicing factors, named for their C-terminal RS domain which consists mostly of arginine and serine residues. They are known for their ability to stimulate inclusion of exons with weak splice sites by binding to exonic splicing enhancer sequences (ESEs). This activity is dependent on bridging interactions between the RS domains of the SR factors and U2AF (bound at 3'SS) or RNA duplexes formed by U2 and U6 snRNAs at the branchpoint and 5'SS (Shen & Green, 2006; Tian & Maniatis, 1993; J. Y. Wu & Maniatis, 1993). In addition to exon

inclusion, SR proteins can also dictate choice of alternative 5'SS selection. While the logic seems to be that SR factor/ESE engagement enhances selection of the intron proximal 5'SS, a mechanistic appreciation for this activity remains elusive (Erkelenz et al., 2013). Orthogonally, splicing factors can also silence SS when bound to pre-mRNAs. For instance, SRSF7 and PTB are examples of factors known to have suppressive effects when bound within introns. Mechanisms of SS inhibition include stabilization of the U1 snRNP binding to the 5'SS, prevent progress of the splicing reaction (Sharma et al., 2011), SS occlusion (Boehm & Gehring, 2016) and others (reviewed in Lee & Rio, 2015).

Regulation of RNA Polymerase II and transcriptional control

RNA Polymerase II (RNAPII) transcribes pre-mRNA and is centrally poised to influence diverse RNA metabolism that pre-mRNA must undergo to become competent mRNA. Consistently, RNAPII has been shown to have direct and indirect effects on co-transcriptional activities such as RNA modifications (5' capping, for example), splicing and cleavage and polyadenylation (McCracken, Fong, Rosonina, et al., 1997; McCracken, Fong, Yankulov, et al., 1997; Mortillaro et al., 1996; Yuryev et al., 1996). Implicit in these connections is the notion that transcriptional control has consequences on regulation of processing. For example, while RNAPII may transcribe different genes at different rates (Fukaya et al., 2017; Jonkers & Lis, 2015), altering the speed of RNAPII can lead to alternative splicing. This is best summarized by the "window of opportunity" or "first come, first served" model, which argues that upstream SS have an advantage as they are transcribed first, and modulating the rate of availability of downstream competing SS through changes in transcription rate will alter SS choice (de la Mata et

al., 2003; Dujardin et al., 2013; Saldi et al., 2016). In support of this model, RNAPII has been observed to temporarily pause at splice sites (Milligan et al., 2017).

Beyond regulation of transcription rate, RNAPII also acts as a modular platform to recruit diverse enzymes and regulators. The C-terminal heptad repeat domain (CTD) of the RNAPII large subunit is appreciated as the main scaffolding unit and can control recruitment through dynamic phosphorylation. The CTD modification state, in turn can be dictated by promoter/enhancer signals, the chromatin environment, and transcriptional dynamics. Changes in any of these elements can have cascading effects on RNA processing, most notably on alternative splicing. These properties have been integrated as the "mRNA factory" model, in which RNAPII is proposed to couple transcription with processing by forming a large complex that brings together factors involved in synthesis as well as processing (Saldi et al., 2016).

Histone modifications and AS

Genomic DNA is typically wrapped around nucleosomes, which consist of histone octamers. The N-terminal tails of histone proteins are exposed as they project outwards from the nucleosome. These tails can be post-translational modified at several positions and constitutes an additional layer of information that is integrated during gene expression (Lawrence et al., 2016). Specific combination of histone modifications have been observed within expressed genes, and recent studies have made noteworthy connections with the RNA splicing as well (Luco & Misteli, 2011).

One mode in which histones may regulate splicing is through RNAPII. It has been suggested that nucleosome density and positioning may alter RNAPII velocity, which in turn could influence splice site choice as described above. Furthermore, histone modifications are also able to recruit splicing factors through reader proteins, thereby

increasing the local concentration of *trans*-acting factors. In support of this model, certain histone modifications are enriched on exons relative to introns (Andersson et al., 2009; Huff et al., 2010; Kolasinska-Zwierz et al., 2009). As an example, the polypyrimidine tract-binding protein (PTB), which binds silencing elements surrounding exons and causes exon skipping, can be recruited to specific pre-mRNAs through the H3K36me3 reader protein MRG15 (Luco et al., 2010). In a similar vein, one mechanism of recruitment of U2snRNP to pre-mRNAs of active genes is through the H3K4me3 reader protein CHD1, which binds at the 5' end of expressed genes (Sims et al., 2007). While these models engage ideas related to splice site choice, other studies have pointed towards alternate stages of spliceosome assembly as nodes of regulation. A recent report indicated that the H3K36 methylation reader Eaf3 is required to recruit the NTC complex to the spliceosome. Hence, without the NTC complex, the splicing machinery is unable to form or maintain the catalytic B spliceosome, and the result is intron retention (Leung et al., 2019). While this discovery was not made within the context of regulated AS, it is certainly plausible that histone modifications could influence AS through recruitment of factors required to assemble catalytically active spliceosomes.

RNA modifications

Over 100 types of chemical modifications have been identified on cellular RNAs on all four canonical nucleotide residues. While ribosomal RNA (rRNA) and transfer RNA (tRNA) are believed to be the most heavily modified RNAs in the cell, mRNAs are also modified – the 5' cap structure being the best recognized. Remarkably, there are also "internal" mRNA modifications such as N^6-methyladenosine (m6A), N^1-methyladenosine (m1A), 2'-O-methylation, 5-methylcytosine (m5C), pseudouridine and many others (reviewed in Li & Mason, 2014; Roundtree et al., 2017). An important question, that

remains to be fully explored is whether these modifications can influence RNA processing and gene expression.

Recent technical advances have provided elaboration of several modification pathways. This includes a deeper and higher resolution mapping of modification landscapes, elucidation of factors present in “writer” complexes that catalyze RNA modifications, discovery of “reader” proteins that bind to modified RNAs, as well as investigations of potential functions. These studies have suggested strong links between internal RNA modifications and intron processing. The direct interaction between splicing factor WTAP and the METTL3/METTL14 m6A writer complex is a striking example of this concept. Loss of WTAP results in lower proportions of methylated adenosine and molecular evidence indicates that WTAP is required to effectively recruit to RNA to the METTL3/METTL14 complex (Lence et al., 2016; J. Liu et al., 2014; Ping et al., 2014). Similarly, the transcriptomes of cells that lack functional writer complexes has provided meaningful insights into direct and indirect consequences on RNA metabolism, including AS. An instructive example is the sex-specific processing of *Sxl,* the master regulator of sex determination in *Drosophila.* XY flies (males) include cassette exon 4 in *Sxl*, whereas XX flies (females) skip the same. However, female flies that lack METTL3/METTL14-dependent m6A have altered *Sxl* splicing with cassette exon 4 inclusion, and display a number of male characteristics (Haussmann et al., 2016; Kan et al., 2017; Lence et al., 2016). The same studies also demonstrated that the nuclear m6A reader YTHDC – a splicing factor – is also required for correct female-specific *Sxl* splicing. Thus, an important model that has emerged through such studies is that modified RNAs act as dynamic SREs that are binding platforms for specific reader proteins.

It is noteworthy that modified SREs offer more dynamic control over gene regulation. *cis*-SREs that are based on canonical nucleotides are always included on

pre-mRNAs, thus regulation of AS may rely mainly on the availability of *trans*-acting splicing factors. However, since m6A is dynamic, there is potential for control based on both SRE and *trans*-acting factors.

Deviations from canonical splicing

Splice site definition and intron removal are thought to occur as a single splicing reaction using one pre-mRNA substrate (**Figure 1.8A**, *cis*-splicing). However, it is quite possible to imagine scenarios where potentially more complex phenomenon might occur. For instance, in the human genome, there are at least 1200 introns that are longer than 100000 nt (Shepard et al., 2009). Finding precise splice sites within such large search spaces may pose challenges, especially considering that SS motifs appear degenerate as introns get longer. Furthermore, unusual attributes suggest additional layers of regulation may be required to facilitate pre-mRNA maturation. Indeed, there are observable RNA intermediates and products that indicate complex RNA processing. For example, intragenic trans-splicing, a process in which two distinct pre-mRNA molecules from the same genes are spliced together has been observed at low levels in many eukaryotic species (reviewed in Hastings, 2005; Lei et al., 2016) and could enhance accurate splicing of long introns. Along the same lines, recursive splicing (RS) is a phenomenon that suggests a single intron can be removed as multiple smaller segments and has been observed in metazoans (**Figure 1.8B**, recursive splicing) (Georgomanolis et al., 2016). In fact, splicing need not even be linear, as backsplicing within long introns has also been noted to produce circular RNA in eukaryotes (Salzman et al., 2012, 2013; P. L. Wang et al., 2014; Westholm et al., 2014). While these complex modes of splicing challenge existing models: namely, 5' to 3' and one intron/pre-mRNA per reaction, they may provide additional control over intron processing and gene expression not afforded

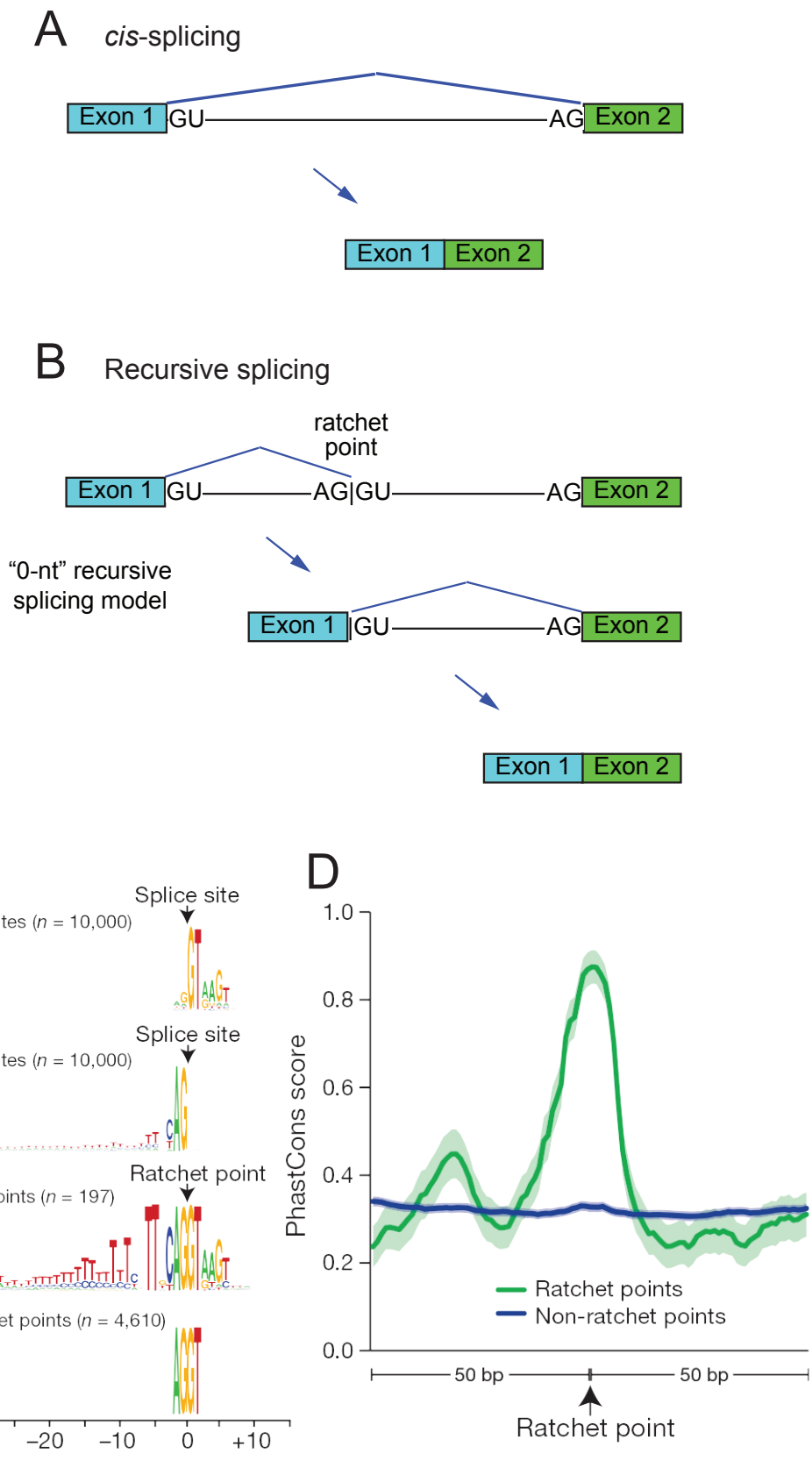

A
cis-splicing
Exon 1
GU
AG
Exon 2
Exon 1
Exon 2
B
Recursive splicing
ratchet point
Exon 1
GU
AG|GU
AG
Exon 2
"0-nt" recursive splicing model
Exon 1
GU
AG
Exon 2
Exon 1
Exon 2
C
Splice site
5′ splice sites (n = 10,000)
Bits
Splice site
3′ splice sites (n = 10,000)
Bits
Ratchet point
Ratchet points (n = 197)
Bits
Non-ratchet points (n = 4,610)
Bits
−40
−30
−20
−10
0
+10
D
PhastCons score
1.0
0.8
0.6
0.4
0.2
0.0
Ratchet points
Non-ratchet points
50 bp
50 bp
Ratchet point

Figure 1.8. Recursive splicing removes a large inton in two or more steps.
(A) cis-splicing. The canonical mode of splicing. The three splicing elements (5'SS, 3'SS and BPS) are used to remove the intron in one step.
(B) Recursive splicing. In toy example, the same intron removal occurs in two smaller steps. In the first, the ratchet point 3'SS is activated along with the canonical 5'SS to remove a fragment of the intron. This regenerates a 5'SS, which is used subsequent step along with the canonical 3'SS to remove the remaining intronic sequence.
(C) Comparison of splice motifs. Note that ratchet point (recursive splice sites) are tandem 3'SS and 5'SS.
(D) Conservation of sequence around the RP.
C and D are used with permission (Duff, 2015)

by canonical splicing. That this may be the case is supported by the conservation of these processes within species separated by millions of years of evolution (Georgomanolis et al., 2016; Lei et al., 2016). Nevertheless, very little is known about the mechanism and function of non-canonical modes of splicing and their requirements for long intron processing. As my work is generally focused on cryptic splices sites and recursive splicing, I pay special attention to these in the sections below.

Cryptic splice sites in pre-mRNA

Intron removal is typically schematized using a toy model where SS mark intron-exon boundaries and the only other *cis*-elements depicted are perhaps the branch point sequence and SREs. However, in practice picking out a "real" SS can be a daunting task because splice motifs are short, and there are many similar sequences within pre-mRNA transcripts and certainly within the transcriptome (Roca et al., 2013). Such sequences that match consensus but do not show activity are sometimes referred to as pseudo, cryptic or decoy splice sites and they are opportunistically activated when canonical SS are mutated, such as in human diseases (Anna & Monika, 2018; Kahles et al., 2018; Roca et al., 2013). However, these adjectives are perhaps misleading as they communicate a state of silence or inactivity. In fact, cryptic splice sites tend to interact with RBPs and are typically involved in alternative splicing through competition with canonical splice sites (Coté et al., 2001; Ule & Blencowe, 2019). Beyond splicing regulation, cryptic splice sites are also involved in other processes.

Transposable elements are an interesting case study as they are a source for cryptic SS and exons. For example, *Alu* elements in primates are known to contain cryptic exons (Keren et al., 2010; Sibley et al., 2016). When in reverse orientation within a gene, these sequences contain rich poly(U) tracts that can be binding platforms for

splicing factors U2AF2 and T cell-restricted intracellular antigen (TIA) proteins. The recruitment of these factors can facilitate cryptic exon definition but is typically prevented by the repressive interactor hnRNPC (Zarnack et al., 2013). Similarly, antisense L1 elements were recently reported to contain cryptic splice sites that are silenced by a host of RBPs (Attig et al., 2018). Within this context, cryptic splice sites and exons appear to be involved in neofunctionalization (Attig et al., 2018). Thus, even though silent, cryptic splice sties require distinct regulatory programs during intron removal.

Cryptic splice sites can also have other pre-mRNA processing activity. The most famous role for cryptic 5'SS outside of splicing lies in the prevention of premature cleavage and polyadenylation (PCPA) within long introns, a process called telescripting (Berg et al., 2012; Oh et al., 2017). While this mechanism may involve prevention of transcript cleavage, U1 snRNP interaction with cryptic 5'SS on RNA can also inhibit the polyadenylation machinery (Furth et al., 1994; Guan et al., 2007). Thus, availability of cryptic 5'SS may also be involved in the regulation of transcriptional termination and mRNA stability.

Recursive Splicing

Recursive splicing is a phenomenon in which an intron is removed as multiple smaller fragments (Burnette et al., 2005; Georgomanolis et al., 2016). It is distinguished by characteristic splice substrates, called recursive splice sites (RSS) or ratchet points (RP), consisting of tandem splice acceptor and donor sequences (**Figure 1.8B-C**). As these sites cannot be inferred by sequencing mRNA, they represent yet another class of cryptic splice sites. The unique architecture of these elements resemble exons of length zero, hence they have been commonly referred as zero-nucleotide exons (Burnette et al., 2005; Duff et al., 2015; Hatton et al., 1998). Similar to splicing via exon definition, it

has been proposed that the RSS first functions as a 3'SS, permitting removal of the upstream intron segment and regenerating a 5'SS. The regenerated 5'SS pairs with a downstream splice acceptor to excise the remaining intron (**Figure 1.8B**).

RS was first detected within the *Ultrabithorax (Ubx)* gene in *Drosophila melanogaster* (Hatton et al., 1998). RSS were observed at the 5' junction of two 51 nt cassette microexons and was demonstrated as the mechanism of exon skipping. Subsequently, computational searches combined with molecular analysis have verified RSS at a handful of *Drosophila* cassette exons (Burnette et al., 2005; Conklin et al., 2005). The same computational efforts lead to the discovery of 165 intronic RPs that did not appear to be associated with annotated exons and were located within long introns > 10000 nt (Burnette et al., 2005). This initial scope of RS reported by the Javier Lopez lab indicated a largely intronic phenomenon, subdividing long introns into smaller fragments. This view has been strengthened and expanded by subsequent reports that have found hundreds of RPs within long introns (average length ~ 50000 nt) in *Drosophila* (Duff et al., 2015; Pai et al., 2018) and within introns and exons in mammals (Blazquez et al., 2018; Boehm et al., 2018; Sibley et al., 2015; X.-O. Zhang et al., 2018). Phylogenetic analysis has revealed that intronic RPs are deeply conserved, with the strongest signals concentrated at the AG|GU sequence that represents the invariant signatures of the 3' and 5'SS (**Figure 1.8D**) (Duff et al., 2015; Sibley et al., 2015).

Despite an expanding annotation of these sites, at the onset of my research, very little was known about *Drosophila* intronic RPs beyond phenomenology. In fact, dissecting this pathway with molecular biology proved technically challenging since recursive RNA intermediates are typically transient. Furthermore, RP cloning and manipulation was also difficult because of the unusually long sequences involved. Hence, elementary questions regarding RS were unanswered. For example, are recursive intermediate pre-mRNA converted to mRNA, or are these intermediates

accidental and unstable products? Furthermore, if these sites embody *bona fide* pre-mRNA intermediates, what is the basis for their requirement, and how is processing facilitated by RS? Even at the mechanistic level, intronic RPs fall outside the canons of splicing – because the 3' and 5'SS are fused together, these substrates appeared suboptimal for exon definition. However, despite the challenges, finding solutions to the above questions may reveal novel principles about pre-mRNA splicing, especially within long introns.

Aberrant RNA splicing as a basis for disease and disorder

The majority of human genes require intron removal; hence, splicing is a fundamental organismal requirement. However, research on human disease strongly suggests that many pathological conditions have underlying splicing defects. Consequently, a significant portion of RNA research is dedicated to identifying functionally relevant and disease-specific splicing targets, as well as rationale-based therapeutic strategies.

There are distinct classes of splicing errors that can lead to disease. The most common cause is pre-mRNA mutations that lead to mis-splicing. These can either be mutations of the three critical splicing elements, or alterations of regulatory sequences that assist during SS definition. For example, a point mutation in the *HBB* gene that encodes β-globin causes β^+-thalassaemia due to a splicing defect (Busslinger et al., 1981; Maquat et al., 1980; Spritz et al., 1981). Another example is the *LMNA* gene, where distinct splicing mutations result in multiple pathological phenotypes, referred to as laminopathies (Scotti & Swanson, 2016).

Splicing errors can also be cause by mutations in the core spliceosome. Impaired constitutive and alternative splicing have been observed in these instances. Examples

include diseases, retinal degenerative disorders as well as cancer. Interestingly, these mutations reveal distinct tissue sensitivities to splicing perturbations. For instance, mutations in the *PRPF6* gene have been shown to cause Retinitis pigmentosa, a disease that leads to blindness. Conversely, mutations in *U2AF1* have been found in myelodysplastic syndromes (MDS) (Scotti & Swanson, 2016). Another form of global splicing errors can be caused by mutations in *trans*-acting factors. *TARDP* (TDP43) and *FUS* mutations are a known causal lesion in patients with amyotrophic lateral sclerosis (Arnold et al., 2013; Sun et al., 2015).

The broad role that splicing can play in the disorder will continue to expand as more is understood about the splicing reaction.

Thesis objectives

My thesis explores the landscape, mechanism and function of recursive splice sites. In this process, I focused on three important concepts:

1. How are intronic recursive splice sites (zero-nucleotide exons) defined?

In this chapter, I use a combination of genetics, molecular biology and bioinformatics to show that intronic RPs are actually defined using short cryptic RS-exons. The short exons are formed the RP 3'SS and a previously unknown downstream cryptic 5'SS (RS-exon 5'SS). I show that mutation of the RS-exon 5'SS results in loss of recursive splicing.

2. How are cryptic RS-exons regulated and what is their contribution to long intron removal?

In this chapter, I investigate mechanisms that regulate RP 5'SS versus RS-exon 5'SS choice. My experiments indicate that 5'SS strength, splicing regulatory elements as well as the EJC may determine the output of RS-exon alternative splicing. Additionally, I also

make the first ever intronic RP deletions in any model organism. I studying RNA processing in these alleles and using rt-PCR, I note that mRNA production is unaffected. Encouragingly, deletion of the expressed *Ubx* m1 and m2 recursive exons produce phenotypes and splicing changes.

3. The EJC silences cryptic splice sites that found at or near exon junction sequences.

In the final chapter, I surprisingly find that EJC loss results in the activation of hundreds of unannotated, weak 5' and 3' splice sites. The derepressed cryptic splice sites are most commonly found near exon junction sequences. I show that the EJC suppresses these sites via cryptic SS occlusion. Critically, I determine that exon junction sequences are a hub for weak cryptic 5' and 3' recursive splice sites, and the EJC has a conserved role in silencing such sequences.

Chapter 2

Short cryptic exons mediate recursive splicing in *Drosophila*[1]

Summary

Many long *Drosophila* introns are processed by an unusual recursive strategy. The presence of ~200 adjacent splice acceptor and splice donor sites, termed ratchet points (RPs), were inferred to reflect "zero nucleotide exons" whose sequential processing subdivides removal of long host introns. I used CRISPR-Cas9 to disrupt several intronic RPs in the animal, and some recapitulated characteristic loss-of-function phenotypes. Unexpectedly, selective disruption of RP splice donors revealed constitutive retention of unannotated short exons. Functional minigene tests confirm that unannotated cryptic splice donor sites are critical for recognition of intronic RPs, demonstrating that recursive splicing involves the recognition of cryptic RP-exons. I generalize this mechanism, since canonical, conserved, splice donors are specifically enriched in a +40-80 nt window downstream of known and newly-annotated intronic RPs, and exhibit similar properties to a newly-recognized class of expressed RP-exons. Overall, these studies unify the mechanism of *Drosophila* recursive splicing with that in mammals.

[1] Reprinted from *Joseph, B., Kondo, S.* & Lai, E.C. Short cryptic exons mediate recursive splicing in Drosophila. Nat Struct Mol Biol.* ***25,*** *365–371 (2018).*
*KS generated the *kuz[ΔRP]* and *Bx[ΔRP]* alleles

Introduction

Large introns create challenges for accurate processing, due to seemingly modest information encoded by minimal splice donor (GU) and acceptor (AG) sites. One established concept is the splicing machinery defines the smallest available unit; thus, introns are defined when they are relatively small, but exon definition becomes critical when flanking introns are larger (De Conti et al., 2013). In this way, many cryptic splicing signals within intronic context might be avoided. Still, it is puzzling how proper junctions are decoded as introns increase from tens to hundreds of kilobases, even megabases in mammalian genomes, given that splicing rates of short and very long introns are similar (Singh & Padgett, 2009). Coupling of RNA Polymerase II, chromatin, and factors involved in splice site recognition and spliceosome assembly, may facilitate processing at long introns (Hollander et al., 2016). For example, factors involved in splice site pairing, including U1 snRNP and U2AF65, associate with Pol II. At the same time, exons can preferentially associate with nucleosomes, are marked by distinctive histone modifications, and associate with U2 snRNP. Thus, organization and scaffolding afforded by Pol II and chromatin can aid the specificity and efficiency of splicing across long distances (Hollander et al., 2016).

Another consideration is the process of recursive splicing, whereby splicing of long introns is achieved in stepwise fashion. This breaks up the daunting task of processing a larger intron into several smaller, more manageable segments. The Lopez lab first recognized this mechanism during processing of the 77kb intron of *Drosophila Ultrabithorax* (*Ubx*). This intronic space includes two short cassette exons (mI and mII) that can be present or absent in different isoforms. However, unlike typical alternative splicing reactions, careful analysis showed that processing of these *Ubx* microexons involves splicing that regenerates 5' splice sites at their junctions (Hatton et al., 1998).

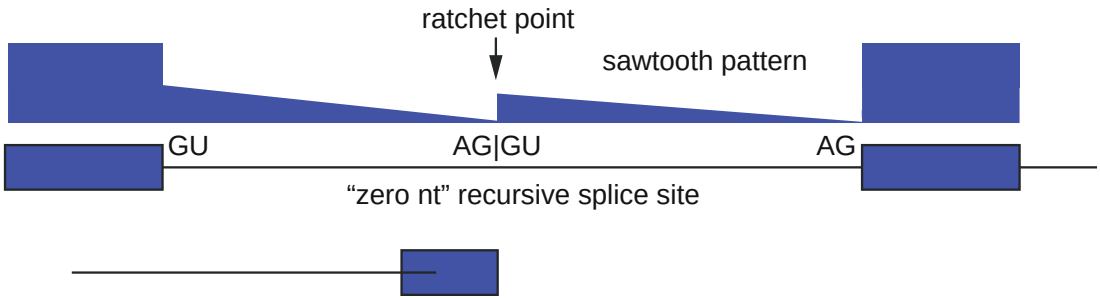
The same *Ubx* intron was subsequently shown to contain a "ratchet point" (RP) without a recognizable microexon; thus, it is marked only by a juxtaposed AG:GU splice acceptor-donor pair (Burnette et al., 2005). This same study predicted 165 candidate RPs within long introns of >100 genes, suggesting that recursive splicing is utilized broadly to process long introns in *Drosophila* (Burnette et al., 2005).

Of note, the vast majority of predicted RPs (155/165) were not associated with known exons and therefore do not appear in mature mRNA; seven novel RPs were validated using rt-PCR assays (Burnette et al., 2005). It would take another decade, until the advent of deep RNA-sequencing surveys, for broad experimental confirmation of recursive splicing. In particular, total RNA-seq data from diverse *Drosophila* stages, tissues and cell types permitted de novo annotation of 197 "zero nucleotide exon" RPs from 130 introns of 115 genes (Duff et al., 2015). Little is known about the recursive splicing mechanism, although the process is believed constitutive and appears especially sensitive to U2AF activity (Duff et al., 2015). How an exon of zero nucleotides would be recognized by the splicing machinery is mysterious, and previous sequence analysis downstream of known intronic ratchet points did not reveal sequence motifs or compelling conserved regions (Duff et al., 2015) .

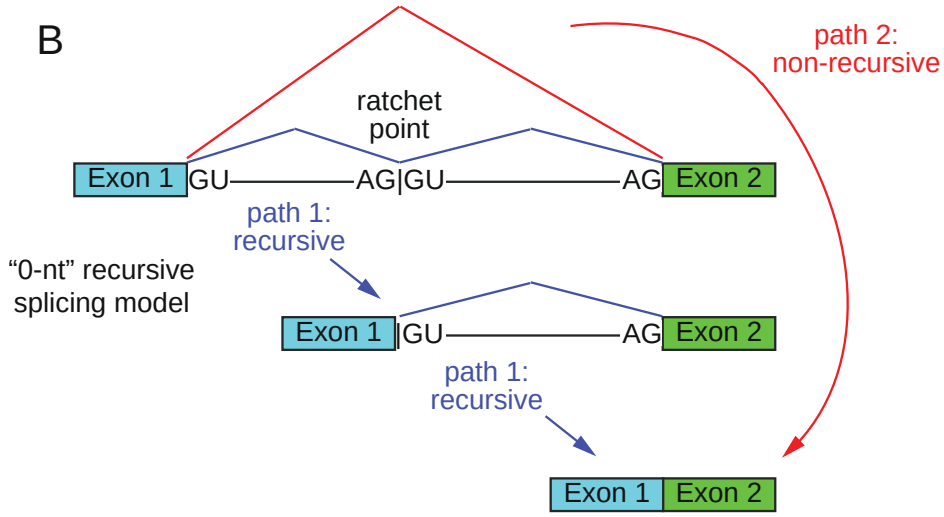
Mammals also utilize recursive splicing, but seem to harbor far fewer intronic recursive splice sites (Duff et al., 2015; Sibley et al., 2015). Notably, while exon signatures were not identified at *Drosophila* "0-nt exon" RPs, functional studies in human cells provided evidence that exon definition via recursive splicing exons ("RS-exons") is critical for recursive splicing (Sibley et al., 2015). Thus, mammalian RS-exon splicing is analogous to the strategy described for the initial *Ubx* microexons (Hatton et al., 1998).

Many fundamental questions regarding recursive splicing remain. For example, it is not resolved whether recursively spliced products represent obligate intermediates on the way to mature mRNAs (**Figure 2.1**). The "sawtooth" pattern of total RNA-seq reads

A Annotation of ratchet points from RNA-seq reads



Models for recursive splicing



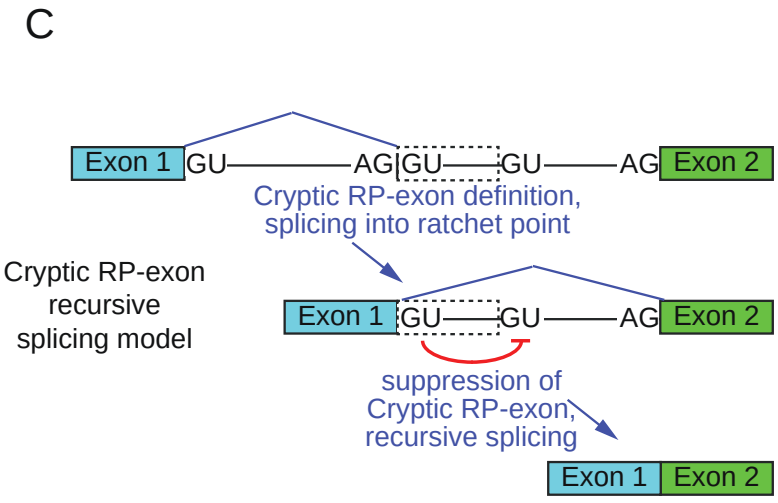


Figure 2.1. Evidence and mechanistic models for recursive splicing. (A) Sawtooth RNA-seq patterns are indicative of recursive splicing intermediates. Generally, RNA-seq coverage in introns is reflective of nascent transcription and resembles right-angled triangles, with highest coverage at the 5' end and lowest at the 3' end of the intron. However, introns that undergo recursive splicing consist of multiple intronic segments, each with its own right-angled triangle coverage, producing a sawtooth pattern. This property has been exploited to infer recursive splicing and annotate RPs. (B) Models for processing introns with RPs. It is conceivable that introns that contain RPs will be processed in one of two ways. First, the RP is utilized constitutively (path 1) and the intron is removed in two sequential steps. Second, the RP is skipped such that the entire intron is spliced out in one step (path 2). (C) A molecular model for recursive splicing. I propose that recursive splicing proceeds by first defining a cryptic RP-exon, which is specified by the RP splice acceptor and a downstream cryptic splice donor. Definition of the cryptic RP-exon allows removal of the first intron segment and production of the recursive intermediate. In the second splicing reaction, we propose that the regenerated RP splice donor outcompetes the cryptic splice donor, thereby removing the whole intron and ligating neighboring exons.

at certain loci, which dips characteristically at RP sites, is consistent with these being co-transcriptional splicing intermediates (Duff et al., 2015) (**Figure 2.1**). However, it is possible that recursive splicing is a pathway parallel to non-recursive splicing. The residence of RPs in the very longest transcription units does not render them amenable to direct mechanistic observation using *in vitro* splicing assays, and recursive splicing has otherwise been studied only with minigenes. Other basic questions include what effect recursive splicing has on gene expression, whether recursive splicing matters *in vivo*, and whether there truly are fundamental differences in recursive splicing mechanism between flies and mammals, as suggested by available literature (Cook-Andersen & Wilkinson, 2015).

In this study, I use molecular genetic information from CRISPR engineering to reveal that recursive splicing in *Drosophila* proceeds via unannotated cryptic RP-exons. I validate this model using functional tests, and extend it genomewide, by showing that unannotated, high-scoring, conserved splice donors are present in a distinctive length window downstream of known and novel intronic RPs. Recursive splicing is utilized on a continuum, since beyond the hundreds of splicing events inferred to involve cryptic RP-exons, I also recognize scores of expressed RP-exons. These findings now unify the mechanism of recursive splicing in flies with mammals.

Results

In vivo mutagenesis of ratchet points in *Drosophila*

To my knowledge, no studies of recursive splicing have yet involved endogenous sites in intact animals (Burnette et al., 2005; Duff et al., 2015; Hatton et al., 1998; Kelly et al., 2015; Sibley et al., 2015). Therefore, I exploited CRISPR-Cas9 to mutagenize ratchet points (RPs) in *Drosophila* and assess consequences on phenotypes and RNA processing. I successfully targeted intronic RPs in *Beadex* (*Bx*), *Ultrabithorax* (*Ubx*) and

kuzbanian (*kuz*); the former lies between non-coding exons while the latter two reside between coding exons. In all three cases, I characterized alleles that selectively disrupted RP splice donor sites (**Figure 2.2A**, molecular details provided in **Figure 2.3**). While the *Bx[RP]* mutant was viable and lacked overt defects, the *Ubx[RP]* and *kuz[RP]* mutants were lethal. My RP mutants failed to complement known amorphic *Ubx* (Bender et al., 1983) and *kuz* (Fambrough et al., 1996) mutants, and I proceeded to detailed phenotypic analyses.

Ubx[RP]/+ animals exhibit mild haltere enlargement, consistent with partial conversion to wing identity (**Figure 2.2B-C**). As the dominant *Ubx* effect can be difficult to visualize, I sensitized the background using the *Ubx* hypomorphic allele on the TM3 balancer. Strikingly, viable *Ubx[RP]/TM3*, *Ubx[bx-34e]* animals exhibit overt transformation of halteres into wings equivalent in severity to amorphic *Ubx[1]* in trans to TM3 (**Figure 2.2D-E**). Moreover, immunostaining of larval CNS showed the normal pattern of Ubx protein in wildtype ventral nerve cord was basically undetectable in lethal *Ubx[RP]* homozygotes (**Figure 2.2F-G**). Thus, *Ubx[RP]* abrogates *Ubx* function and protein accumulation.

With *kuz[RP]*, homozygous mutant embryos phenocopied previously described zygotic defects in CNS axonal patterning (Fambrough et al., 1996). For example, BP102 staining showed reduction of longitudinal bundles and accumulation of commissural material in *kuz[RP]* homozygotes compared to controls (**Figure 2.2H-I**). Fasciclin II (Fas II) staining also recapitulated known *kuz* defects. *kuz[RP]/+* heterozygotes exhibit characteristic Fas II patterns of three longitudinal axonal tracts on either side of the midline (**Figure 2.2J**), while amorphic *kuz[e29-4]* homozygotes fail to elaborate the longitudinal tracts and present midline crossing defects (**Figure 2.2K**). I find similar phenotypes for *kuz[RP]* homozygotes (**Figure 2.2L**) and *kuz[RP]/[e29-4]* trans-heterozygotes (**Figure 2.2M**), indicating *kuz[RP]* is a strong loss-of-function allele.

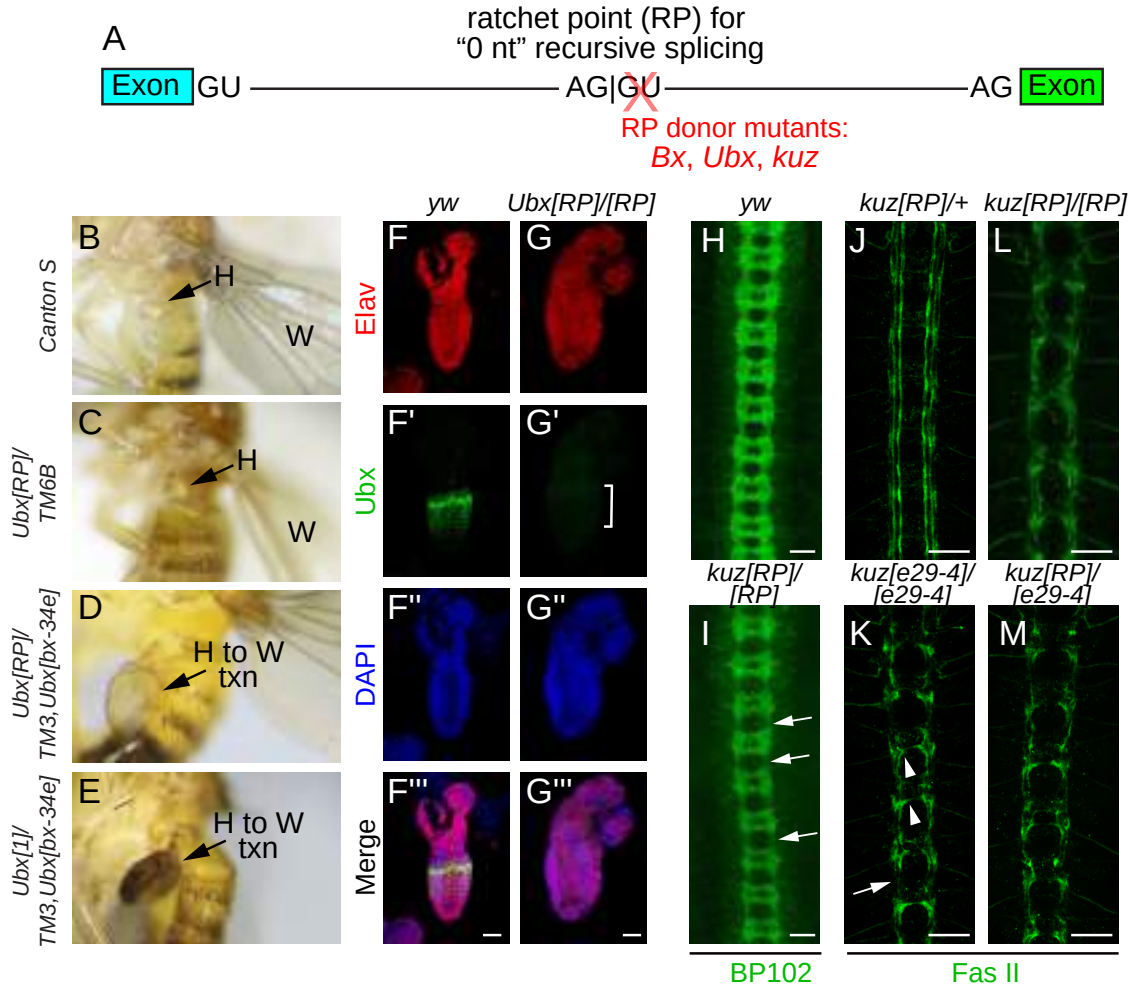
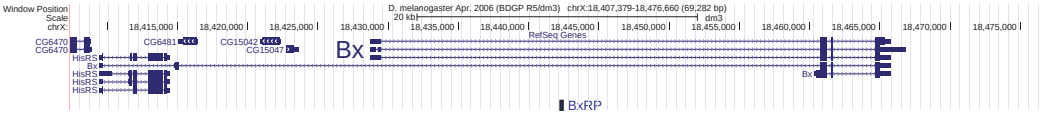

A
ratchet point (RP) for
“0 nt” recursive splicing
Exon
GU
AG|GU
AG
Exon
RP donor mutants:
Bx, Ubx, kuz
yw
Ubx[RP]/[RP]
yw
kuz[RP]/+
kuz[RP]/[RP]
Canton S
Ubx[RP]/ TM6B
Ubx[RP]/ TM3,Ubx[bx-34e]
Ubx[1]/ TM3,Ubx[bx-34e]
Elav
Ubx
DAPI
Merge
H
W
H to W txn
kuz[RP]/ [RP]
kuz[e29-4]/ [e29-4]
kuz[RP]/ [e29-4]
BP102
Fas II

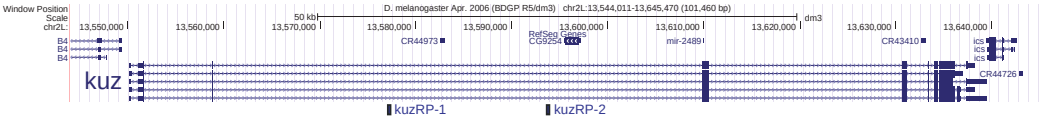
Figure 2.2. Ratchet point donor mutants of *Ubx* and *kuz* are strong loss-of-function alleles. (A) Recursive splicing at long introns sequentially removes intron segments at paired splice acceptor-donor sites (AG:GU), also known as a ratchet point (RP), without leaving behind a mature exon. I used CRISPR-Cas9 to selectively mutagenize several *Drosophila* RP donor sites. (B-E) Lateral images of adult flies that illustrate phenotypes of *Ubx[RP]* mutants. (B) Wildtype (Canton S) fly with wing (W) and haltere (H) labeled. (C) *Ubx[RP]* heterozygote (in trans to TM6B balancer) shows mild enlargement of the haltere, indicative of *Ubx* haploinsufficiency. (D) *Ubx[RP]/[bx-34e]* (in trans to TM3 balancer) shows an overt haltere-to-wing transformation (H to W txn). (E) The phenotype of the RP mutant is similar to the known amorphic allele *Ubx[1]* over TM3. (F-G) Immunostaining of first instar larval CNS. (F) Control *yw* shows the normal segmental pattern of Ubx protein (green) in the ventral nerve cord, counterstained with pan-neuronal Elav (red) and DAPI (blue). (G) *Ubx[RP]* homozygote selectively lacks Ubx protein. (H-M) Ventral images of stage 16 embryos stained with α-BP102 (H, I) or α-Fas II (J, M) to reveal all CNS axons or subsets of ipsilateral axons, respectively. (H) BP102 exhibits a characteristic ladder-like staining pattern in control (yw) embryo. (I). *kuz[RP]* homozygote display thickening of the commissures and thinning of longitudinal connectives (arrows). (J) *kuz[RP]/+* heterozygote exhibits a normal Fas II pattern of three bundles of longitudinal axons on either side of the ventral midline. All three *kuz* mutant combinations, *kuz[e29-4]* homozygotes (K), *kuz[RP]* homozygotes (L) and *kuz[RP]/[e29-4]* trans-heterozygotes (M) exhibit similar Fas II defects. These include failure to establish the longitudinal tracts and midline crossing defects. Scale bars in F-G indicate 40 μm and in H-M indicate 20 μm.

A

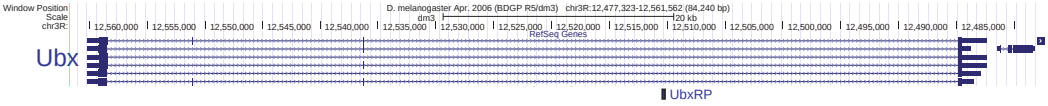
Beadex (Bx)



kuzbanian (kuz)



Ultrabithorax (Ubx)



B

kuz[RP] - 6 nt deletion

WT: GAGCAGCAGACAATGGCATAATAAACCATAATCAAATATATTGTAATGTTTATTTTTCGTCTTCTCTTTAC**AGGTGAGT**GCTCGGTTTCTAA
MT: GAGCAGCAGACAATGGCATAATAAACCATAATCAAATATATTGTAATGTTTATTTTTCGTCTTCTCTTTACAG------GCTCGGTTTCTAA

Bx[RP] – 1 nt substitution, 2 nt insertion

WT: AATACCTTTCTTGTTTTCCTTGTTTTTCC**AGGT--AAGT**GTCAACACCCACCCAATTGCTACAACACACAAGAT
MT: AATACCTTTCTTGTTTTCCTTGTTTTTCCAGGAAAAAGTGTCAACACCCACCCAATTGCTACAACACACAAGAT

Ubx[RP] – 38 nt insertion

WT: TCAAACTATTTTCTTCTCTTTTCT**AG**----------**GTAAGT**GTCAAATATTTAATACACCC
MT TCAAACTATTTTCTTCTCTTTTCTAGAATTCTGTCAAATATTTAATAACCCTTAAACCAAACAGGTAAGTGTCAAATATTTAATACACCC

C

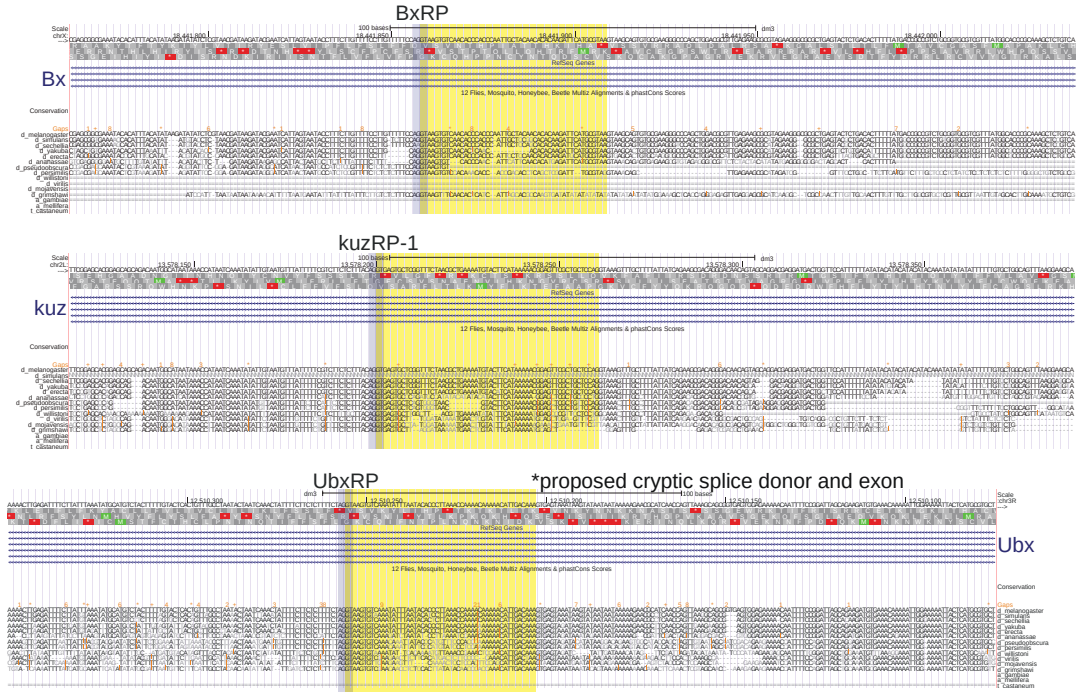


Figure 2.3. Genes with RPs were manipulated to identify cryptic exons. (A) UCSC genome browser screenshots display the three genes (*Bx*, *kuz,* and *Ubx)* manipulated in this study, including the approximate genomic locations of RPs within long host introns. (B) Nature of mutant alleles along with sequence alignments. (C) UCSC genome browser nucleotide-level screenshots of mutated RPs (grey highlight) along with cryptic exons detected in mutants (yellow highlight). **Ubx[RP]* is an insertion mutant, which separates the RP splice acceptor and donor sites by 38nt.This 38nt insertion is retained in mutant animals. However, bioinformatic analysis has identified a naturally occurring cryptic exon and splice donor as indicated.

Altogether, these tests provide first evidence that altering recursive splicing in the animal can disrupt endogenous gene function and generate mutant phenotypes.

Molecular analysis of RP mutants reveals constitutive retention of cryptic exons

I analyzed molecular consequences of RP mutations on RNA processing. Of note, since previous mutational tests of recursive splicing were done with minigenes, it has been difficult to ascertain if this process generates obligate intermediates towards mRNA. Alternatively, there could be parallel processing pathways that skip recursive sites, and/or recursive splicing might theoretically generate dead-end products (**Figure 2.1**).

rt-PCR analyses to detect an intermediate amplicon downstream of the ratchet point (**Figure 2.4A**) yielded specific products from each RP mutant (**Figure 2.4B**), indicating successful splicing into ratchet points. Moreover, I detected mature mRNA products from all three RP mutants. However, mature mRNA amplicons from RP mutants were longer in all three cases (**Figure 2.4C**). Interestingly, sequencing of these products showed mutant transcripts retained sequences that originate from immediately 3' of the RP splice acceptor, and are spliced to cognate downstream exons through cryptic splice donor sites (**Figure 2.4D** and **Figure 2.3**). Since the ectopic exon in *Bx* resides in its 5' UTR, this does not affect protein output. However, the inclusion of novel exons in *kuz* and *Ubx* disrupts their reading frames. Of note, rt-PCR tests exclusively detected RP-mutant transcripts bearing the ectopic exons, whether coding or non-coding. Thus, splicing into these RP sites is constitutive, and they are obligate *in vivo* intermediates toward mature mRNAs.

Recursive splicing is mediated by short cryptic exons

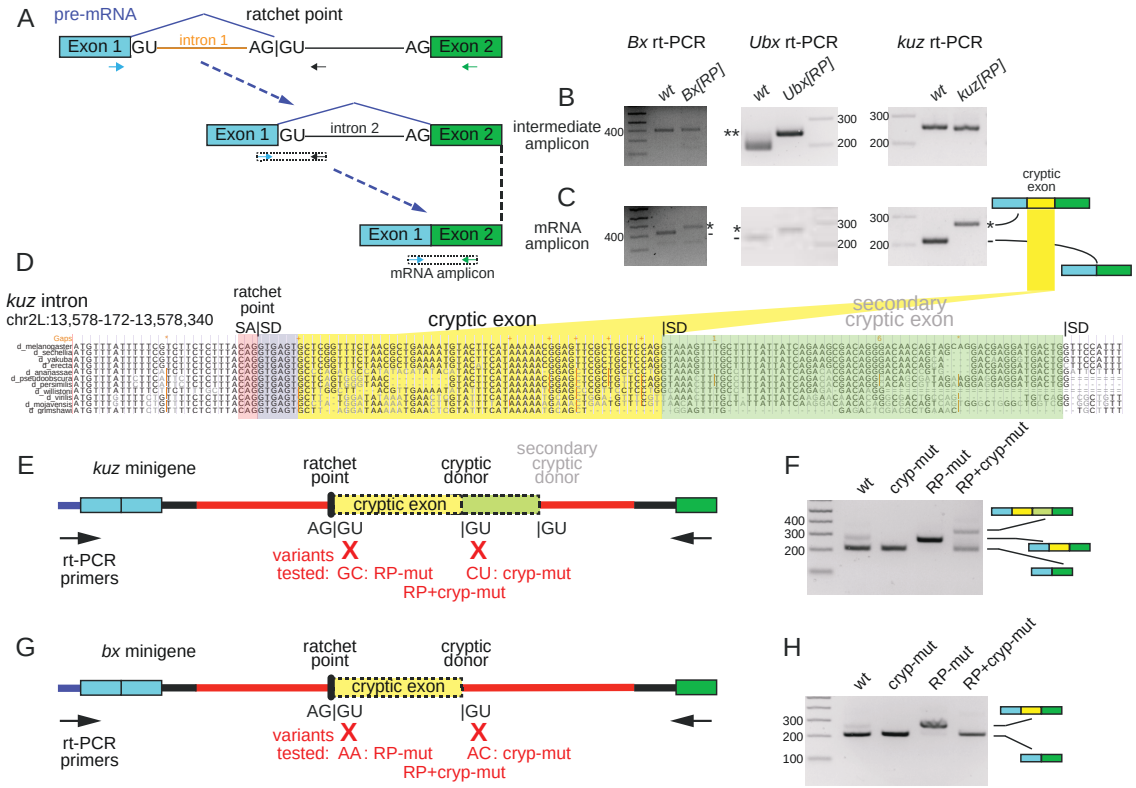

A
pre-mRNA
ratchet point
Exon 1
GU
intron 1
AG|GU
AG
Exon 2
intron 2
mRNA amplicon
B
Bx rt-PCR
Ubx rt-PCR
kuz rt-PCR
wt
Bx[RP]
Ubx[RP]
kuz[RP]
intermediate amplicon
400
300
200
C
mRNA amplicon
cryptic exon
D
kuz intron
chr2L:13,578-172-13,578,340
ratchet point
SA|SD
cryptic exon
|SD
secondary cryptic exon
E
kuz minigene
ratchet point
cryptic donor
secondary cryptic donor
AG|GU
|GU
variants tested: GC: RP-mut
CU: cryp-mut
RP+cryp-mut
rt-PCR primers
F
wt
cryp-mut
RP-mut
RP+cryp-mut
G
bx minigene
variants tested: AA: RP-mut
AC: cryp-mut
H

Figure 2.4. Molecular evaluation of RP mutants reveals existence of cryptic exons. (A) Schematic of recursive splicing depicting removal of an intron in two steps. In the first step, a portion of the intron is removed (orange) resulting in intermediate pre-mRNA with a regenerated 5' splice site. In the second step, the remainder of the intron (black) is removed, producing mRNA. Arrows are used to display primers to specifically amplify intermediate and mRNA transcripts. (B) Both wildtype and homozygous RP mutant animals produced intermediate amplicons, indicating that RP donor mutations did not disrupt recursive splicing. ******_Ubx[RP]_ mutants have a 38nt insertion that disrupts the RP and separates SA and SD by 38nt; thus, lengthening the intermediate amplicon. (C) Compared to wildtype, RP donor mutants consistently had larger mRNA amplicons that contained cryptic exon retention. (D) UCSC genome browser screenshot of the first ratchet point in *kuz* (*kuz-RP1*, with ratchet point splice acceptor [SA] in red and splice donor [SD] in purple; the SD was disrupted in *kuz[RP]* allele), along with highlighted regions showing cryptic exon (yellow) and a secondary cryptic exon (green) revealed in mutagenesis experiments. (E, G) Schematics of minigene constructs. The ~2.5 kb intronic RP locus used in the *kuz* (E) and the *Bx* (G) minigene is shown as a red line. Variants tested and primers used for rt-PCR are as indicated. (F, H) rt-PCR was used to evaluate spliced products from minigenes. (F) S2-R^+ cells were transfected with WT *kuz* minigene (wt), or variants containing mutations in cryptic splice donor (cryp-mut), RP-splice donor (RP-mut), or both (RP+cryp-mut). (H) S2-R^+ cells were transfected with WT *Bx* minigene and an analogous set of variants. In both cases, mutation of RP donor sites resulted in cryptic exon retention, while mutation of both RP+cryp donor sites lowered the efficiency of mature splicing. With *kuz*, the RP+cryp mut construct reveals usage of an extended cryptic exon.

The complete retention and characteristic size of retained cryptic exons implied their involvement in recursive splicing. In particular, I hypothesized that intronic ratchet points do not represent "0-nt exons" as originally suggested (Duff et al., 2015), but actually proceed by an exon definition strategy involving unannotated cryptic exons, whose inclusion is subsequently suppressed. In this model, the strategy of *Drosophila* recursive splicing might resemble that of *Ubx* mini-exons (mI and mII) and mammalian recursive splicing (Sibley et al., 2015) (**Figure 2.1C**). In some cases the putative splice donor site of the cryptic exon is conserved, as with *Ubx-RP*, but the putative splice donors of *Bx-RP* and *kuz-RP* cryptic exons are less conserved; none of these cryptic exons reflect coding constraint (**Figure 2.3C**). By contrast, their companion RP sequences are perfectly constrained in the most distantly aligned Drosophilid genomes. If cryptic exons are integral to the recursive splicing reaction, they would represent an unusual case of RNA processing in which some cis signals are better conserved than others.

I utilized minigenes to test *kuz-RP* and *Bx-RP* intronic processing in S2-R+ cells, along with mutants in RP splice donors, cryptic splice donors, or both (**Figure 2.4E**, **G**). Both *kuz* and *Bx* wildtype minigenes produced expected spliced mRNA products (**Figure 2.4F**, **H**, wt lanes). Unexpectedly, a second band was observed for wildtype *kuz* and to a lesser extent, for wildtype *Bx* and confirmed by sequencing to be spliced mRNA that included the cryptic exon. This further suggests the presence of unannotated cryptic exons at intronic RPs. Indeed, total RNA-seq shows that the cryptic *kuz* exon is partially retained in S2-R^+ but not in tissues (**Figure 2.5**).

Consistent with my animal mutants, mutation of RP donor sites caused constitutive inclusion of cryptic exons from *kuz-RP* and *Bx-RP* minigenes (**Figure 2.4F**, **H**, RP-mut lanes, quantified in **Figure 2.6**). Disruption of cryptic splice donors resulted in exclusion of cryptic exons for ectopic *kuz* and for ectopic *Bx* (**Figure 2.4F**, **H**, cryp-mut

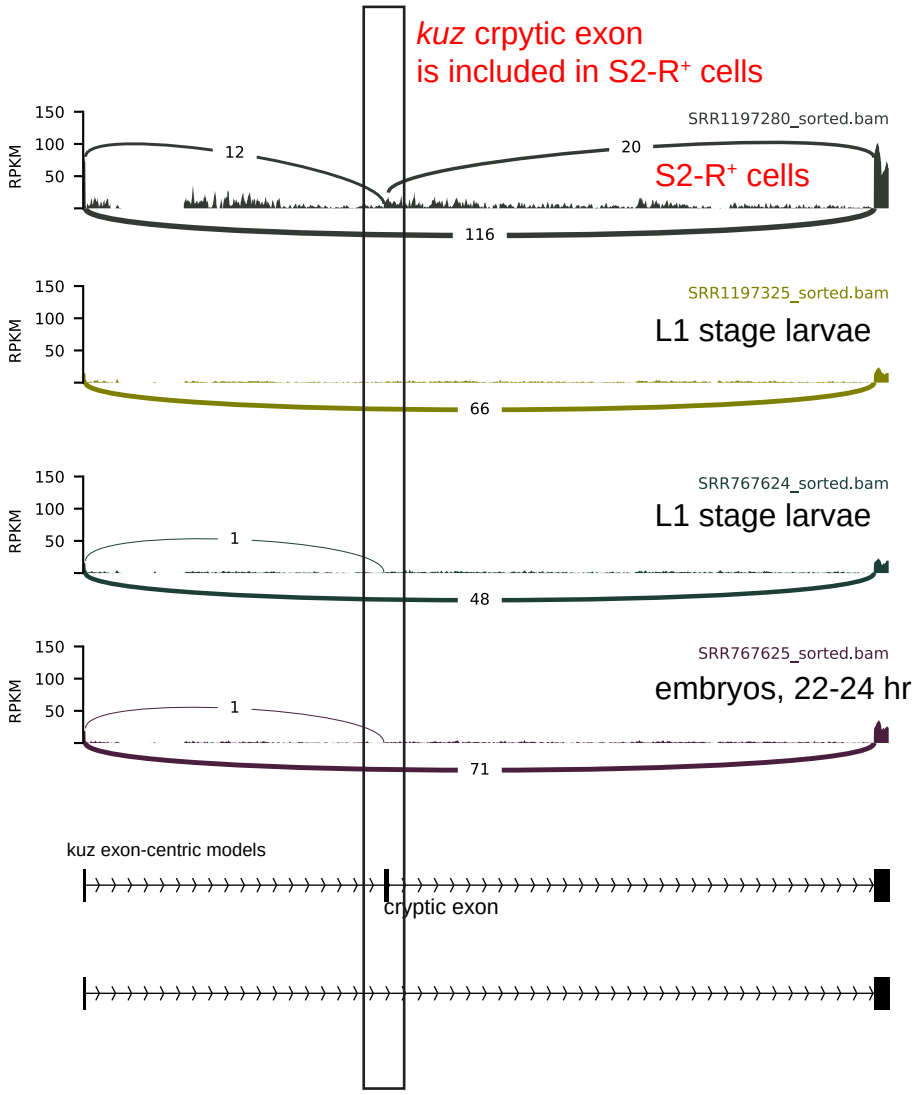


Figure 2.5. *kuz* cryptic exon is retained in S2-R+ cells. Sashimi plots were used to display the usage of the cryptic exon in modENCODE total RNA-seq data from S2-R+ cells, L1 stage larvae and 22-24hr embryos. Spliced reads can only be detected into and out of the cryptic exons in S2-R+ cells, suggesting selective inclusion and/or stabilization here.

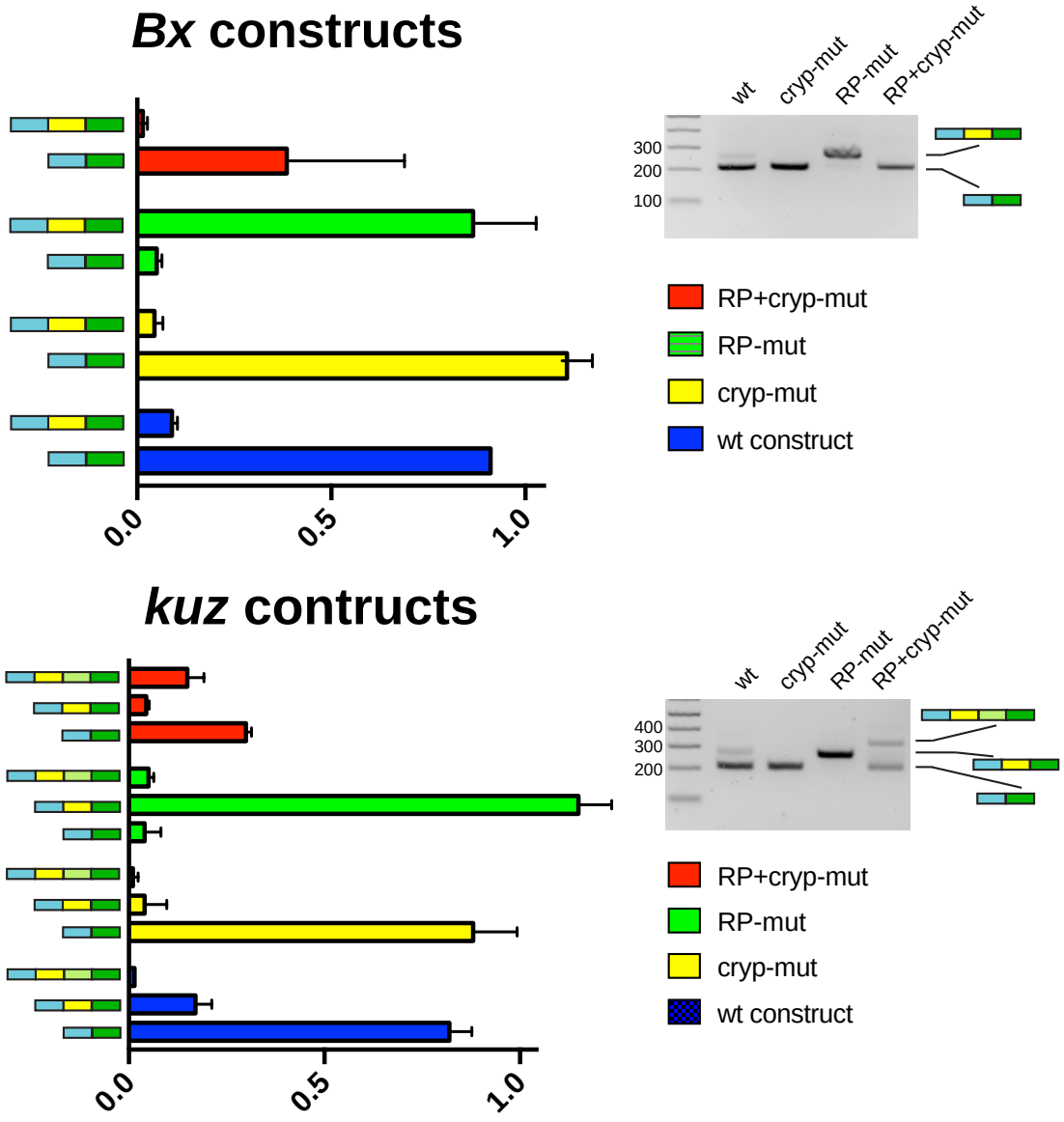


Figure 2.6. Quantification of relative cryptic exon inclusion ratios from wildtype and mutant recursive splicing minigenes. *Bx* and *kuz* minigenes that contained the indicated mutations (see Fig 2.3E, G) were transfected into S2 cells and subjected to rt-PCR analysis. Relative exon inclusion was calculated by normalizing the intensity of the mRNA band to all indicated bands in the same lane, and then scaled to total expression observed in wt lane. Representative gels are shown (see also main Figure 2.2).

lanes). In these tests, cryptic exon skipping, whether through recursive splicing or loss of cryptic exon definition, would yield similar spliced products (see **Figure 2.1B**: path 1 vs. path 2). To distinguish these possibilities, I examined constructs with both cryptic and RP splice donors mutated. If the reaction proceeded via recursive splicing, spliced mRNAs would include cryptic exons, whereas if recursive splicing were abolished due to loss of exon definition, spliced mRNAs would exclude cryptic exons.

The double *Bx* mutant predominantly yielded one spliced product that lacked the cryptic exon (**Figure 2.4H**, RP+cryp-mut lane). The *kuz* double mutant yielded two products – without cryptic exon and with a longer exon (**Figure 2.4F**, RP+cryp-mut lane). Sequencing revealed that retention of an extended cryptic exon in double mutants was due to usage of a downstream, poorly-conserved, secondary cryptic splice donor (**Figure 2.4D**). Cryptic exon skipping in *Bx* and *kuz* suggests loss of recursive splicing due to loss of exon definition, and the lower levels of spliced products in RP+cryp donor mutants indicates that recursive splicing contributes to effective processing. Moreover, activation of a novel cryptic exon in *kuz* double mutants suggests that fortuitous splice elements can easily compensate for disruptions in normally-recognized cryptic splice donor sequences. I return to the latter point in evolutionary analysis. Together, these data provide evidence that exon definition is an important step during recursive splicing.

Genomewide re-annotation of *Drosophila* intronic RPs and RP-exons

I sought to generalize the cryptic RP-exon model for intronic "0-nt" recursive splicing. Before doing so, I aimed to expand the catalog of *Drosophila* ratchet points. Recent efforts used ~11 billion paired end reads from ~100 *Drosophila* stages, tissues, and cell lines to annotate 197 intronic RPs from 130 introns of 115 genes (Duff et al., 2015). However, as these are transient intermediates of co-transcriptionally processed RNA, total RNA-seq data are not optimal for their detection.

I collected data representing actively transcribed RNA (chromatin RNA-seq, nascent RNA-seq, GRO-seq) from S2 cells, embryos, heads and ovaries (**Table 2.1**), and observed enrichment for intronic coverage and junction-spanning reads at known RPs, compared to total RNA and mRNA datasets (**Figure 2.7A-B**). I then developed a pipeline to annotate recursive splicing events (**Figure 2.8**). As before (Duff et al., 2015), I focused on intronic junction spanning reads with tandem splice acceptor and donor motifs at the 3' end, and emphasized loci with sawtooth RNA-seq patterns. I observed strong enrichment of minimal paired splice acceptor and donor motifs (AG:GT) only when the junctions were located within introns >5 kb (**Figure 2.7C**), confirming my pipeline identified genuine splicing events and validating my decision to triage other types of split reads with non-canonical junctions. However, bearing in mind that recursive intermediates are transient, I relaxed the requirement for sawtooth RNA-seq patterns if candidate RPs exhibited strong splice sites (see Methods). I manually examined all candidates to filter potential false positives.

Although I only analyzed 4 tissue types and many-fold fewer mapped RNA-seq reads than previously (Duff et al., 2015), I substantially increase the scope of recursive splicing. My pipeline recovered 187/197 previously annotated RPs (Duff et al., 2015), and in total identified 304 unique RPs in 188 introns of 169 genes (**Figure 2.9A**). The newly recognized RPs share sequence and evolutionary properties of known recursive sites, as shown in partitioned analyses (**Figure 2.10**). Notably, the 93 novel RPs with sawtooth RNA-seq evidence exhibit comparable phyloP conservation scores to known RPs, while 24 novel RPs lacking overt sawtooth RNA-seq patterns were only moderately less constrained (**Figure 2.10A**). Overall, intronic RPs are (1) well-conserved across the Drosophilid phylogeny (**Figure 2.9B**), (2) share consensus splice motif characteristics including strong polypyrimidine tracts (**Figure 2.9C**), and (3) preferentially reside within especially long host introns (**Figure 2.10B**).

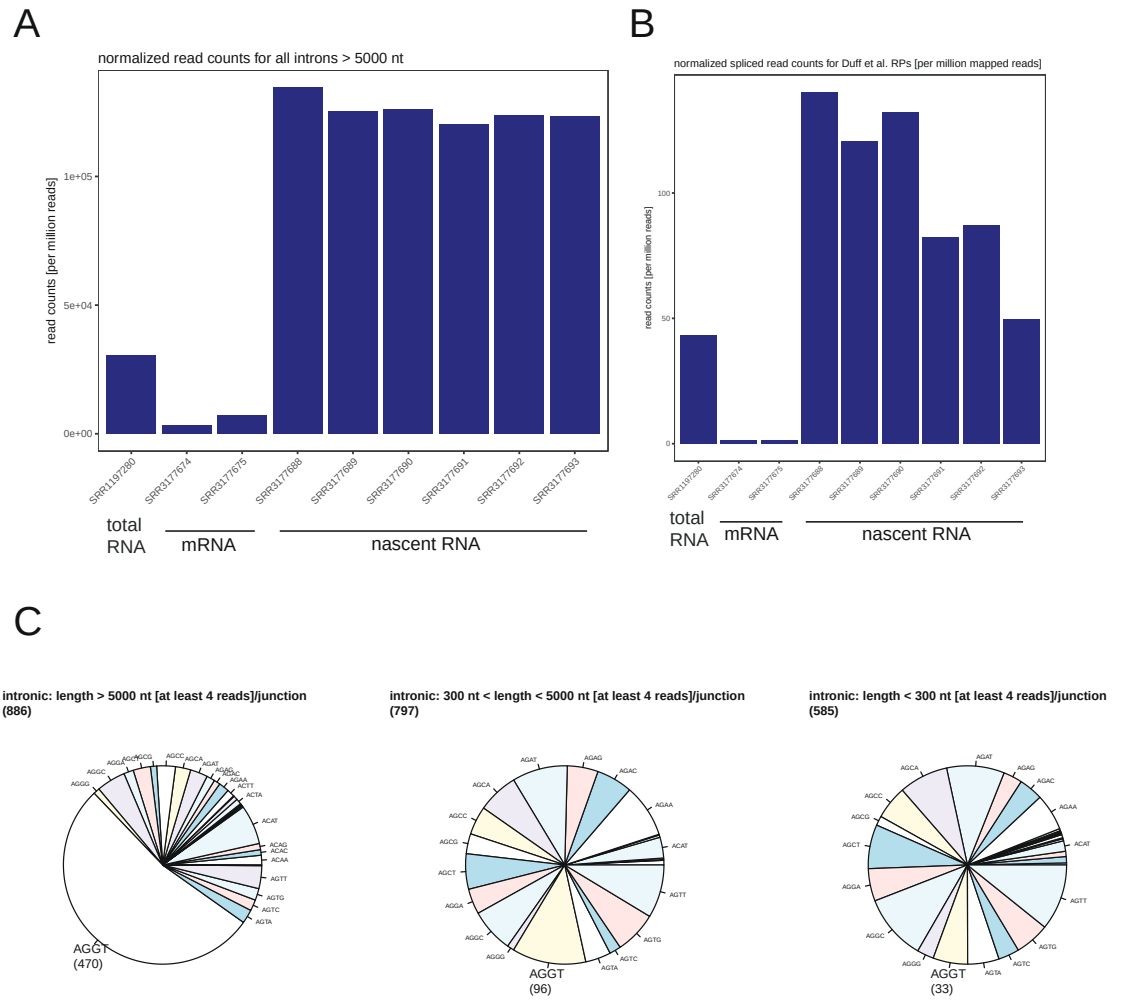


Figure 2.7. Nascent RNA-seq datasets are more suited for detection of RPs. (A, B) Nascent RNA-, total RNA- and mRNA-seq datasets from S2 cells were evaluated for different criteria. (A) Nascent RNA datasets have higher coverage at intronic loci. Long intronic segments - with no overlapping genes - were identified and reads mapping to these regions were summed and normalized. (B) Junction spanning reads that mapped to RPs identified in Duff et al. were summed and normalized. (C) Junction spanning reads found from all nascent RNA-seq and GRO-seq were merged and those with 3' ends mapping to intronic regions were stratified by junction split (intron length) into three categories. For each category, pie charts were drawn to indicate tetranucleotide distributions at the 3' junction end. Note that AGGT, which resembles minimal ratchet point sequences are enriched only within the long intron category.

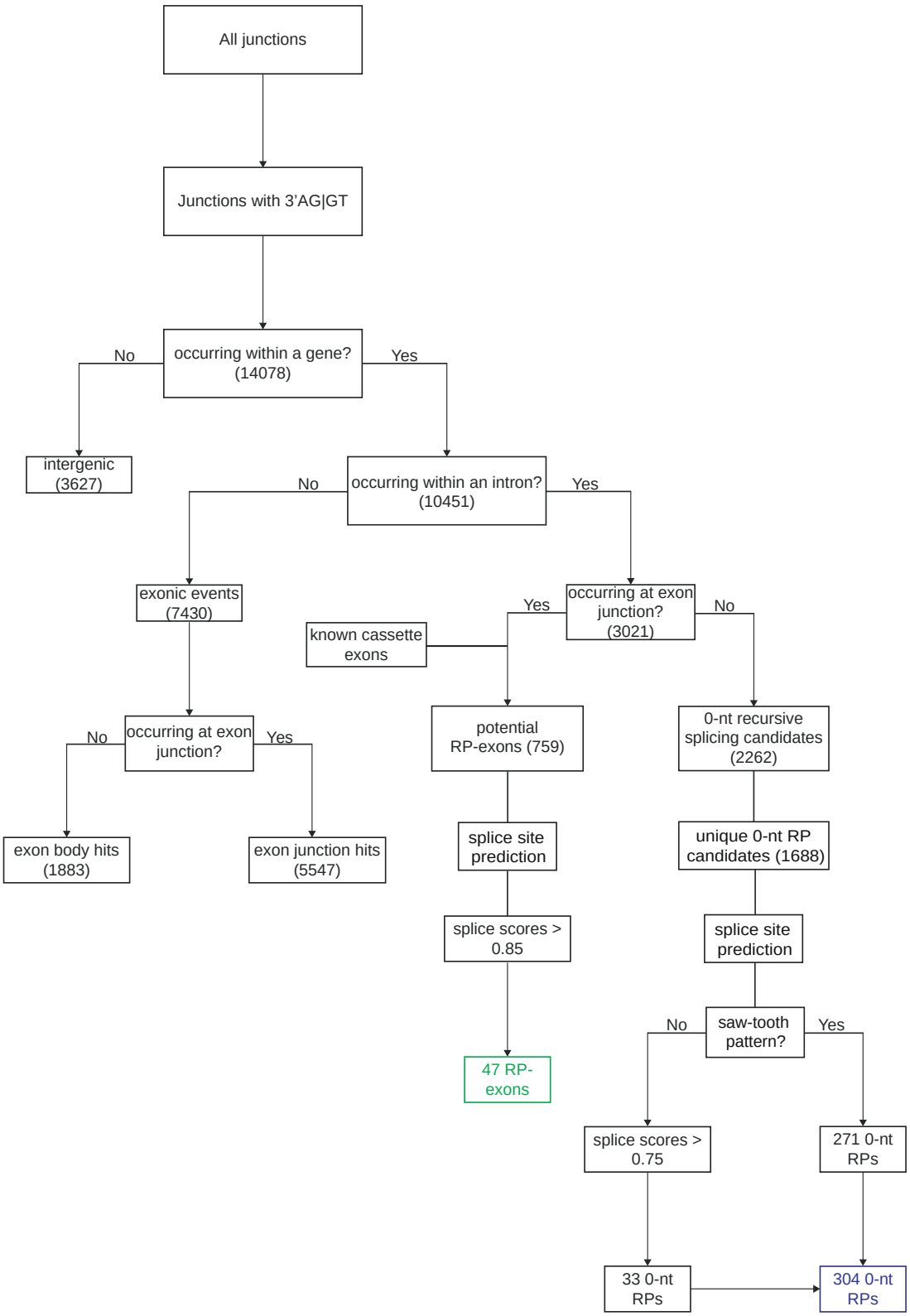

All junctions
Junctions with 3'AG|GT
occurring within a gene? (14078)
No
Yes
intergenic (3627)
occurring within an intron? (10451)
No
Yes
exonic events (7430)
occurring at exon junction? (3021)
Yes
No
known cassette exons
occurring at exon junction?
No
Yes
potential RP-exons (759)
0-nt recursive splicing candidates (2262)
exon body hits (1883)
exon junction hits (5547)
splice site prediction
unique 0-nt RP candidates (1688)
splice scores > 0.85
splice site prediction
No
saw-tooth pattern?
Yes
47 RP-exons
splice scores > 0.75
271 0-nt RPs
33 0-nt RPs
304 0-nt RPs

Figure2.8. Pipeline to annotate novel intronic RPs and RP-exons. Flowchart displays criteria used to classify junction spanning reads into categories such as exon body hits, exon junction hits, intergenic, and candidate RP classes: potential RP-exons and 0nt recursive splicing candidates. Following initial characterization, RP-exon and 0-nt exon candidates were further vetted as indicated by requiring strict splice scores, as quantified by NNSPLICE and manually checking for sawtooth pattern in RNA-seq data.

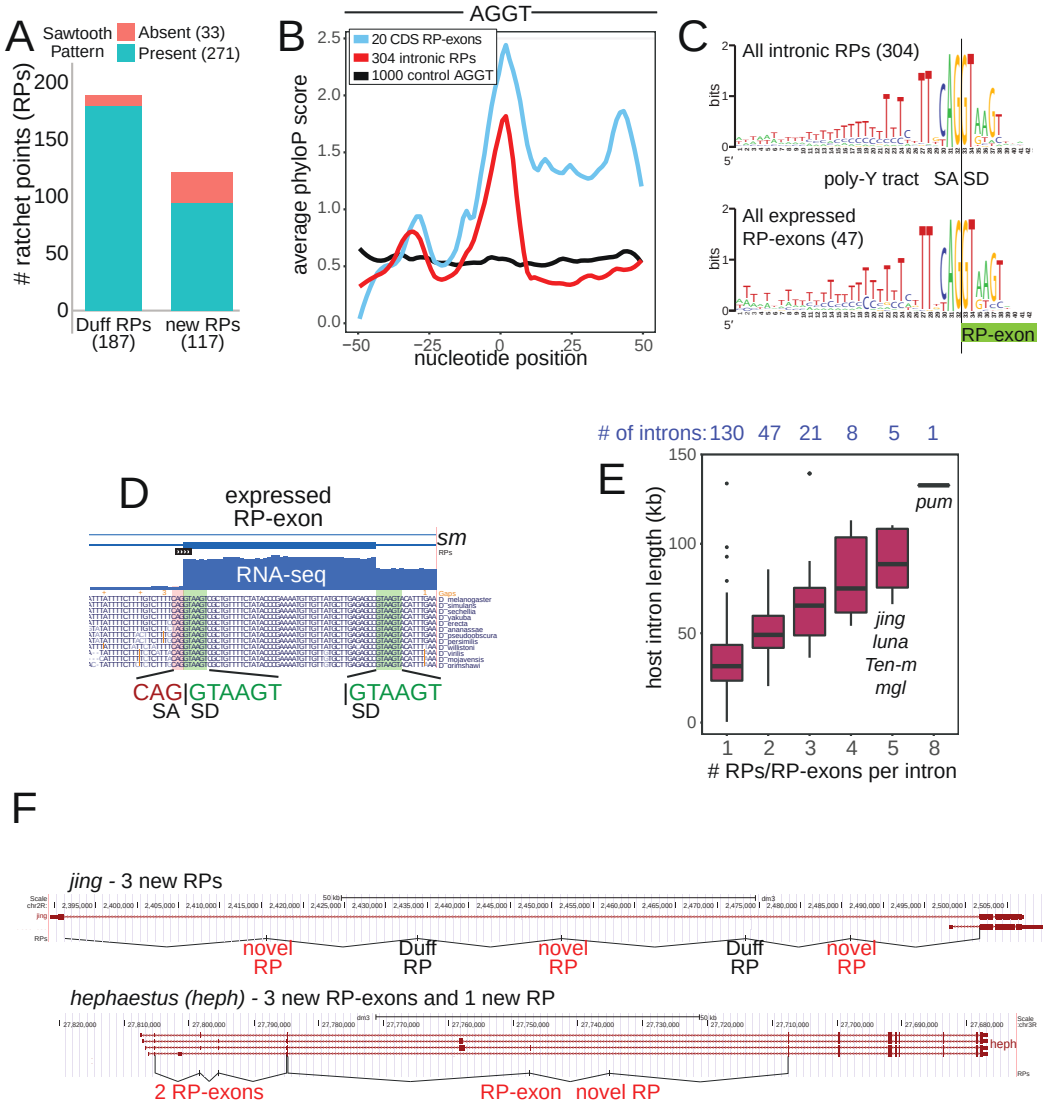
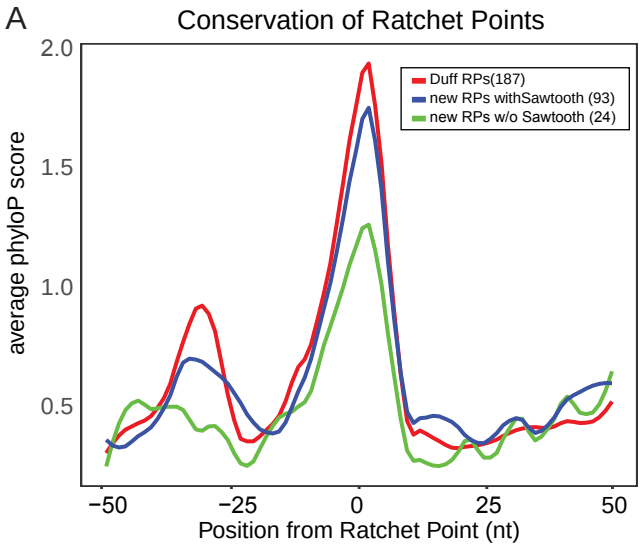

A
Sawtooth Pattern
Absent (33)
Present (271)
ratchet points (RPs)
Duff RPs (187)
new RPs (117)
B
AGGT
20 CDS RP-exons
304 intronic RPs
1000 control AGGT
average phyloP score
nucleotide position
C
All intronic RPs (304)
bits
poly-Y tract SA SD
All expressed RP-exons (47)
RP-exon
D
expressed RP-exon
sm
RNA-seq
CAG|GTAAGT
SA SD
|GTAAGT
SD
E
# of introns: 130 47 21 8 5 1
host intron length (kb)
pum
jing
luna
Ten-m
mgl
RPs/RP-exons per intron
F
jing - 3 new RPs
novel RP
Duff RP
novel RP
Duff RP
novel RP
hephaestus (heph) - 3 new RP-exons and 1 new RP
heph
2 RP-exons
RP-exon
novel RP

Figure 2.9. Genomewide annotation of novel intronic RPs and RP-exons.

(A) Summary of intronic RP annotations (i.e., with no evidence for an expressed exon) made in this study using nascent RNA-seq datasets, 187 of which were previously reported[6] and 117 of which are novel. Presence of sawtooth RNA-seq patterns are noted. (B) Evolutionary conservation for all intronic RPs, RP-exons and control intronic AGGT sites. This was evaluated by averaging phyloP scores at each nucleotide position about the RP. (C) Sequence logos for the aggregate collections of intronic RPs and RP-exon splice junctions. (D) Example of RP-exon in the *sm* gene. It contains a perfectly conserved canonical splice donor (GTAAGT) at the beginning of an expressed cassette exon, which also uses a conserved GTAAGT splice donor. (E) Comparison of intron lengths and number of RPs per intron. Boxes represent the interquartile ranges and n for each group is indicated above plot. (F) Examples of novel intronic RPs and expressed RP-exons identified in *jing* and *hephaestus (heph).*

A



B

	mean length (nt)
Duff et al (187 sites)	~56000
This study (117 sites)	~53500

C

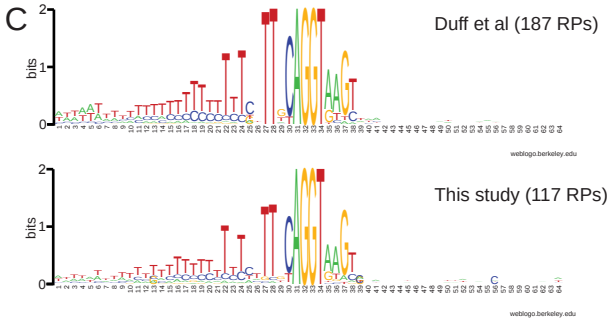


Figure 2.10. Novel ratchet points share sequence, structural, and evolutionary properties of known ratchet points. (A) Comparison of average phyloP scores for “0-nt” RPs. New RPs were grouped into two categories based on whether they had sawtooth patterns in RNA-seq data or not. (B) Average length of host introns for RPs found in Duff et all and this study. (C) Sequence logos for categorized RPs.

The first characterized cases of recursive splicing, *Ubx* microexons mI and mII (Hatton et al., 1998), resemble cryptic exon retention in *[RP]* mutant animals. Moreover, although usage of the *kuz* cryptic exon results in a nonsense product, I detect endogenous inclusion in S2-R^+ (**Figure 2.5**). This led us to suspect there might be a larger class of expressed, recursively-spliced exons beyond *Ubx*, which would exist on a continuum of alternative splicing in *Drosophila* with intronic RPs that putatively proceed via cryptic exons. For example, *msi* is a gene where I detect an intronic RP with sawtooth RNA-seq pattern and a cassette exon that regenerates a 5' splice site (**Figure 2.11**). I adapted my pipeline to annotate cassette exons with very high scoring splice donor sites at their precise 5' ends. I identified 47 expressed RP-exons, nested within 42 introns of 41 genes. Although these exons are skipped in many libraries, all have spliced RNA-seq evidence for expression, and thus represent a class of alternative splicing. A majority of these reside within 5' UTRs, although some specify coding sequences (CDS, 12) and alternative start codons (CDS:5' UTR, 8). The 20 loci that include CDS content exhibit conserved aggregate phyloP profiles indicative of coding sequence (**Figure 2.9B and Figure 2.12)**. I illustrate an RP-exon from *sm* with highly conserved coding sequence and strikingly conserved RP-like tandem SA:SD sequence at its 5' end (**Figure 2.9D**).

Overall, the sequence content at tandem SA:SD sites between the aggregate intronic RPs and expressed "RP-exon" classes is nearly identical (**Figure 2.9C**), and their total host intron lengths are also similar (~55 kb for intronic RPs, ~43 kb for expressed RP-exons). This suggests these are mechanistically similar splicing processes. Interestingly, there is a linear relationship between intron length and total number of RPs per intron (**Figure 2.9E**), and intronic RPs and expressed RP-exons tend to be evenly distributed throughout their resident introns (**Figure 2.9F**). Together, these

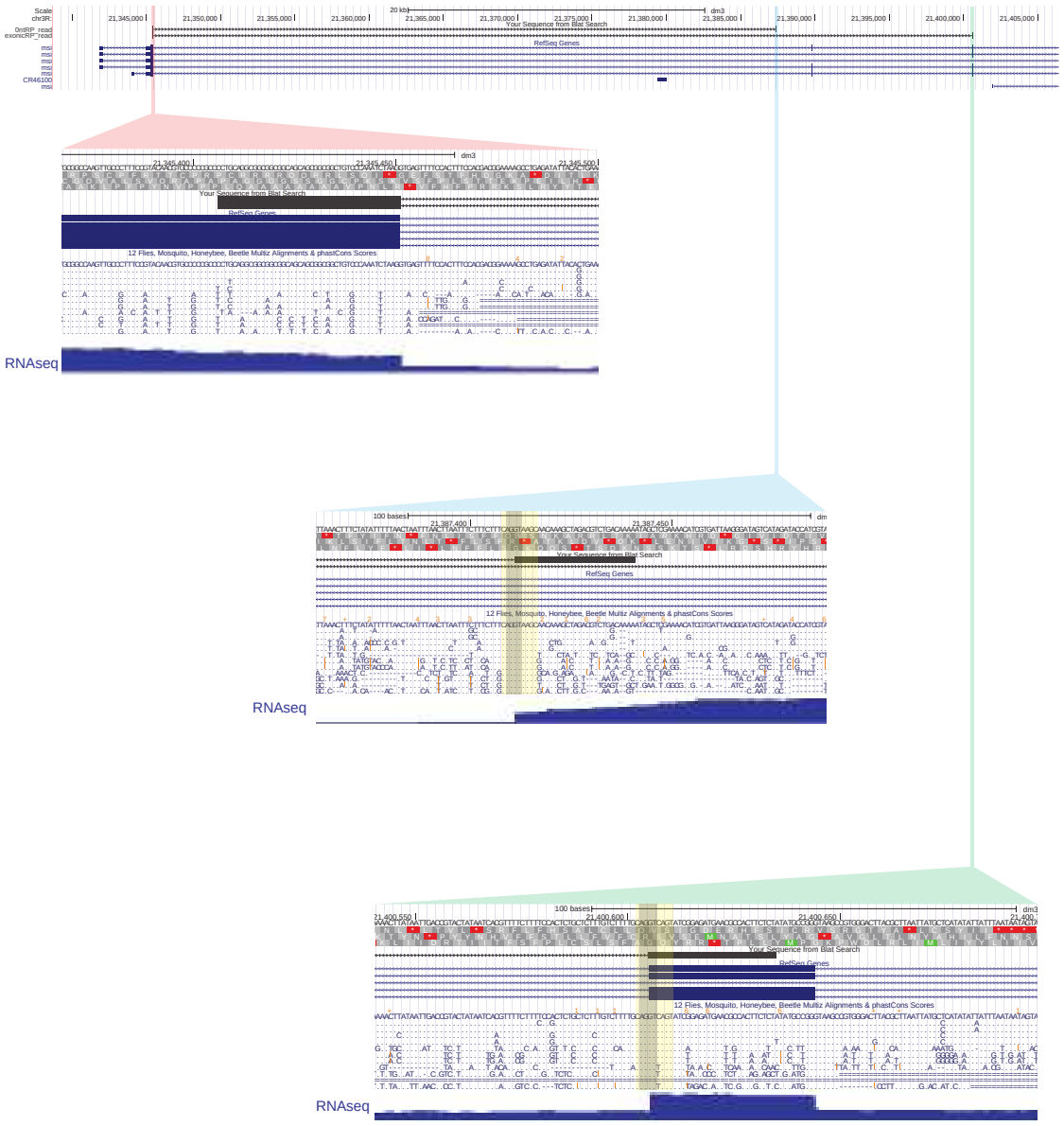


Figure 2.11. Example of Intronic RP and RP-exon annotation in the *msi* gene. The BLAT tool in UCSC genome browser was used to map RNA-seq reads to *musashi (msi)*. 5’ ends of reads map to a *msi* 5' exon (zoomed in shot highlighted in red). 3’ ends of one read maps to an intronic RP (blue highlight) and a zoomed in nucleotide-level screenshot is included in blue. 3’ ends of the other read maps to an RP-exon (green highlight) and a zoomed in nucleotide-level screenshot is included in green. Note the RNA-seq coverage in screenshots and that RP-exons have distinct exon coverage, whereas intronic RPs have sawtooth coverage pattern. The core AGGT splice acceptor-donor pairs are marked in gray, while the larger splice consensus motifs are highlighted in yellow.

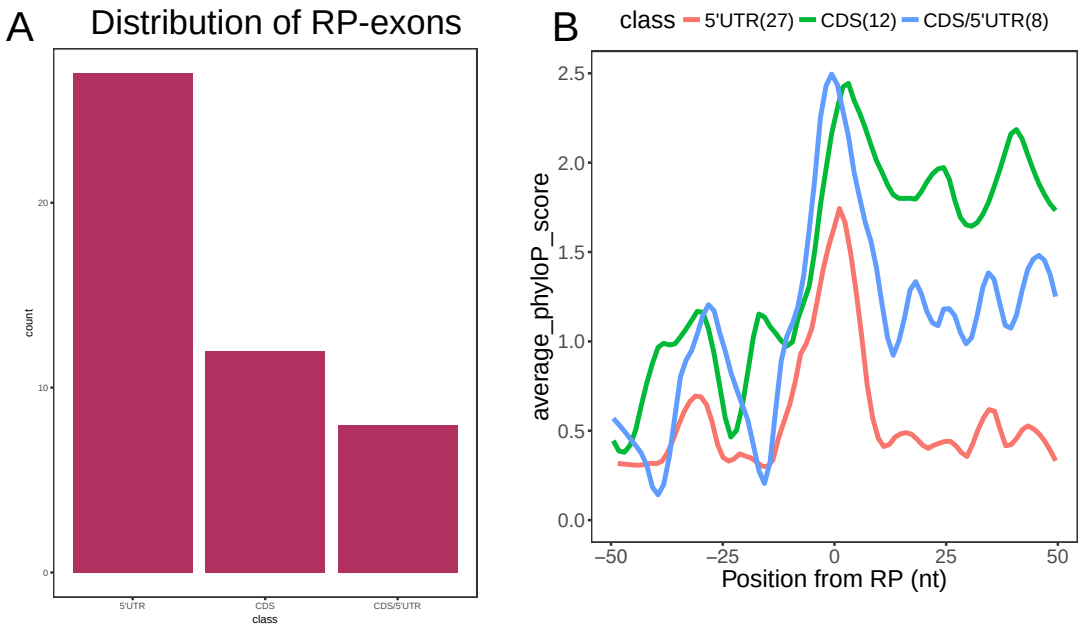


Figure 2.12. Conservation and coding properties of RP-exons by subcategory. (A) Distribution of RP-exons according to location in gene models. (B) RP-exons were divided according to their location in 5'UTR, CDS, and 5'UTR/CDS (ones that contained alternate 5'UTR/start sites). The fully coding RP-exons have a high level of evolutionary conservation, and the set with partial coding potential exhibit an intermediate level of conservation.

findings suggest that recursive splicing preferentially aids processing of long *Drosophila* introns.

Drosophila ratchet points are associated with cryptic exons genomewide

With my expanded annotation of recursive splice sites in hand, I assessed the breadth of the cryptic exon model for processing intronic RPs. I used NNSPLICE to score potential splice donors (SDs) in a 1 kb window downstream of intronic RP sites. Notably, within 100 nt of RPs, >1/3 of RPs had very high-scoring SDs (>0.8), and >1/2 of RPs scored >0.7 (**Figure 2.13A**). To investigate a potential positional bias of these SDs, I plotted their locations at various thresholds, and compared them against SDs downstream of 1000 control intronic AGGT sites. Amongst high scoring (>0.8) SDs, I observe a clear positional bias ~40-80 nt downstream of RP sites (**Figure 2.13B**), while background levels of high-scoring SDs were seen throughout the query window downstream of control AGGT sites. Analysis of other bins of SD scores showed similar positional bias, with modest enrichment even at the 0.5-0.6 range (**Figure 2.14A**). Thus, while my main analyses focus on the top-scoring sites, I conclude the strong majority of recursive sites utilize a positionally constrained cryptic donor.

I used phyloP to assess conservation of cryptic exon splice donors. I emphasize that when centering such analysis on RPs themselves (Duff et al., 2015), no positionally-biased conservation is apparent downstream (**Figure 2.9B**). However, bearing in mind that RP cryptic exons are heterogeneous in length, I reconfigured conservation analyses by centering on cryptic exon splice-site donors, segregated by distance from RPs (**Figure 2.13C and Figure 2.14B**). Satisfyingly, I now observe that high-scoring cryptic exon donor splice sites are highly conserved if they are within 100 nts of RPs, while those located further away are not conserved. There was lesser constraint on lower-scoring bins of cryptic splice donors, but clear selection remained at the same position

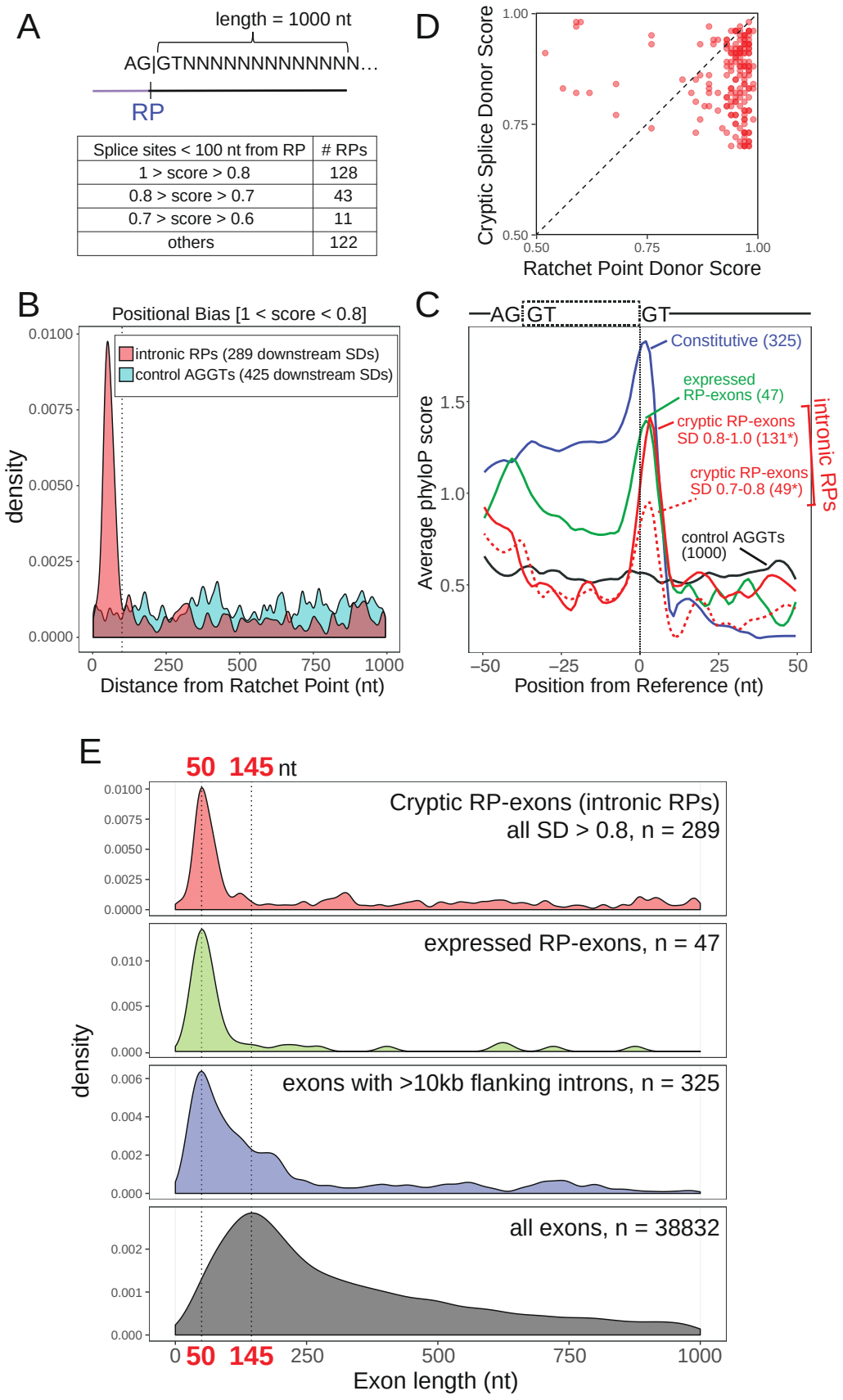
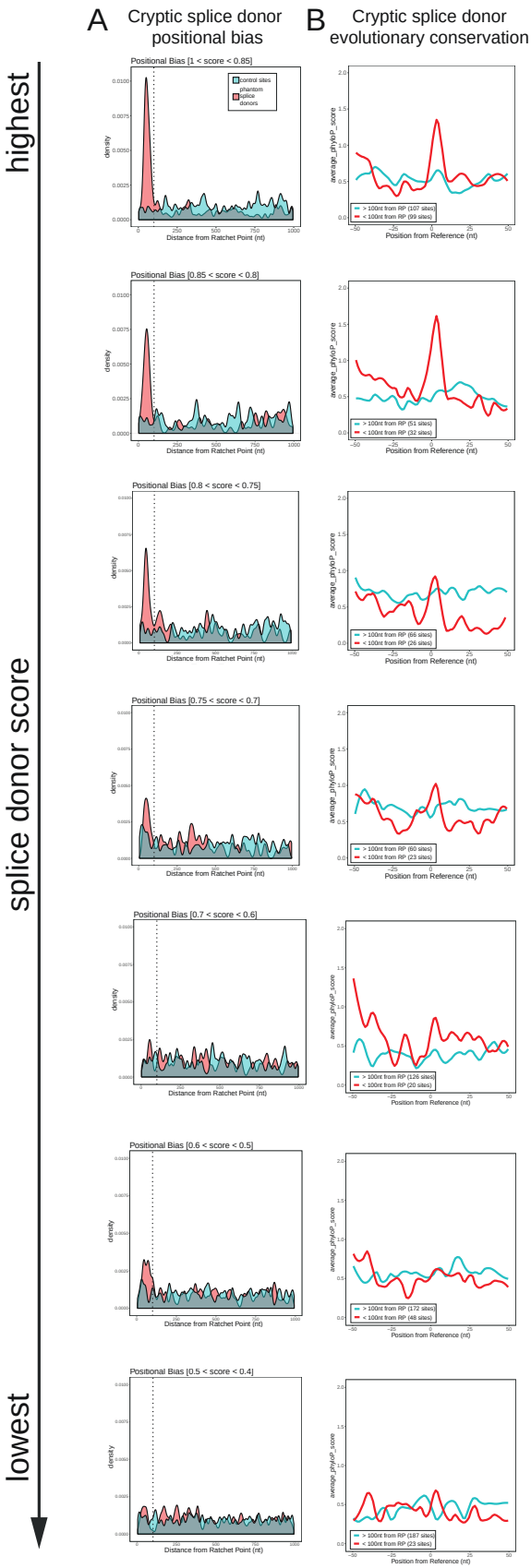

A
length = 1000 nt
AG|GTNNNNNNNNNNNNN…
RP
Splice sites < 100 nt from RP | # RPs
1 > score > 0.8 | 128
0.8 > score > 0.7 | 43
0.7 > score > 0.6 | 11
others | 122
D
Cryptic Splice Donor Score
Ratchet Point Donor Score
1.00
0.75
0.50
B
Positional Bias [1 < score < 0.8]
intronic RPs (289 downstream SDs)
control AGGTs (425 downstream SDs)
density
Distance from Ratchet Point (nt)
C
AG GT GT
Constitutive (325)
expressed RP-exons (47)
cryptic RP-exons SD 0.8-1.0 (131*)
cryptic RP-exons SD 0.7-0.8 (49*)
intronic RPs
control AGGTs (1000)
Average phyloP score
Position from Reference (nt)
E
50 145 nt
Cryptic RP-exons (intronic RPs) all SD > 0.8, n = 289
expressed RP-exons, n = 47
exons with >10kb flanking introns, n = 325
all exons, n = 38832
density
Exon length (nt)

Figure 2.13. Genomewide identification of RP-associated cryptic exons. (A) Schematic of strategy used to identify potential splice sites downstream of intronic RPs within a 1000nt window and a summary of non-overlapping RPs that were found to have high-scoring cryptic splice donors <100nt from the RPs. (B) Positional density of high-scoring splice donor sites downstream of intronic RPs (red) and control AGGT (cyan). A total of 289 SDs were found within a 1 kb region for all 304 RPs, whereas 485 SDs were found within a 1 kb region of 1000 control AGGT sites. Distance of splice site from RP is indicated on the x-axis and the dotted line marks 100nt from RPs. Similar positional bias was observed for lower-scoring bins of cryptic splice donors (**Figure 2.14**). (C) Evolutionary conservation of splice donors from constitutive exons [with flanking intron length >10 kb] (blue), RP-exons (green) and potential cryptic exons [SDs <100nt from intronic RPs] (red line – high score splice sites, red dotted line – moderate score splice sites). *Note that some intronic RPs had >1 high scoring SD within 100nt. (D) Cryptic splice donors are generally weaker than their paired RP donors. Each dot represents the NNSPLICE scores predicted at a given intronic recursive site; only the highest-scoring (>0.7) cryptic splice donors <100 nt from RPs were considered in this analysis. (E) Preferred length of exon definition for cryptic RP-exons, RP-exon and constitutive exons within long intronic contexts. Plotted are exon lengths inferred from intronic RPs + cryptic donor (red), RP-exons (green), constitutive exons with flanking intron length >10 kb (blue) and all *Drosophila* exons with flanking introns (black). In the case of intronic RPs, all SDs (>0.8 score) found within 1 kb of RPs were used to generate positional density plots (B, E); however, only proximal SDs are predominantly likely to contribute to cryptic exon processing.



C # unique RPs with at least one cryptic splice donor in the score category. Categories contain non-overlapping sets.

score	counts
>=0.85	98
0.8-0.85	30
0.75-0.8	24
0.7-0.75	19
0.6-0.70	11
0.5-0.60	24
0.4-0.50	9
others	89
total	304

Figure 2.14. Positional bias and conservation of cryptic donors stratified by splice scores. (A) Splice donors found downstream of RPs (cryptic splice donors) or 1000 control AGGTs sites were grouped based on NNSPLICE splice site strength. Plotted is the positional bias of splice donor site position relative to ratchet points. Substantial enrichment is observed in the ~40-80 window downstream of ratchet points, but not control AGGT sequences, at NNSPLICE scores down to 0.5-0.6 (B) Average phyloP scores of splice donor sites downstream of ratchet points (RPs) segregated into those that are <100nt from RPs and >100nt away from RPs. Clear local conservation is observed amongst groups of cryptic donors scored down to ~0.6. (C) Table showing the number of non-over-lapping RPs with cryptic splice sites grouped by splice site score.

relative to RPs (**Figure 2.14**). Thus, there is a strong evolutionary constraint on cryptic splice donors, even though their exons are ultimately not utilized in mature mRNAs. Overall, the strong sequence and positional constraint on cryptic splice donor sites, indicates their general importance for recursive splicing. Nevertheless, cryptic donor sequences are often less constrained than their partner RP sequences, as reflected by phyloP analyses.

In mammals, usage of cryptic recursively-spliced exons is suppressed following exon definition by competition (Sibley et al., 2015). As mentioned, *Drosophila* differs from mammals in that intronic RPs comprise an extremely abundant class of recursive splicing events. To test if suppression of cryptic exons at RPs can be accommodated by splice donor competition, I compared relative NNSPLICE strengths of pairs of intronic RP and cryptic donors. Of course, I could only do so for identified cryptic SD sites, which progressively get more modest in strength (**Figure 2.14**). I therefore focused this analysis on the very best cryptic splice donors, i.e., those with scores >0.7. Indeed, for the vast majority of cases, the cryptic donor is weaker than its partner RP donor (**Figure 2.13D**), consistent with the model for SD competition.

Finally, I plotted cryptic RP-exon sizes and found them to be tightly distributed around 40-80 nt in length. Notably, this matches the size range of my newly-recognized, broad class of recursive cassette exons (RP-exons). By comparison, the average length of *Drosophila* exons is about three times larger (**Figure 2.13E**). In light of this, I wondered whether recursive splicing exons utilize unique architectural properties. Broadly speaking, intron definition prevails when introns are short, while exon definition prevails when introns are long. However, to my knowledge, the correlation of flanking exon lengths has not been examined systematically in *Drosophila*. To evaluate this, I identified a set of constitutive exons that are embedded within long flanking introns (>10 kb flanking on either side). Interestingly, the length profiles of these exons mirror those

of recursive exons, both cryptic and RP-exons (**Figure 2.13E**). The preferred length of constitutive and recursive exons, and their distinct size profile, unifies the strategy of exon definition within long introns in *Drosophila*.

Discussion

In a prescient discussion on the first case of recursive splicing, Lopez and colleagues eloquently stated that "The internal exons of *Ubx* are not simply small ships adrift in an intronic ocean, precarious recognition of their splice sites causing more or less frequent skipping in different cellular contexts" (Hatton et al., 1998). Instead, *Ubx* microexons proved to be recursively processed, thereby regenerating 5' splice sites, such that these short exons could also be alternatively spliced in some isoforms. They concluded that "This mechanism has important implications not only for understanding alternative splicing regulation but also the processing of long introns in complex transcription units" (Hatton et al., 1998). Nearly twenty years later, I use genetic, molecular, and computational analyses that elaborate on the far-reaching implications of these statements.

In particular, following the recent molecular validation of ~200 ratchet point events in *Drosophila* (Duff et al., 2015), many of which were computationally predicted (Burnette et al., 2005), I provide evidence that redefines the mechanism of "0-nucleotide" recursive splicing and broadly extends the scope of both constitutive and alternative recursive splicing in *Drosophila*. I perform the first *in vivo* RP mutagenesis to demonstrate that disruption of the second step of recursive splicing that is required to skip the cryptic RP-exon can interfere with endogenous gene function. Notably, I show that characteristically-sized, conserved, cryptic exons are critical for recursive splicing via exon definition. That is, *Drosophila* has a preponderance of introns for which inclusion of the cryptic exon is constitutively suppressed, even though its recognition is

central to the recursive splicing process. Not only do I greatly expand the catalog of recursive splicing events that proceed via cryptic RP-exons, I annotate scores of recursive cassette exons in flies. Thus, *Ubx* recursive microexons are not a lone case, and these events represent a continuum of specialized alternative splicing. Importantly, my studies unify this mRNA processing strategy between flies and mammals, the latter of which may also have hundreds of recursive splice sites that are utilized under certain circumstances (Sibley et al., 2015).

Interestingly, the sequence content and length of cryptic exons are evolving, and their splice donor sites turn over more quickly than their companion RP sequences. Even when I experimentally manipulate a reasonably well-conserved cryptic splice donor site, the spliceosome can utilize a fortuitous, non-conserved, donor site. Therefore, RP cryptic exons harbor curious properties: they are functionally critical, yet relatively less conserved modules, in an otherwise well-conserved process of long intron splicing control. The evolutionary plasticity of recursive splicing could potentially lead to the interchange of cryptic and alternative RP-exons at long introns. I observe that progressively longer introns tend to have more recursive sites, that recursive sites are not randomly distributed within such introns but tend to subdivide them. Moreover, GO analysis indicates genes undergoing recursive splicing are enriched for developmental processes, especially neurogenesis and neuronal differentiation. Thus, recursive splicing may preferentially aid the processing of certain types of neural genes.

Methods

CRISPR-Cas9-mediated mutagenesis

kuz[RP] and Bx[RP]: Shu Kondo, a collaborator, used the transgenic Cas9-gRNA system (Kondo & Ueda, 2013) to perform mutagenesis in the *yw* background. In the case of *Bx*, a single gRNA was directed at the recursive splice site using a PAM

proximal to the AGGT sequence. In the case of *kuz,* two gRNAs were directed to flank the recursive splice site. In a typical transgenic CRISPR pipeline, 8 lines of candidate mutagenized chromosomes are established, and evaluated by PCR and Sanger sequencing (Kondo et al., 2017). For *Bx*, 5/8 candidates contained mutations in the vicinity of the ratchet point, but only one mutant disrupted the site, and contained an alteration in the RP donor. For *kuz*, 7/8 candidates contained mutations in the vicinity of the ratchet point, but again only one mutant actually disrupted the site and contained an alteration in the RP donor.

Ubx[RP]: I mutagenized the *Ubx-RP* using CRISPR-Cas9 and a single stranded oligo donor (ssODN) that abolished the ratchet point splice donor site. The gRNA was directed at the ratchet point and cloned into pCFD3. Injections were performed into *yw; nos-Cas9[II-attP40]* (BestGene Inc., Chino Hills) and the progeny of surviving animals were screened for site-specific incorporation of the ssODN. These experiments were far less efficient than the transgenic approach. I screened ~600 candidate lines, and recovered a single RP mutant.

All gRNA sequences, screening oligos and ssODN details are provided in **Table 2.2**.

Immunostaining

To study *kuz* phenotypes, I balanced *kuz[RP]* and the amorphic allele *kuz[e29-4]* (BDSC#5804) over *Cyo[Ubi-GFP]*. I collected embryos from the homozygous and trans-heterozygous crosses as well as control Canton S animals at 25°C. Embryos were aged, fixed and stained using the following primary antibodies: chicken anti-GFP (1:1000, Abcam #ab13970), mouse anti-BP102 (1:10, DSHB) and anti mouse-Fas II (1:100, 1D4, DSHB). Secondary antibodies used were made in donkey and conjugated to Alexa-488, -568 or -647 (Jackson ImmunoResearch). Stacks of images were obtained using a Leica

confocal microscope using a 40x oil immersion objective and maximum projections were generated using ImageJ-LOCI plugin. *kuz* mutant animals were identified by the lack of GFP staining.

To study *Ubx* phenotypes, I used the amorphic allele *Ubx[1]* (BDSC#2866) and the balancer chromosome TM3, *Sb*, *Ser*, which also carries the hypomorphic allele *Ubx[bx-34e]*. To stain for Ubx, *Ubx[RP]/TM6B-[ubi-GFP]* or *yw* flies were allowed to lay eggs in cages for 24 hrs at 25°C. After sufficient time, GFP-negative 1st instar larvae were hand-picked under a fluorescence microscope and dissected to obtain CNS. The samples were fixed and incubated with the following primary antibodies: rat anti-Elav (1:100, 7E8A10, DSHB) and mouse anti-Ubx (1:10, FP3.38, DSHB).

Constructs and cell culture

I created a minimal construct consisting of the following fragments stitched together from *kuz*: fragment of exon 2 (from start codon to end of exon 2), exon 3, a reduced version of intron 3 (131 nt of 5' end and 290 nt of 3' end), and 150 nt of exon 4. NotI and EcoRV restriction sites were added between the two intron 3 fragments to allow for further modifications. All fragments were cloned into pAC-5.1-V5-His using Gibson Assembly®.

To create pAC-kuzMG-kuzRI, a ~2.6 kb fragment surrounding *kuz-RP1* was PCRed from wildtype animals and cloned into the minimal vector using NotI and EcoRV sites. Similarly, to create pAC-kuzMG-BxRI, I used ~2.5 kb fragment surrounding *Bx-RP*. To obtain ratchet point splice donor mutants, PCR was performed on RP mutant animals and cloned into the minimal vector. I used site directed mutagenesis to make all other mutants constructs. All primers used for cloning and mutagenesis can be found in **Table 2.2**.

All transfections in this study were performed using S2-R^+ cells cultured in Schneider *Drosophila* medium with 10% fetal Bovine serum. Cells were seeded in 6-well plates at a density of 1 million/mL and transfected with 100 ng of construct using the Effectene transfection kit [Qiagen]. Cells were harvested following three days of incubation.

rt-PCR of mRNA and recursive intermediates

Ubx[RP] and *kuz[RP]*: mutant stocks were made with GFP balancers. Homozygous 1st instar larval mutants (GFP-) were hand-picked under a fluorescence microscope. Animals were homogenized and RNA was extracted using the standard Trizol protocol. 2 µg of RNA was treated with Turbo DNase [Ambion] for 45 min before cDNA synthesis using SuperScript III [Life Technology] with random hexamers. rt-PCRs were done using AccuPrime™ Pfx DNA polymerase [ThermoFisher Scientific] with standard protocol using 28 cycles for mRNA and 34 cycles for intermediates.

Bx[RP] - similar to *Ubx[RP]* and *kuz[RP]*, except for the following differences: homozygous mutant adult flies were homogenized and RNA was extracted using Trizol. 5 µg of RNA was DNase treated and reverse transcribed using random hexamers. rt-PCRs were done using 35 cycles for mRNA and intermediates.

Cell culture: RNA was collected from transfected cells using Trizol. 5 µg of RNA was treated with Turbo DNase [Ambion] for 45 min before cDNA synthesis using SuperScript III [Life Technology] with random hexamers. rt-PCRs were done using AccuPrime™ Pfx DNA polymerase [ThermoFisher Scientific] with standard protocol using 26 cycles and primers that were specific to minigene construct. All primers with descriptions can be found in **Table 2.2**.

Bioinformatic annotation of putative ratchet points from nascent RNA datasets

I used publicly available nascent RNA-seq and genomic run-on sequencing (GRO-seq) datasets from NCBI's sequence read archive (Chen et al., 2016; Ferrari et al., 2013; McMahon et al., 2016; Mohn et al., 2014; Rodriguez et al., 2012; Rozhkov et al., 2013; Sienski et al., 2012; Wang et al., 2015), described in Table 2.1. The datasets were mapped to the *Drosophila melanogaster* BDGP R5 (dm3) reference genome using TopHat2 (Kim et al., 2013) under default settings.

In theory, recursive splice sites should contain tandem splice acceptor and donor sequences (Figure 2.1A). Therefore, to identify putative recursive splice sites, I first collated all junction-spanning loci (in the case of splice junctions, introns) and kept those that contained the AGGT tetranucleotide across the 3' end (Figure 2.8). This ensures that the junctions have tandem minimal consensus splice acceptor (AG) and splice donor (GT) motifs. I then classified the 3' ends of AGGT junctions based on where they occur, such as exon junctions, exon body, introns, cassette exon junctions or intergenic space, using RefSeq Gene annotations (Figure 2.8). Since recursive splice sites should occur within intronic regions of the transcriptome, I only further analyzed events that were unambiguously intronic ("0-nt" recursive splicing candidates) or mapping to cassette exon junctions (RP-exon candidates – see below). These loci were examined in depth for potential ratchet points.

Up to this point, to cast a wide net, I had qualified all AGGT junctions as candidates. However, to narrow down to a set of likely candidates, I employed the splice site prediction tool, NNSPLICE (Reese et al., 1997), to quantify 3' and regenerated 5' splice site scores (Figure 2.8). Simultaneously, I merged and converted all nascent RNA and GRO sequencing datasets into the browser-friendly bigWig format, and manually inspected all intronic recursive splicing candidates for the characteristic saw-tooth pattern expected of ratchet points. I found 271 sites that had high splice site scores and were supported by clear saw-tooth patterns (Figure 2.9A, Figure 2.8).

However, I observed that the strength of the saw-tooth pattern was a continuum, which likely depends on library properties such as coverage and inherent host gene properties, such as recursive intermediate stability. Therefore, I reasoned that I were systematically removing potential RPs by requiring a saw-tooth pattern, and sought to acquire these by selecting sites with high splice site scores. I set splice score cutoffs to mirror the scores of RPs that were supported by saw-tooth patterns (< 0.75) and found a total of 33 additional intronic RP loci (Figure 2.9A, Figure 2.8).

Bioinformatic annotation of RP-exons from nascent RNA datasets

After identifying cryptic exons associated with intronic RPs through bioinformatics and experimentation, I inferred that some known cassette exons might also be processed by recursive splicing. I utilized an analogous strategy to identify a set of expressed RP-exons for further inspection (see above) and supplemented these with all annotated cassette exons that were not sampled due to sample or tissue type. Typically, saw-tooth patterns are used in the annotation of RPs. However, since expressed RP-exons are stable exons recovered in transcriptomic data, I had to rely on splice site score alone to predict potential recursive splicing. Therefore, I scored all 3' and regenerated 5' splice sites from cassette exon junctions using NNSPLICE, and employed a strict cutoff of 0.85, resulting in the annotation of 47 expressed RP-exons (Figure 2.8). These were generally alternatively spliced.

Identification of potential cryptic exon donor splice sites

Following the observation of intron retention in *kuz[RP]*, *Bx[RP]* and *Ubx[RP]*, I manually browsed other RPs and noticed a similar and regular occurrence of 5' splice sites downstream of the AGGT site. To formalize this observation, I used NNSPLICE to search for splice sites of varying strengths within a 1 kb region downstream of all

putative intronic RPs. As control, I used a set of AGGT sites from introns of matched length as the RPs, and likewise looked for splice sites.

Evaluation of recursive splice site and cryptic exon donor site conservation

I used phyloP scores from the UCSC Genome Browser to assess conservation. Briefly, potential ratchet points from intronic cryptic RP-exon and expressed RP-exon categories were anchored at position 0 and phyloP scores for all sites were summed and averaged at each position from -50 to +50.

To calculate conservation of cryptic exon splice donors, I split all cryptic splice sites into two categories: those occurring <100 nt of RPs and those occurring >100 nt from RP. For each set of cryptic sites, I anchored each cryptic splice site at position 0 and calculated phyloP scores for each nucleotide position from -50 to +50. To graph conservation, I averaged the phyloP scores at each position per set.

Statistics and Reproducibility

Adult fly phenotypes in Figure 2.2B-D were evaluated using >100 animals from each genotype, and the phenotypes were completely penetrant. For immunostaining experiments in Figure 2.2F-G, I imaged 3 *yw* and 2 *Ubx[RP/RP]* 1st instar larval CNS, and the normal Ubx pattern was observed in all wildtype CNS and completely absent in *Ubx* mutants. For immunostaining experiments in Figure 2.2H-I, I imaged 7 *yw* and 7 *kuz[RP]/[RP]* embryonic CNS, and the mutant phenotypes were completely penetrant. For Figure 2.2J-M, I imaged 7 *kuz[RP]/+* (J), 8 *kuz[e29-4]/[e29-4]* (K), 9 *kuz[RP]/[RP]* (L) and 3 *kuz[RP]/[e29-4]* embryonic CNS. All control CNS were normal whereas all *kuz* mutant combinations exhibited the defects shown in the representative images.

For Figure 2.4, the rt-PCR experiments were performed in biological triplicates and the cell culture reporter experiments were performed in biological duplicates.

Table 2.1 Nascent RNA mapping statistics from data obtained from previously published work

SRR ID (single)	Lab	Tissue	Type	total reads	mapped reads	Unmapped rea	non-unique	unique	unique map%	PMID
SRR1187957	Brennecke	Ovary	stranded	96701231	4023105	92678126	2660776	1362329	1.41	24906153
SRR2033380	Brenner	S2	stranded	31605464	24054601	7550863	17303504	6751097	21.36	NA
SRR2033381	Brenner	S2	stranded	34209235	5810181	28399054	3359720	2450461	7.16	NA
SRR3177688	Rosbash	S2	unstranded	141978235	137050807	4927428	108567691	28483116	20.06	27040499
SRR3177689	Rosbash	S2	unstranded	187909521	180650355	7259166	143082768	37567587	19.99	27040499
SRR3177690	Rosbash	S2	unstranded	181104807	174264024	6840783	126394549	47869475	26.43	27040499
SRR3177691	Rosbash	S2	unstranded	73611053	71848652	1762401	49448361	22400291	30.43	27040499
SRR3177692	Rosbash	S2	unstranded	45974714	45062795	911919	32888626	12174169	26.48	27040499
SRR3177693	Rosbash	S2	unstranded	83139627	80676422	2463205	49742772	30933650	37.21	27040499
SRR3476588	Aravin	Ovary	stranded	51916059	50592760	1323299	46436181	4156579	8.01	27292797
SRR3476589	Aravin	Ovary	stranded	39131961	38026599	1105362	34250512	3776087	9.65	27292797
SRR485849	Rosbash	Head	unstranded	16656421	14647509	2008912	8614738	6032771	36.22	22658416
SRR485850	Rosbash	Head	unstranded	17159544	15020668	2138876	7016063	8004605	46.65	22658416
SRR485851	Rosbash	Head	unstranded	17701096	15802492	1898604	8157018	7645474	43.19	22658416
SRR485852	Rosbash	Head	unstranded	17215251	14920102	2295149	8610110	6309992	36.65	22658416
SRR485853	Rosbash	Head	unstranded	17189093	15407246	1781847	7900900	7506346	43.67	22658416
SRR485854	Rosbash	Head	unstranded	26218224	22826693	3391531	14499045	8327648	31.76	22658416
SRR485855	Rosbash	Head	unstranded	26316602	23850022	2466580	15828630	8021392	30.48	22658416
SRR485856	Rosbash	Head	unstranded	27941339	24311525	3629814	17245062	7066463	25.29	22658416
SRR485857	Rosbash	Head	unstranded	28962116	25664273	3297843	16899335	8764938	30.26	22658416
SRR485858	Rosbash	Head	unstranded	28374323	25710023	2664300	17317257	8392766	29.58	22658416
SRR485859	Rosbash	Head	unstranded	28819929	24821772	3998157	17638864	7182908	24.92	22658416
SRR485860	Rosbash	Head	unstranded	29834822	24816347	5018475	20892381	3923966	13.15	22658416
SRR485861	Rosbash	Head	unstranded	72707476	65650810	7056666	58582419	7068391	9.72	22658416
SRR485862	Rosbash	Head	unstranded	71446759	63246824	8199935	57359546	5887278	8.24	22658416
SRR485863	Rosbash	Head	unstranded	83266432	73155930	10110502	65093044	8062886	9.68	22658416
SRR609665	Brennecke	OSC	stranded	64287263	8283591	56003672	2233351	6050240	9.41	23159368
SRR609666	Brennecke	OSC	stranded	35016070	3409408	31606662	1066540	2342868	6.69	23159368
SRR646576	Hannon	Ovary	stranded	22891787	5102086	17789701	2337337	2764749	12.08	23392609
SRR646577	Hannon	Ovary	stranded	24861675	5165579	19696096	2133214	3032365	12.2	23392609
SRR646578	Hannon	Ovary	stranded	21075508	5002958	16072550	2377204	2625754	12.46	23392609
SRR646579	Hannon	Ovary	stranded	20435211	4294399	16140812	2204854	2089545	10.23	23392609
SRR836452	Kuroda	S2	stranded	10881777	682398	10199379	299284	383114	3.52	24183666
SRR836453	Kuroda	Kc-167 cells	stranded	13440030	2116022	11324008	1047288	1068734	7.95	24183666

SRR ID (paired)	Lab	Tissue	Type	eft total	eft mapped	eft unmapped	eft non-uniqu	eft unique	eft unique m	ight total	ight mapped	ight unmapped	ight non-unique	ight unique	ight unique map%	PMID
SRR1999059	Zamore	Ovary	stranded	43108388	41039681	2068707	28805897	12233784	28.38	43108388	40105036	3003352	28082075	12022961	27.89	26340424
SRR1999060	Zamore	Ovary	stranded	41181542	38082094	3099448	9867635	28214459	68.51	41181542	36993683	4187859	9524095	27469588	66.7	26340424
SRR1999061	Zamore	Ovary	stranded	43681414	41116370	2565044	14268857	26847513	61.46	43681414	39959410	3722004	13807662	26151748	59.87	26340424
SRR1999062	Zamore	Ovary	stranded	43025345	38689739	4335606	10479304	28210435	65.57	43025345	37704482	5320863	10154401	27550081	64.03	26340424
SRR1999063	Zamore	Ovary	stranded	42641295	39467465	3173830	13666245	25801220	60.51	42641295	38466367	4174928	13248522	25217845	59.14	26340424
SRR1999064	Zamore	Ovary	stranded	43978787	40141073	3837714	14468104	25672969	58.38	43978787	39182705	4796082	14084001	25098704	57.07	26340424
SRR1999065	Zamore	Ovary	stranded	40773873	37781982	2991891	13451250	24330732	59.67	40773873	36768918	4004955	13028851	23740067	58.22	26340424

Table 2.2 Primer sequences (Page 1 of 2)

gRNA and ssODN	**sequence**
Bx_gRNA	GGTGGGTGTTGACACTTACC
kuz_gRNA1	CATTACAATATATTTGATTA
kuz_gRNA2	TCTCTTTACAGGTGAGTGCT
Ubx_gRNA	ACTATTTTCTTCTCTTTTCT
Ubx_ssODN	TAAAACTTGAGATTTTCTTATTTAAATATG CATGTCTACTTTTTGTACTCACTGTTTGC CTAATACTA ATCAAACTATTTTCTTCTCTTTTCTAGAAT TCT GTCAAATATTTAATACACCCTTAAACCAA

Genotyping	**sequence**
Bx.RI-CHKF	CTGGGTGCCAAGGGTGATGATGAATGT
Bx.RI-CHKR	AGCCAGTCAGCGGCAGCGGCGACAAAAC
kuz.RI-CHKF	ATGGACCTCTTTATCTGCACGGTTTTG
kuz.RI-CHKR	ACGGCCTGCCCGCAGAAAAGCTGCTAAC
Ubx.RI-CHKF	CTTTACACCTTTACACGGGCGTATTTTC
Ubx.RI-CHKR	GGATGGCAGGGGTGTGTGGGTGCTATG

RT-PCRs

kuzbanian (kuz)	**sequence**	**Description**
kuzexon2-new.fwd	CGACAGCCATCCACGTTGGATC	endogenous mRNA
kuzexon4rvs	CGCTCTATTGTGACTAGCTCGGATG	endogenous mRNA
kuzexon2-new.fwd	CGACAGCCATCCACGTTGGATC	endogenous recursive intermediate 1
kuzRI1rvs	ATGGAACCAGTCATCCTCGTCC	endogenous recursive intermediate 1
kuzexon2-new.fwd	CGACAGCCATCCACGTTGGATC	endogenous recursive intermediate 2
kuzRI2rvs	TCCAGCTCGATTAAGATGTCTTCC	endogenous recursive intermediate 2

Beadex (Bx)	**sequence**	**Description**
bxdistE1fwd	CGAACCGACCGCAAAAGC	**endogenous mRNA**
bxE2rvs	AGTTGACCACATTGACCACG	**endogenous mRNA**
bxdistE1fwd	CGAACCGACCGCAAAAGC	**endogenous recursive intermediate**
bxRIrvs	CTAATTGTTGTTGTGCTGCCG	**endogenous recursive intermediate**

Ultrabithorax (Ubx)	**sequence**	**Description**
ubx.dM2.fwd	GCTATCGCAGGTAAGAGATACTC	**endogenous mRNA - isoform D**
ubx.mRNA.rev	CATCTCGATTCTCCGTCTG	**endogenous mRNA - isoform D**
ubx.m2int.fwd	GCTCACTTCTACCAGACTG	**endogenous recursive intermediate**
ubx.int.rvs	CTTTGCCCAGCACGCATGAG	**endogenous recursive intermediate**

Table 2.2 Primer sequences (Page 2 of 2)

Cell culture primers

splicing reporter rtPCR primers

kuz.mg.fwd1	TACTAGTCCAGTGTGGTGG
kuzexon4rvs	CGCTCTATTGTGACTAGCTCGGATG

Gibson assembly of splicing minigene

Gibson Assembly of minimal region with kuz exon2,3 and 4 as well as short intronic regions

E3.NotI.junction_fwd	tagtccagtgtggtggaattctgcaGTTACGCAAAA GATATTTCTGGAGTTAAAAG
E3.NotI.junction_rev	CGATATCTACGATATGCATTGCGGCCGC TACGAACACCTAGTTGAAATCC
EcoRV.E4.junction_fwd	GGCCGCAATGCATATCGTAGATATCGTG TGGACTGGTCTGGTCTG
EcoRV.E4.junction_rev	ccttcgaagggccctctagactcgaGGAAGTCTCT TCCATGTG

To insert RP locus into the above vector using restriction enzyme cloning, the following primers were used

Bx_Not1.Bx.fwd	ATGCTTGAGCGGCCGCACAATCACGAT CCGCTGTTG
Bx_EcorV.Bx.rvs	ATGCTTGAGATATCATTTTCGTTTGTTG CACTGCC
kuz_Not1.kuzRI.fwd	ATGCTTGAGCGGCCGCCAGTGGCGAA AAGGCAATGG
kuz_EcorV.kuzRI.rvs	ATGCTTGAGATATCCCACACACTACAG CACTAC

To mutagenize phantom donor in wildtype and RP mutant, the following primers were used

Bxphant.SDM_F	CAGTGTGCGAaGGGCCCA
Bxphant.SDM_R	tGCATGAATCTTGTGTGTTGTAGCAATTG
kuz.phantom.mut_F	GCTGCTCCAGctacaGTTTGCTTTTATTATC
kuz.phantom.mut_R	GAACTCCGTTTTTATGAAGTAC

Chapter 3

Molecular and genetic dissection of intron recursive splice sites in *Drosophila*

Summary

Intronic RPs are now appreciated to be fragments of cryptic RS-exons. Thus, during RS a cryptic exon is defined prior to removal of the upstream intron. Notably, this step results in the regeneration of a 5'SS that predominantly outcompetes the cryptic RS-exon 5'SS. In this chapter, I first investigate mechanisms that regulate the choice between the RP 5'SS and the RS-exon 5'SS. Using a combination of *in vivo* RP mutagenesis and cell culture tests, I identify 5'SS strengths and exonic splicing regulatory elements as factors that can alternate choice of 5'SS. In contrast, I find that the exon junction complex may specifically suppress activation of the RP-5'SS within the same context. Lastly, I use multiple CRISPR/Cas9 mutagenesis strategies to create the first ever panel of *in vivo* RS-exon deletions. I remove 9 RPs from 5 genes and identify *Ubx* as a particularly sensitive target. Overall, these studies represent the first collection of tests designed and executed to understand the regulation and function of recursive splicing.

Introduction

Biological processes are enriched by a number of distinct regulatory pathways. These interpret biological state and fine tune signals. In the case of intron removal, alternative splicing is a regulatory layer that provides gene expression control, and dysregulation of splicing can often lead to disorder and disease (Scotti & Swanson, 2016). It is within this context that understanding the bounds of splicing regulation can hold immense significance and facilitate mechanism-based targeted therapeutics. In the last chapter, I discovered that intronic RPs are fragment of short cryptic RP-exon (hereafter referred to as RS-exons). Thus, in recursive splicing, short cryptic RS-exon definition precedes removal of the upstream intronic segment and regenerates a RP 5'SS (**Figure 2.1C**). In the subsequent step, the RP 5'SS is activated (instead of the RS-exon 5'SS), and this results in processing of the remaining intronic sequence. Thus, the RS-exon is defined, but the zero-nucleotide exon is activated (Joseph et al., 2018).

Canonical cassette exons are typically frame preserving and symmetric. This means that their lengths are a multiple of three and that they occur in the phase zero format (Long et al., 1995). Intriguingly, RS-exons do not generally share these properties, and I previously proposed that these exons do not appear to be under the same constraints as coding exons, presumably because they are never included in mRNA (Joseph et al., 2018). However, it is worth appreciating the severe consequences of accidental cryptic RS-exon inclusion in mRNA, especially those that lie between coding sequence, as these may alter translational reading frame or contain premature stop codons (Joseph et al., 2018; Sibley et al., 2015). A good illustration of this can be found in chapter 2, where disruptions of the intronic RP 5'SS in *kuz* and *Ubx* lead to host gene loss-of-function via inclusion of frame changing cryptic RS-exons. Implicit in this argument is the notion that cryptic RS-exon skipping must be tightly regulated. However, little is known about the regulation of intronic RS-exon skipping.

The most important indications come from quantification of RP 5'SS versus the RS-exon 5'SS. In both insects and mammals, it appears that the intronic RPs have stronger 5'SS than their corresponding RS-exons. Hence SS competition has been proposed as one basis for zero-nucleotide splicing. A functional test in cell culture using a minigene reporter with weakened RP 5'SS resulted in RS-exon inclusion, providing support for the SS competition model (Joseph et al., 2018; Sibley et al., 2015). Nevertheless, it remains unclear if the reduced introns of minigene reporters can appropriately mimic the challenges of long introns, or if SS competition matters in the context of endogenous genes.

Other clues come from studies of expressed RS-exons in *Drosophila* which are predominantly focused on the *Ultrabithorax (Ubx)* microexons m1 and m2. Originally identified as 51 nt cassette exons within a ~73 kb long intron, these exons were reannotated as recursively spliced by the Lopez lab (Hatton et al., 1998). Furthermore, since these are coding exons and alternatively spliced (specifically in the fly nervous system (Artero et al., 1992)), attention has been paid to factors that may regulate alternative splicing. Consequently, splicing factors such as *virilizer, fl(2)D*, and *hrp48* have emerged as regulators in *trans* (James M. Burnette et al., 1999). Consistent with this notion, *cis* elements on the *Ubx* m1 microexon have also been identified that potentially enhance inclusion of the m1 RS-exon (Hatton et al., 1998). As regulation through *cis/trans* elements appears to regulate expressed RS-exons, it is worth examining if such mechanisms can also regulate *Drosophila* intronic RPs (cryptic RS-exons).

More recently, the Ule lab has discovered that splice donors of certain mammalian recursive splice sites are constitutively suppressed through the action of the core exon junction complex (EJC) and peripheral factor RNPS1. These recursive splice sites are never used as zero nucleotide exons, and can only be detected to do so under

EJC loss-of-function (Blazquez et al., 2018). These factors are – in theory – stabilized ~20-24 nt upstream of the RSS splice donor and are positionally poised to influence splice site choice. However, as the splice donors of *Drosophila* intronic RPs are constitutively activated, it is reasonable to ask if *Drosophila* RS-exons are also sensitive to EJC recruitment.

Beyond regulation, the function of recursive splicing/RS-exons remains a mystery. Since the initial discovery of recursive splicing, there have been no *in vivo* tests of the requirements or function of cryptic or expressed RS-exons. Nevertheless, their deep conservation suggests value for some aspect of host gene expression. The only instance of a recursive splice site ablation is the hypomorphic allele, Ubx^{MX17} (Busturia et al., 1990). This X-ray induced mutant allele constitutively skips microexons m1 and m2 in *Ubx* mRNA, and displays classic phenotypes, including the haltere to wing transformation, as well as changes in the positional identity of neuroblasts in the embryonic CNS (Busturia et al., 1990; de Navas et al., 2011; Geyer et al., 2015; Subramaniam et al., 1994). Besides homeosis, Ubx^{MX17} also displays flight and behavioral defects, arguing that the RS-exons have an important role in the function of the protein (Subramaniam et al., 1994).

RS may also have roles in RNA processing. Originally, it was proposed that RS was required to ease the challenges of splicing long introns. This view was supported by the observation of a linear correlation between host intron length and number of RPs (Joseph et al., 2018; Pai et al., 2018). However, a recent study interested in splicing kinetics found that RP containing introns are in fact processed slower than length-matched control introns (Pai et al., 2018). Nevertheless, RS may also influence alternative splicing. The Ubx^{MX17} allele is a large ~18 kb inversion of sequence that surrounds the m2 exon. Yet somehow, loss of the m2 at the DNA level also yields loss of m1, but during pre-mRNA processing (de Navas et al., 2011; Subramaniam et al., 1994).

That the m2 microexon may influence the inclusion of m1 in mRNA argues that RS-exons may have a role in RNA processing. However, these interesting functional alterations – attributed to RS-exons – could also have been caused by other disruptions contained within the ~18 kb inversion allele. Thus, irrespective of the result, it seems worthwhile to develop a system to functionally examine RS-exons in the context of pre-mRNA processing and animal development.

In this chapter, I study the regulation of SS selection within cryptic and RS-exons. I consider the influence of 5'SS strength, exonic *cis*-elements and upstream intron removal (EJC deposition). My results suggest a role for all three in the contextual regulation of 5'SS choice. Additionally, I use two different strategies to delete RPs in the fruit fly and generate the first ever panel of RP deletions. Animals with deletions of *Ubx* m1 and m2 display homeotic transformations, with the m2 deletion expressing stronger phenotypes. While deleted RPs generally appear dispensable for host pre-mRNA processing, loss of the m2 microexon in *Ubx* induces changes in RNA processing. Overall, this work provides a broad view of control and function of recursive splicing.

Results

in vivo RP mutageneses verifies 5'SS competition as a determinant of RS-exon inclusion

If the decision of RS-exon inclusion were determined based on consensus match, or SS strength, weakening the dominant SS should produce a reversal of the wildtype RS-exon inclusion outcome. This strategy was successfully adopted by Sibley and colleagues using a minigene RS reporter in mammalian cell culture (Sibley et al., 2015). While their work strongly supports the SS competition model, it is currently

unclear whether the same mechanism controls RS-exon inclusion within the endogenous context of unusually long introns.

In Chapter 2, I demonstrated that a transgenic CRISPR-Cas9 approach can be employed to precisely target and mutagenize RP 5'SS in the *kuz* and *Bx* loci. As the guide RNAs used either directly targeted the recursive splice sites or nearby sequences, the same reagents were reused to generate a new panel of variants that specifically weakened the RP 5'SS but kept all other elements intact. Progeny of animals carrying the CRISPR and Cas9 transgenes were screened and alleles harboring mutations that altered the RP splice donor were selected. Even though these animals carry a diverse set of *cis*-indels, the overall effect is a change in the RP 5'SS. For *Bx,* I identified seven mutants that preserved the GU dinucleotide (+1 to +2 position) of the RP splice donor but contained deviations in positions +3 to +8 (**Figure 3.1A, Table 3.1**). Similar to *Bx[∆RP]* (Chapter 2), all seven mutants appeared to be viable in homozygosis. Indeed, since the *Bx* RS-exon resides in the 5'UTR*,* alternative splicing has no effect on the reading frame of the coding sequence. Quantification of SS match using NNSPLICE confirmed that the mutations resulted in a range of RP splice donor strengths, from moderate (#s 13 and 20), to weak (# 16), and poor (#s 12, 21, 23 and 24) as well. Notably, all seven mutant RP splice donors are theoretically weaker than the cognate RS-exon 5'SS, which remained unchanged (**Figure 3.1A**).

I analyzed molecular consequences of RP mutations on RNA processing. rt-PCR analyses to detect the intermediate amplicon downstream of the ratchet point yielded the expected products for all *Bx* mutants (**Figure 3.1B-C**). The intermediate amplicon indicates usage of the recursive splice acceptor. Since the induced mutations did not damage the recursive 3'SS (including the +1 to +2 position), activation of this SS was not expected to be altered. Instead, changes in RS-exon inclusion levels were evaluated on mature transcripts. Remarkably, rt-PCR of mRNA amplicons indicated that all

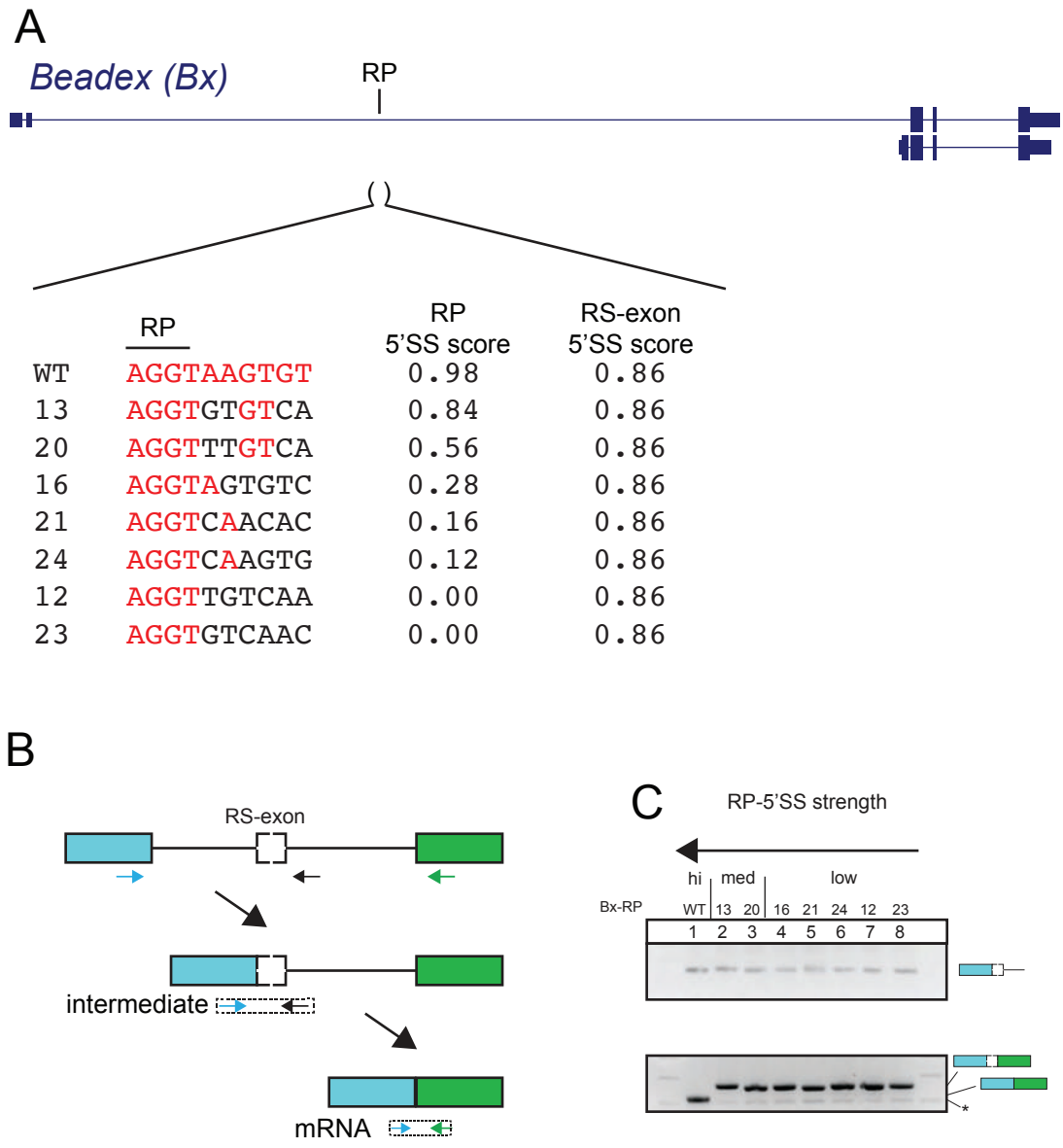


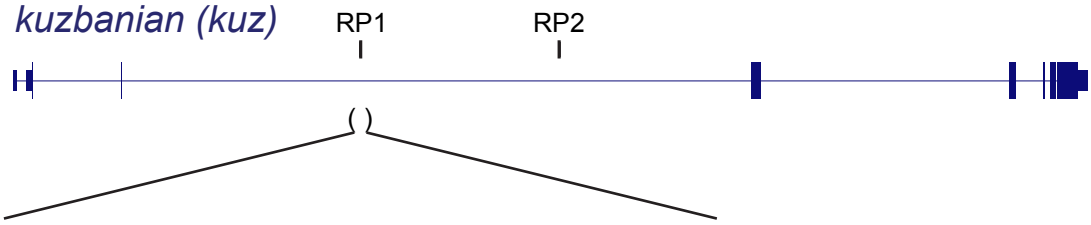
Figure 3.1. Weakening the *Bx* RP 5'SS *in vivo* results in RS-exon inclusion.
(A) *Bx* gene models displaying isoforms that use different transcription start sites. Notably, the RP is found within the longer isoform in an ~31 kb intron 2. Precisely targeted mutations that alter the RP 5'SS are listed below. The red color conveys match to wildtype RP 5'SS whereas black indicates nucleotide changes. The allele ID is left of the sequence and changes to RP 5'SS score on the right. The unchanged RS-exon 5'SS score is also included.
(B) A model for *Bx* intronic recursive splicing. PCR amplicons are displayed using dotted boxes and primers as arrows.
(C) Wildtype and RP 5'SS mutants yield RS intermediate amplicons. However, unlike wildtype, all weakened RP mutants include the cryptic RS-exon.

changes to the RP splice donor strength (moderate, weak or poor) resulted in a complete switch to RS-exon inclusion (**Figure 3.1B-C**). As all RP 5'SS variants generated were weaker than the RS-exon splice donor, these data support a model in which SS strength drives alternative splicing. Thus, *Bx* RS-exon skipping *in vivo* appears to be a result of activation of the stronger RP splice donor.

The same strategy was used to obtain RP1 5'SS variants in the *kuz* locus. Six mutants were identified that gradually weakened this optimal SS. This included variant #14, which bears a 2 nt substitutions at positions +6 and +7 of the SS and induces a slight decrease in splice score from 0.97 to 0.94 (1.00 being the highest). Another variant (#30) contained substitutions at additional positions, resulting in a moderate score (0.55). Finally, a set of four mutants bear changes in position +3 to +8, measuring weak SS scores in the 0- 0.21 range. A control allele that contained no mutations in the RP 5'SS was maintained as a control (#24). Similar to the *Bx* subjects, the RS-exon 5'SS for these mutants remained unchanged. Critically, only #s 24 (control) and 14 had RP 5'SS that were still significantly stronger than the RS-exon 5'SS (**Figure 3.2A, Table 3.2**).

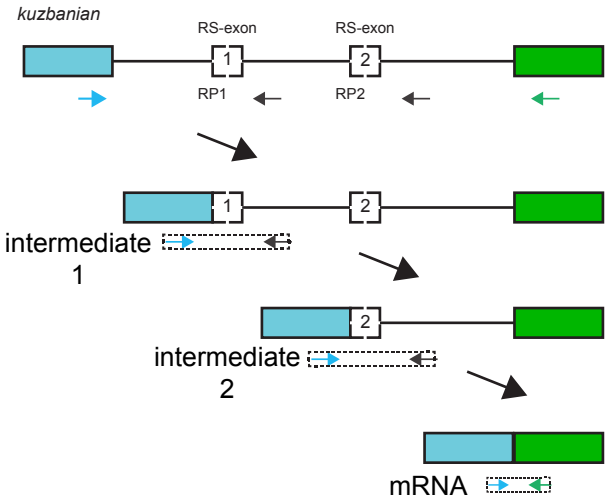
Again, I used a set of rt-PCR assays to inspect the molecular consequences of mutating the RP splice donor. Using the examples of RP 5'SS disruptions in *kuz*, *Bx* and *Ubx*, I previously showed that recursive splicing is constitutive (**Figure 2.4**) (Joseph et al., 2018). Therefore, *kuz* intron 3 is processed as three smaller fragments using two RPs (**Figure 3.2B**). I first examined the two obligate splicing intermediates that arise out of activation of RP1 and RP2 (**Figure 3.2A-B**). The first intermediate, which indicates processing of *kuz* RP1 (and mutant RP1), was unaffected by mutations to the 5'SS (**Figure 3.2C, Intermediate 1**). However, the second intermediate amplicon (indicating processing of *kuz* RP2), yielded an additional band from samples that had moderate to poor RP1-5'SS scores (**Figure 3.2C, Intermediate 2**). The additional product was longer

A



	RP	RP1 5'SS score	RS-exon 1 5'SS score
WT	AGGTGAGTGC	0.97	0.77
24	AGGTGAGTGC	0.97	0.77
14	AGGTGAGCTC	0.94	0.77
30	AGGTGAAAAC	0.55	0.77
26	AGGTAGTGCT	0.21	0.77
15	AGGTTTCTAA	0.21	0.77
16	AGGTCTTCTC	0.00	0.77
31	AGGTTTCTCT	0.00	0.77

B



C

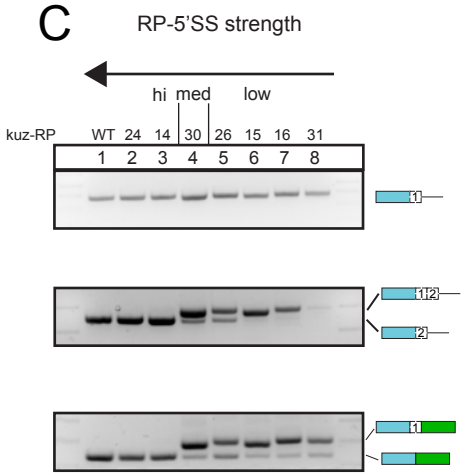


Figure 3.2. Weakening the *kuz* RP1 5'SS *in vivo* results in RS-exon inclusion.
(A) *kuz* gene models displaying two evenly spaced RPs wihin the ~50 kb intron 3. Precisely targeted mutations that alter the RP1 5'SS are listed below. The red color conveys match to wildtype RP1 5'SS whereas black indicates nucleotide changes. The allele ID is left of the sequence and changes to RP1 5'SS score on the right. The unchanged RS-exon 5'SS score is also included.
(B) A model for *kuz* intronic recursive splicing. PCR amplicons are displayed using dotted boxes and primers as arrows.
(C) Wildtype and RP1 5'SS mutants yield similar RP1 intermediate amplicons. However, differences can be observed for RP2 intermediate and mRNA amplicons. Conversion of the high scoring RP1 5'SS to a medium or low scoring SS results in cryptic exon inclusion in RP2 intermediates and mRNA. Interestingly, while RP2 intermediates display a steady conversion, from cryptic exon skipping to fully cryptic exon inclusion as the RP1 5'SS weakens, mRNA amplicons always yield a minor level of cryptic exon skipped products. As I have previously demonstrated that *kuz* recursive splicing is constitutive, the data suggests that weakened *kuz* RP1 5'SS can be activated to produce exon skipped products (see **Figure 3.3**)

in length than expected and verified to contain the addition of RS-exon 1 – a clear indication of RS-exon 1 5'SS usage instead of the RP1 5'SS. The inclusion of RS exon 1 in the second intermediate amplicon was found to increase with each stepwise decrease in RP1 5'SS strength and only began once the RP1 5'SS was significantly weaker than the RS-exon 1 splice donor (**Figure 3.2A-C, Intermediate 2** lanes 4-8). Together, these results also support RP 5'SS strength as a major determinant of RS-exon inclusion.

Lastly, I examined the molecular consequences of RP1 5'SS mutations on *kuz* mRNA. Here, the objective was to understand the conversion of the second intermediate into mRNA (**Figure 3.2B**). Under wildtype conditions, the RP2 5'SS outcompetes the RS-exon 2 5'SS, producing mRNA that skips RS-exon 2. In fact, rt-PCR of *kuz* mRNA confirms this for wildtype as well as mutants #s 24 (control) and 14, which still have strong RP1 5'SS and yield only canonical second intermediate (**Figure 3.2C, mRNA**, lanes 1-3). However, since lanes 4-8 (moderate to poor RP1 5'SS) contained RS-exon 1 inclusion in the second intermediate (**Figure 3.2C, Intermediate 2**), I wondered how this would affect downstream intron removal. Since these intermediates will contain three splice donors (RP1 5'SS, RP2 5'SS and RS-exon 2 5'SS) (**Figure 3.3**), my expectation was that the strongest SS would be activated most frequently. Of the three, RP2 5'SS is stronger than the RS-exon 2 5'SS, as well as mutant RP1 5'SS. Therefore, I expected the RS-exon-1-included second intermediates to be converted to RS-exon-1-included mRNA (**Figure 3.3**). This prediction was supported by rt-PCR tests that showed RS-exon 1 inclusion in mRNA (**Figure 3.2C, mRNA**, lanes 4-8). Surprisingly, while mutants #s 15, 16 and 31 only produced RS-exon-1-included second intermediates, a small fraction of these appear to get convert to RS-exon-1-skipped mRNAs (**Figure 3.2C, mRNA**, lanes 6-8). This suggests that the significantly weaker RP1 5'SS can also get activated during conversion to mRNA and hints that other factors may also regulate RS-exon inclusion. Overall, I provide the first *in vivo* evidence that 5'SS strength is a potent

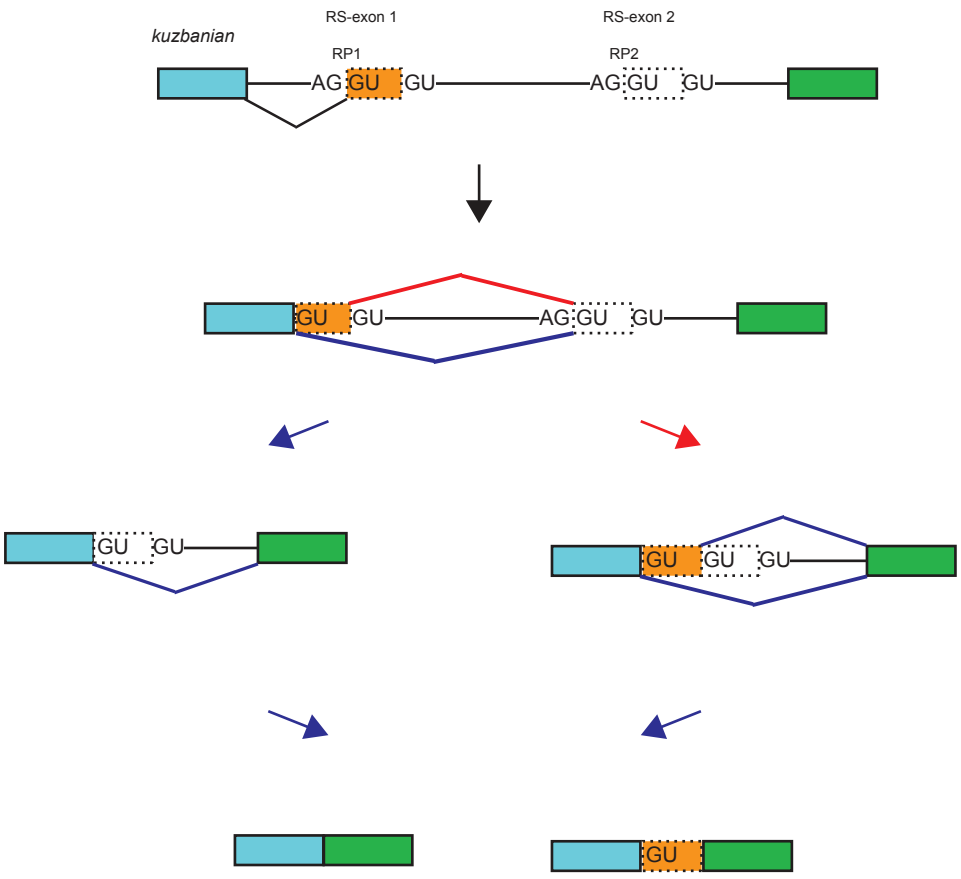


Figure 3.3. Intron removal trajectories that explain *kuz* RP2 intermediate and mRNA rt-PCR products from RP 5'SS mutants. The schematic depicts the activation of SS that yield the observed pre-mRNA intermediates and mRNA. The use of RP 5'SS is indicated with dark blue edges and arrows, whereas usage of the RS-exon 5'SS is indicated in red. When *kuz* RP1 5'SS is mutated to a poor splice site, the RS-exon 5'SS is activated (red) and the RP2 intermediate includes the cryptic RS-exon 1 (dotted orange box). However, in the next step (conversion to mRNA), one of the two remaining RP 5'SS is used to generate canonical mRNA or mRNA with cryptic RS-exon 1 retention. Suprisingly, weak and poor RP1 5'SS are also able to activate at the RP2 intermediate stage, despite the presence of two other strong 5'SS.

determinant of RS exon inclusion or skipping. Furthermore, as most RPs tend to have stronger regenerated 5'SS (**Figure 2.13D**), this largely results in RS-exon skipping.

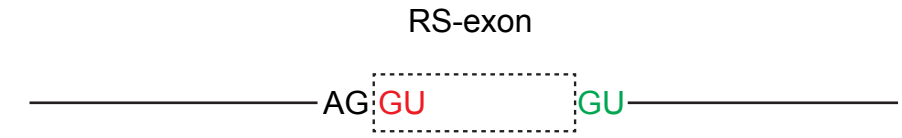
Cryptic RS- and RS-exon reporters display a range of alternative splicing in cell culture

Generating and testing RS mutants in live animals proved valuable, but for further dissection, I resorted to the more tractable *Drosophila* S2 cell culture system using my modular minigene RS reporter generated in Chapter 2. The design of this reporter allowed an easy one-step cloning of RS loci into a minimal splicing minigene (see Chapter 2, Methods), and was previously used to test *kuz* and *Bx* RS, both of which predominantly skipped the RS-exon. To distinguish other mechanisms of RS-exon alternative splicing, it seemed reasonable to first identify examples of RPs with differential RS-exon inclusion. Therefore, I cloned a set of seven cryptic RS-exons and eight RS-exons into the splicing backbone (**Figure 3.4**). Additionally, in order to mimic the architectural properties of RS, the regions cloned spanned ~3 kb flanking each RS-exon (**Figure 3.5A**).

Expression of all seven cryptic RS-exon reporters predominantly yielded the expected product in which the RS-exon was skipped (**Figure 3.5B**). This was the case for reporters from genes including *chinmo, Egfr, shep, Ubx, ct* and *nmo,* all of which had stronger RP 5'SS compared to RS-exon 5'SS (**Figure 3.4**). Interestingly, the *homothorax (hth)* RS reporter also yielded an exon skipped product despite having a substantially weaker RP 5'SS compared to RS-exon 5'SS (**Figure 3.5B**). In three of seven instances (*ct, Ubx* and *nmo)* it was possible to detect a small proportion of RS-inclusion, but this was not related to difference in 5'SS scores (**Figure 3.4 and Figure 3.5B**). In theory, the RS-exon skipped products could be obtained through exon skipping rather than RS. To

account for this possibility, I generated mutant versions of two RS reporters (*ct-RP and Ubx-0nt-RP)* in which the RP 5'SS were disrupted (**Figure 3.6A**). Under conditions of exon skipping, such mutations should not alter the reporter products. However, if spliced via RS, the mutant reporter should have constitutive inclusion of the RS-exon (**Figure 3.6A**). Both mutant reporters display a switch from exon skipped to included (**Figure 3.6B**). Therefore, I conclude that cryptic RS-exon reporters yield skipped products via recursive splicing.

Next, I examined the products of expressed RS-exon reporters (**Figure 3.4**). Here products revealed different proportions of RS-exon inclusion and skipped amplicons (**Figure 3.5C**). For genes *sm, heph* (RP2) and *mub*, the dominant amplicon was the exon skipped product, while *Ubx* (m1) and *msi* yielded mostly the exon included product. The remainder, reporters of *ps, fra and heph* (RP1) produced an even proportion of skipped and included amplicons (**Figure 3.5C**). Remarkably, the predicted RS-exon 5'SS was activated in most cases with RS-exon inclusion, the only exception being the *pasilla (ps)* reporter, which in addition to the predicted RS-exon 5'SS, also activated a weak 5'SS ~170 nt downstream. As with the cryptic reporters, I verified that RS was the basis for the observed AS (**Figure 3.6C**). Interestingly, comparison of 5'SS strengths revealed that seven out of eight reporters in this category have stronger RP-5'SS (**Figure 3.4**). The *Ubx* (m1) and *msi* reporters are particularly noteworthy as these mostly yield the exon inclusion product despite having stronger RP 5'SS. Overall, these experiments i. provide a set of fifteen RS reporters and ii. clearly indicate that other mechanisms may also regulate RS-exon inclusion.



RP name	RP 5'SS score	RS-exon 5'SS score	
chinmo-RP	0.95	0.91	Cryptic RS-exons
Egfr-RP	0.97	0.69	
shep-RP	0.98	0.98	
Ubx-0nt-RP	0.97	0.90	
ct-RP	0.96	0.91	
hth-RP	0.60	0.98	
nmo-RP	0.98	0.96	
sm-RP	0.98	0.85	RS-exons
heph-RP2	0.92	0.82	
mub-RP	0.94	0.67	
ps-RP	0.96	0.93	
fra-RP	0.95	0.96	
heph-RP1	0.99	0.94	
Ubx-m1-RP	0.96	0.90	
msi-RP	0.92	0.85	

Figure 3.4. 5'SS quantifications for RS substrates tested in cell culture. Top: schematic of an intronic RS-exon, with the RP 5'SS in red and the RS-exon 5'SS in greeen. Below: the names and 5'SS scores for 7 cryptic RS-exons and 8 expressed RS-exons cloned into cell culture reporters. Quantifications were performed using the NNSPLICE algorithm. Boxed candidates display unexpected RS-exon alternative splicing based on 5'SS scores (see **Figure 3.5**).

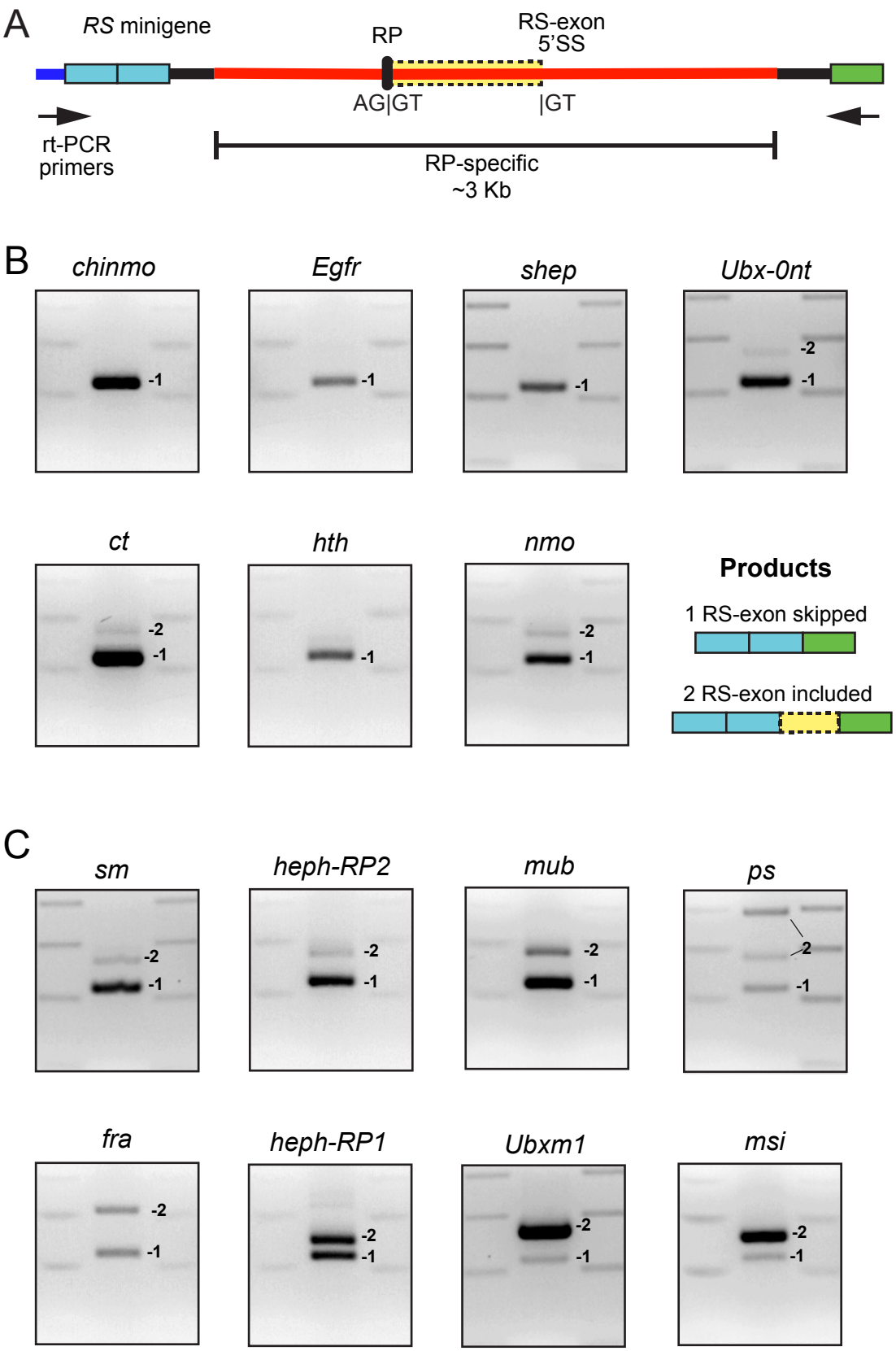

A
RS minigene
RP
RS-exon
5'SS
AG|GT
|GT
rt-PCR
primers
RP-specific
~3 Kb
B
chinmo
Egfr
shep
Ubx-0nt
ct
hth
nmo
Products
1 RS-exon skipped
2 RS-exon included
C
sm
heph-RP2
mub
ps
fra
heph-RP1
Ubxm1
msi

Figure 3.5. Cloning and testing of 15 RS splicing minigene reporters.
(A) A schematic of the splicing minigene. The detailed construction of this reporter is listed in Chapter 2. In brief, ~3 kb of intronic fragment (red) containing RS-exons was cloned into the intron (black) of the *kuz* minimal splicing minigene.
(B) rt-PCR of splicing reporters containing cryptic RS-exons. For all seven substrates tested, I observe the expected exon skipped amplicon as the major product.
(C) rt-PCR of splicing reporters containing expressed RS-exons. A range of RS-exon inclusion levels can be observed for these RS substrates. Notably, some do not match expectations based on 5'SS scores (see **Figure 3.4**). For instance, *msi* and *Ubxm1* are dominantly included despite having weaker RS-exon 5'SS

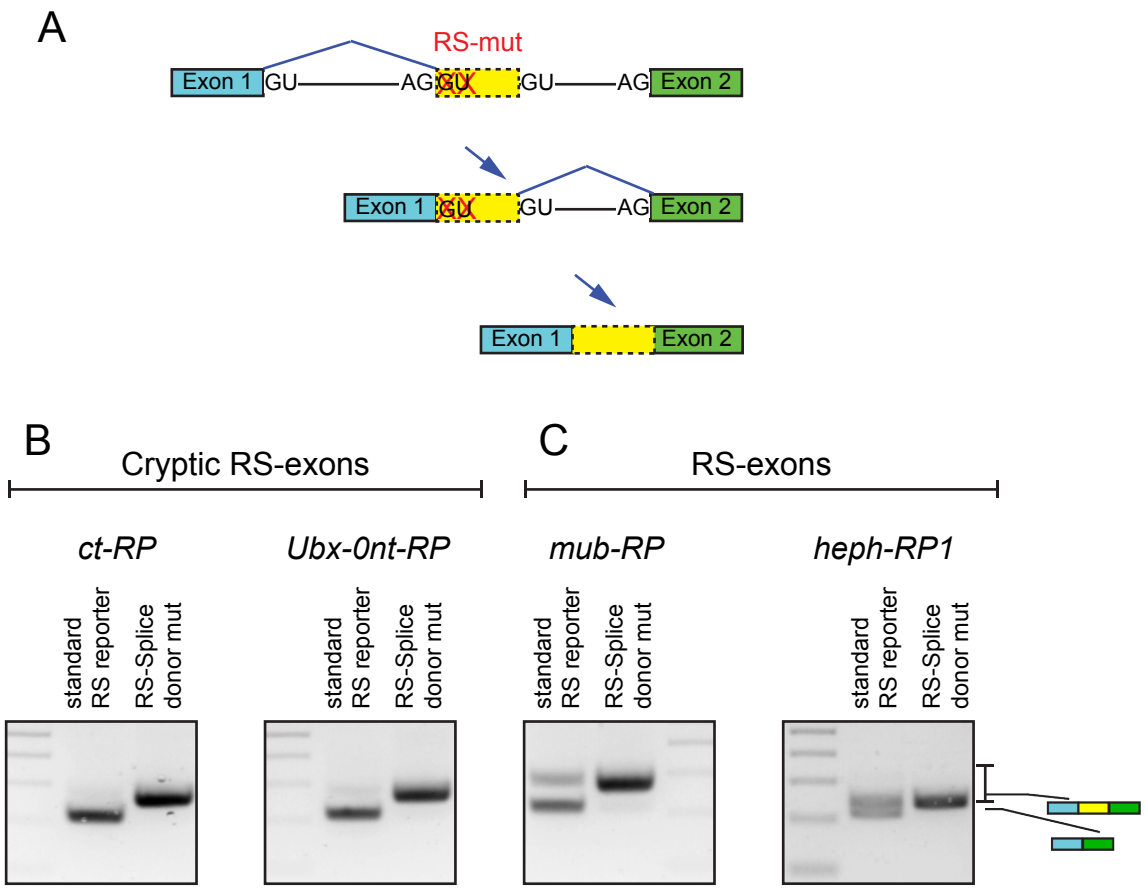


Figure 3.6. Demonstration of recursive splicing in minigene splicing reporters.
(A) Schematic of the recursive splicing pathway under RP 5'SS disruption. Critically, the skipped cryptic RS-exon is converted to constitutively included.
(B) As in Chapter 2, RS-5'SS mutations in cryptic RS-exon substrates leads to complete inclusion of the RS-exon in mRNA.
(C) Similarly, expressed RS-exons display the same logic as in (B), indicating they are also spliced via recursive splicing.

Exonic SREs regulate RS-exon alternative splicing

Several mechanisms of regulated alternative splicing have been reviewed in the introduction to this thesis. In fact, some of these ideas, such as the roles of *trans*-acting factors and histone modifications have been previously considered (Duff et al., 2015). Since my RS reporters only differ in the content of intronic RS sequence but lead to different processing, I first tested the possibility that these could be the effects of splicing regulatory elements found within the reporter. Typical SREs are found either within exons, or proximal to exons within introns. However, as the introns in these constructs are fairly long, I decided to first examine exonic sequence.

Constitutively expressed exons are thought to contain an abundance of conserved exonic splicing enhancers (ESEs) (Zefeng Wang & Burge, 2008). Therefore, I first examined RS-exons for their conservation patterns. Most cryptic RS-exons are poorly conserved, but RS-exons with coding potential demonstrate higher conservation (**Figure 2.9B**). Through sequence gazing, I noted that the known and expressed *Ubx* microexons (m1 and m2), and the RS-exon from *smooth (sm)* are deeply conserved, so these appeared good candidates for further evaluation. Of these three, I had RS reporters for *Ubx* (m1) and *sm*, but as the *sm* RS-exon is not abundantly included in S2 cells (**Figure 3.5C**), I restricted my attention to the *Ubx-m1* reporter (**Figure 3.7A**).

All 51 nt of the m1 exon are ultraconserved (including the wobble position of all 17 amino acids codons), suggesting that information beyond the coding potential is under strong selection (Bomze & López, 1994). Therefore, one method to examine if RS-exons contains important regulatory information is to selectively mutate its contents. Based on the motif preferences of *Drosophila* serine/arginine (SR) proteins that bind ESE elements (Bradley et al., 2015), I made a set of spaced nucleotide substitutions to the m1 RS-exon. As some SR proteins prefer guanosine rich elements, I substituted several central guanosines to either adenosine or thymidine (**Figure 3.7B**, Gdep). These

sequences have previously been mutated in a *Ubx* minigene splicing reporter from the Lopez laboratory and shown to influence m1 inclusion and therefore serve as a good control (Hatton et al., 1998a). With these mutations, the *Ubx-m1* RS-exon reporter was converted to predominantly m1 skipping (**Figure 3.7C**, lanes 1 vs 5), clearly indicating that exonic *cis*-elements can regulate RS-exon alternative splicing. Moreover, I wondered if the sequence content in RS-exons was sufficient to explain the behaviors of RS-exon reporters in cell culture. I chose RS-exons with extreme behaviors, showing either full inclusion (*Ubx-m2* RS-exon) or none (*Ubx-0nt* and *chinmo* cryptic RS-exons) and swapped these into the *Ubx-m1* reporter (**Figure 3.7B**). Importantly, these reporters contain the same flanking exons and the ~3 kb of intronic sequence as the unmodified *Ubx-m1* reporter, but selectively swapped the RS-exon sequence. Furthermore, these sequence swaps alter the RP-5'SS but not the RS-exon 5'SS as this lies in the intron. Remarkably, the modified reporters behaved in accordance with the RS-exon swap. For example, the *Ubx-m2* RS-exon swap yielded predominantly exon inclusion (**Figure 3.7C**, lane 2). In stark contrast, the *Ubx-0nt* and *chinmo* RS-exon swaps produced exon skipping (**Figure 3.7C**, lanes 3 and 4). Since all RS-swaps maintained the stronger RP 5'SS (**Figure 3.4**), these results argue that elements within the RS-exon are additional determinants of RS alternative splicing.

A prediction based on this model is that swapping an expressed RS-exon into a skipped RS-exon reporter should result in a switch to exon inclusion. To test this out, I utilized my *Ubx-0nt* reporter, which is predominantly skipped (**Figure 3.8A**). The RS-exon in this reporter was swapped with those that are fully included (*Ubx-m1* and *Ubx-m2* RS-exons), or completely skipped (*chinmo* RS-exon) (**Figure 3.8B**). Surprisingly, for this reporter as well, the modifications mirrored the known behaviors of the swapped RS-exons. While *chinmo* RS-exon was mostly skipped, *Ubx-m1* produced a switch to an even proportion of both products. Meanwhile, the *Ubx-m2* swap yielded a complete

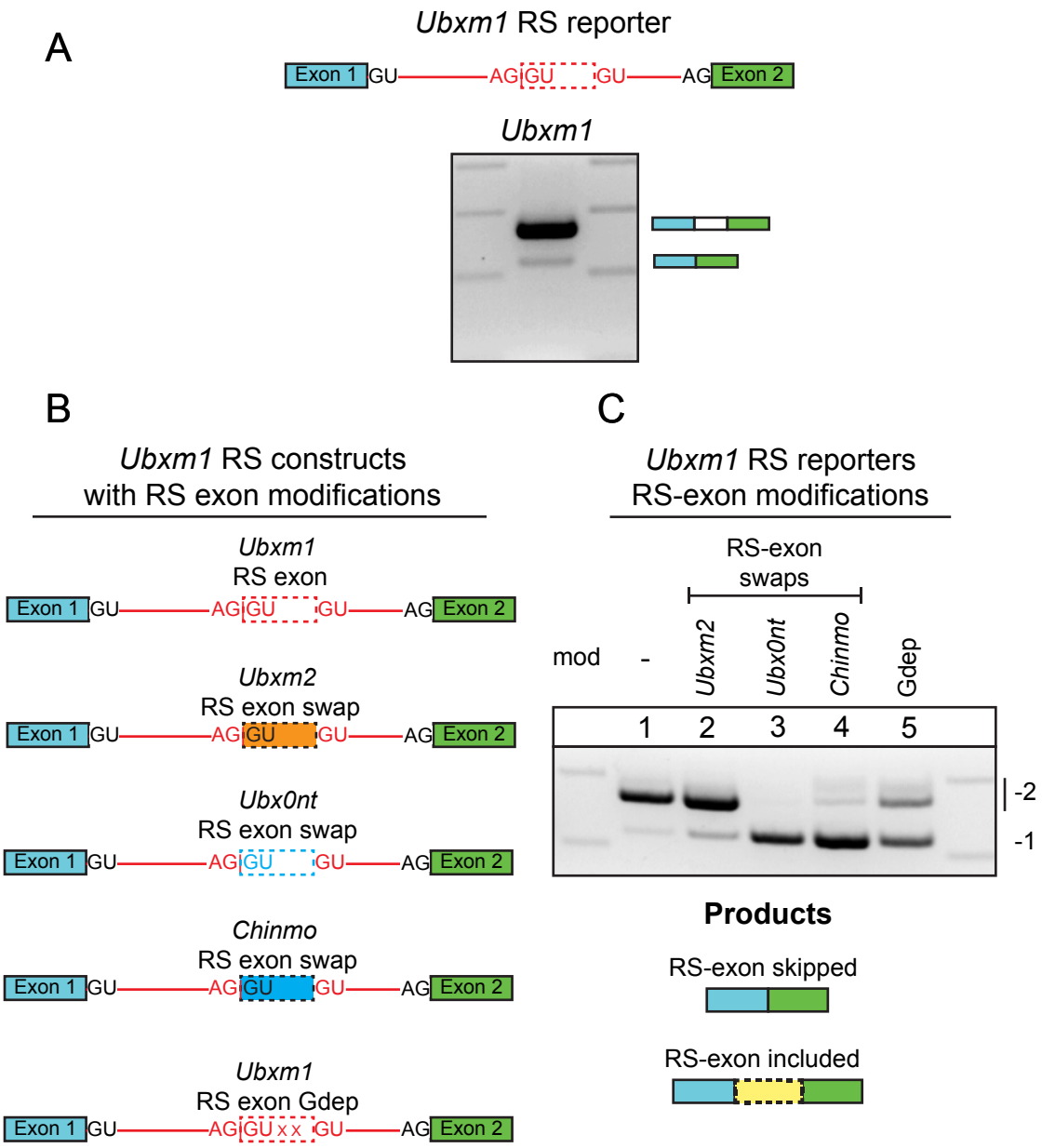


Figure 3.7. RS-exon swap in *Ubxm1* alters RS-exon inclusion levels.
(A) Top: Schematic of the *Ubxm1* reporter. *Ubxm1* specific intronic sequence in red. Bottom: This reporter mostly yield RS-exon inclusion.
(B) Schematic of modifications to the *Ubxm1* reporter. Only the *Ubxm1* RS-exon portion of the reporter was swapped with the RS-exons of *Ubxm2, Ubx0nt* or *Chinmo.* These changes are displayed via the limited changes to the RS-exon on the original reporter. Central G/C nucleotides on the RS-exon were substituted with A/T to generate the Gdep RS-exon construct.
(C) RS-exons seem to contain information regulating RS-exon AS. Swapping the *Ubxm1* RS-exon with others mimics their inclusion or skipping behaviors. Moreover, the Gdep reporter indicates that short RNA elements (SREs) may regulate the cassette exon properties of RS-exons.

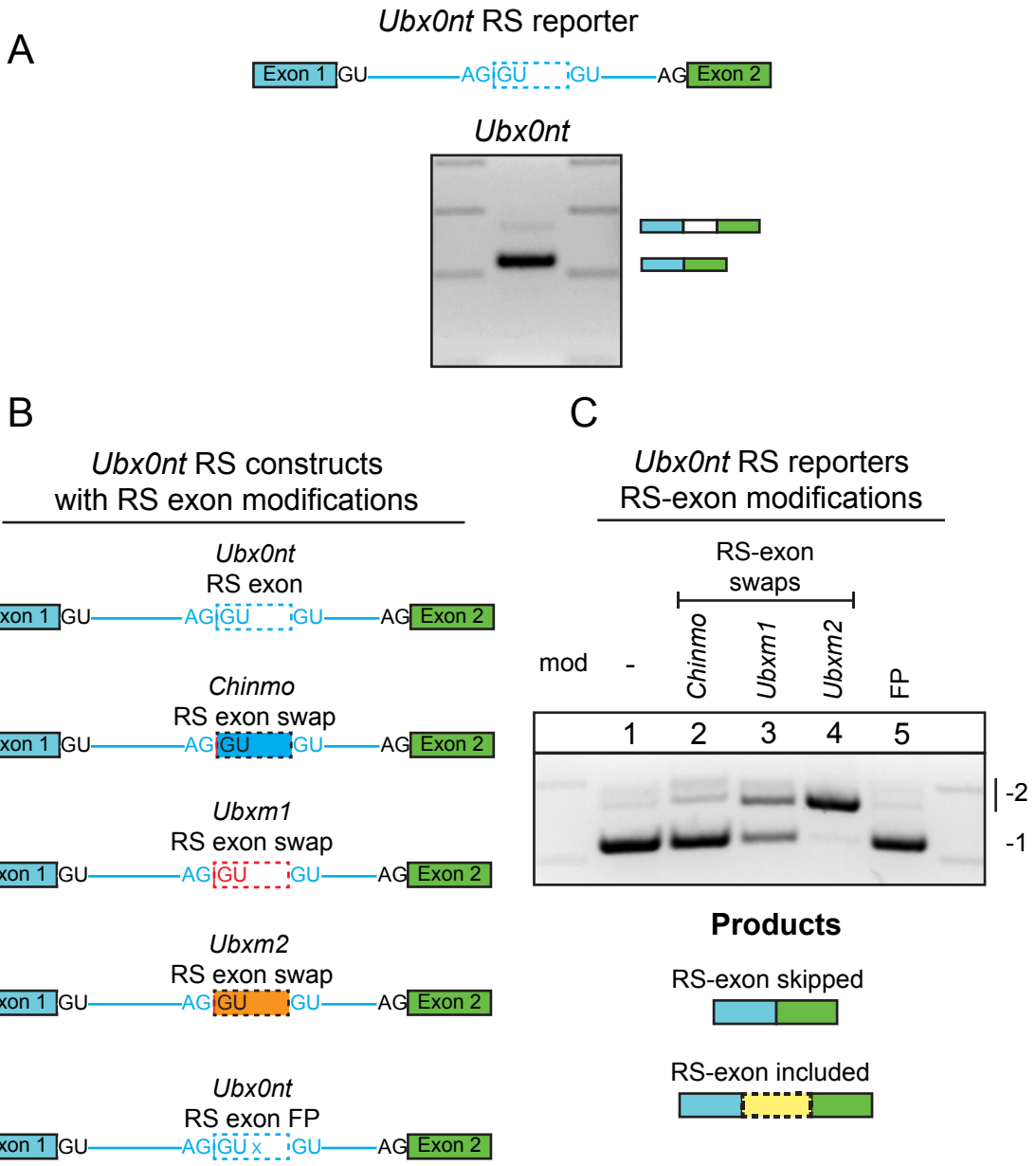


Figure 3.8. RS-exon swap in *Ubx0nt* alters RS-exon inclusion levels.
(A) Top: Schematic of the *Ubx0nt* reporter. *Ubx0nt* specific intronic sequence in blue. Bottom: This reporter mostly yield RS-exon skipping.
(B) Schematic of modifications to the *Ubx0nt* reporter. Only the *Ubx0nt* RS-exon portion of the reporter was swapped with the RS-exons of *Ubxm2, Ubxm1* or *Chinmo.* These changes are displayed via the limited changes to the RS-exon on the original reporter. The *Ubx0nt* RS-exon size was converted to a frame preserving length (FP).
(C) RS-exons seem to contain information regulating RS-exon AS. Swapping the *Ubx0nt* RS-exon with others mimics their inclusion or skipping behavior-sproperties. The FP reporter is largely exon-skipped, suggesting that mRNA stability is not a major confounding factor.

switch to exon included (**Figure 3.8C**, lanes 1-4). Notably, a longer, minor product can be observed for both reporters (**Figure 3.7C** and **Figure 3.8C**). This is due to the unexpected activation of weak 5'SS downstream of the annotated RS-exon.

An important caveat in these experiments is that both *Ubx* m1 and m2 RS-exons are frame preserving (51 nt, each), whereas the *Ubx-0nt* and *chinmo* RS-exons are not (53 and 56 nt). Therefore, the results of the RS-exon swap experiments could also be explained by differences in stability of cryptic RS-exon vs RS-exon inclusion mRNA. To assess this possibility, I modified the *Ubx-0nt* reporter to make the RS-exon frame preserving (**Figure 3.8B-C**, FP – 54 nt). As this reporter was still skipped, I conclude that exonic SREs regulate RS-exon inclusion.

The Exon Junction Complex may stimulate RS-exon inclusion

The EJC is deposited ~20-24 nt upstream of exon junctions during the splicing reaction (Schlautmann & Gehring, 2020). If RS is similar to canonical splicing, removal of the upstream intron fragment should deposit the EJC ~20-24 nt upstream of the RP 5'SS. Hence it is quite reasonable to consider if this complex may regulate RS. Two sources of evidence suggest a likely relationship. Firstly, in *Drosophila*, the EJC has been reported for its role in the accurate processing of long introns (Ashton-Beaucage et al., 2010; Roignant & Treisman, 2010), and otherwise in the regulation of splice site activation (Hayashi et al., 2014; Malone et al., 2014). Secondly, in the mammalian system, the Ule and Gehring labs recently demonstrated that the EJC suppresses RS on constitutive exons (Blazquez et al., 2018; Boehm et al., 2018). Therefore, I sought to examine how the EJC may influence *Drosophila* intronic RS.

I selected a set of four reporters that yielded a range of RS-exon inclusion (low to high). In order to precisely model loss of EJC recruitment on these reporters, I deleted

the upstream intron segment 1 (**Figure 3.9A**, ∆ intron segment 1). Deletion of the intron segment mimics the RS-intermediate pre-mRNA without actually undergoing the splicing reaction, so these reporters are not expected to recruit the EJC. All four deletion constructs displayed an overall increase in RS-exon skipping (**Figure 3.9B**). The *sm* and *heph* pre-spliced reporters, in fact, only produced the exon skipped amplicon. Moreover, the *Ubxm1* and *msi* reporters (normally included), also yielded predominantly skipped products (**Figure 3.9B**).

To examine if the EJC regulates cryptic RS-exons, I also generated a pre-spliced version of the *Ubx-0nt* reporter. Deletion of the upstream intron in this reporter had no discernable effects in comparison to the unmodified construct (**Figure 3.9C**). Overall, these results suggest that the EJC also influences RS-exon AS. Potential mechanisms are discussed following the results section.

in vivo deletion of intronic RPs and RS-exons in *Drosophila*

So, what is the function of intronic recursive splicing? Due to their deep conservation, the current view holds that intronic RPs must be important for host gene expression. However, no functional tests have been performed on recursive splicing. My mutagenesis of RPs represents the only reported tests on recursive splicing in intact animals (Chapter 2, Joseph et al., 2018). However, only RP splice donors were disrupted in this study, as well as the experiments in **Figures 3.1 and 3.2**. Therefore, while partial disruptions can lead to alternative splicing (Chapter 2), I wondered about the consequences of complete RP disruption.

Initially, I chose a set of four RPs within as many genes. As partial mutations of *kuz* RP1 and *Ubx-0nt* RP had strong molecular and developmental phenotypes (Chapter 2), these SS were first on my list for complete deletion. However, the disadvantage of

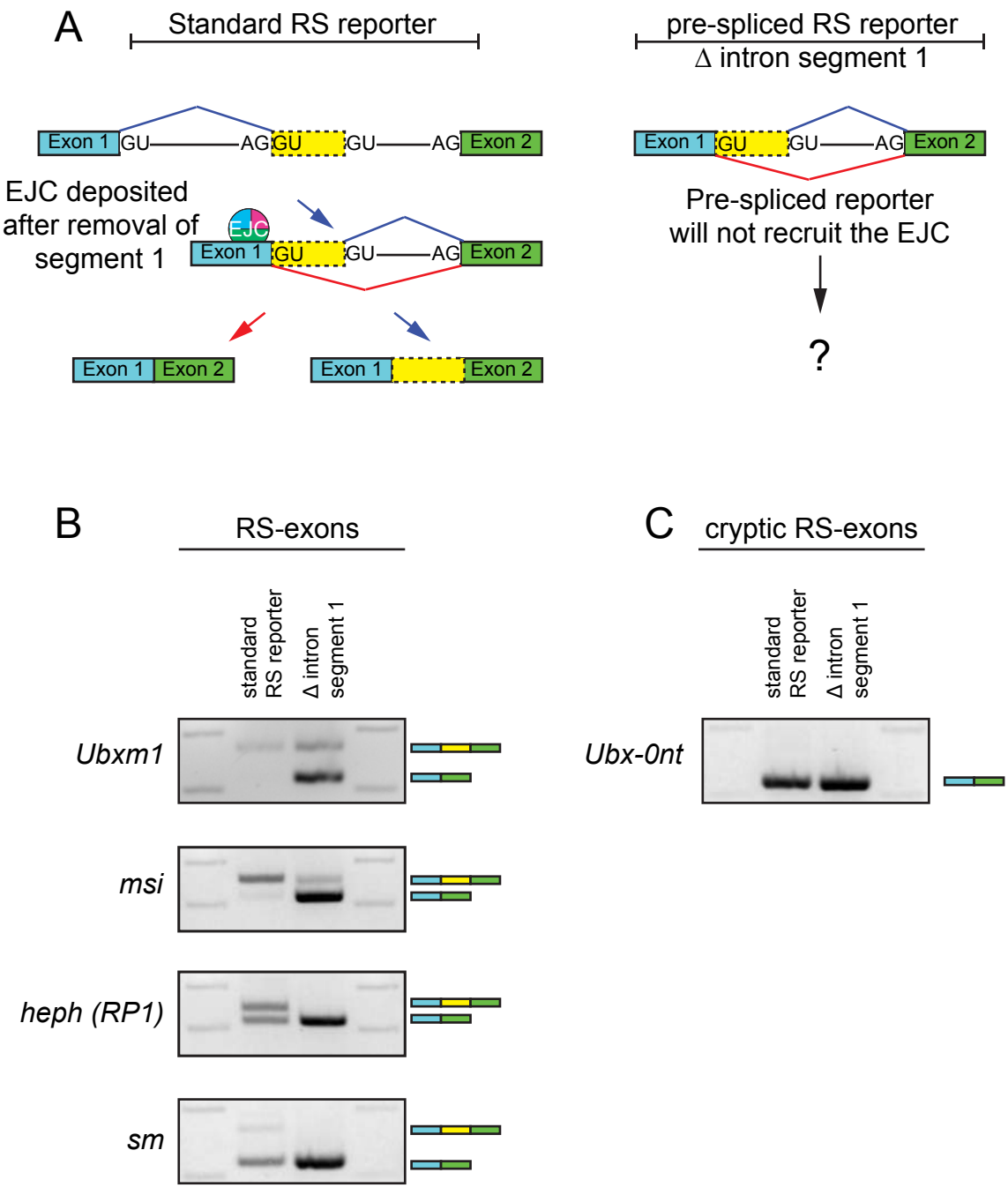


Figure 3.9. Intron pre-removal causes RS-exon skipping. (A) Left: model of RS-exon splicing, including the deposition of the EJC after removal of intron segment 1. Right: schematic of pre-spliced RS reporters. These will not recruit EJC prior to removal of intron segment 2. (B) Pre-spliced RS-exon reporters display higher levels of exon skipping. (C) Cryptic RS-exon reporter from *Ubx-0nt* that is normally skipped is unaffected by pre-removal of intron segment 1.

selecting these genes is that they contain multiple RPs, so a single deletion will still leave the host introns with at least one more RP. Therefore, I chose to make deletions in *Epidermal growth factor receptor (Egfr)* and *dachsous (ds),* which have just one RP each. Like *kuz* and *Ubx,* the RPs in these genes are hosted within unusually long introns (**Figure 3.10**), so they seemed ideal substrates. Moreover, it was quite relevant that these genes are important signaling components (González-Morales et al., 2015; Malartre, 2016; Saavedra et al., 2016) and offered the opportunity to examine effects of RP deletions in multiple developmental contexts.

I used a transgenic Cas9 system with two or four guide RNAs (see Methods) to delete intronic RPs in animals. Use of the same combination of targeting guides was employed for *kuz* RP1, and screening of 40 animals via a PCR assay yielded two mutants with short deletions that included 5' and 3'SS components of the RP (**Figure 3.10A**). Unlike the partial mutants previously reported (Joseph et al., 2018), complete disruptions of RP1 had no consequences on animal viability.

Similarly, deletions of the *Egfr, ds* and *Ubx* RPs were obtained by PCR screening 8, 30 and 120 animals respectively (**Figure 3.10B-D**). While $Egfr^{\Delta RP}$ *and* $ds^{\Delta RP}$ were homozygous viable and lacked overt defects, $Ubx^{\Delta RP}$ was lethal. The cause appears to be an off-target lesion, as $Ubx^{\Delta RP}$ complements known amorphic *Ubx* alleles (Ubx^{1} and $Ubx^{6\text{-}28}$) (Weinzierl et al., 1987).

Molecular characterization of RP deletion mutants

I sought to examine RNA processing of the mutated genes. Of note, previous partial RP mutagenesis did not disrupt recursive splicing, but instead resulted in the inclusion of a short cryptic exon in mRNA. rt-PCR analyses to detect the intermediate amplicon (**Figure 3.11A**) downstream of the RPs yielded no products from each mutant

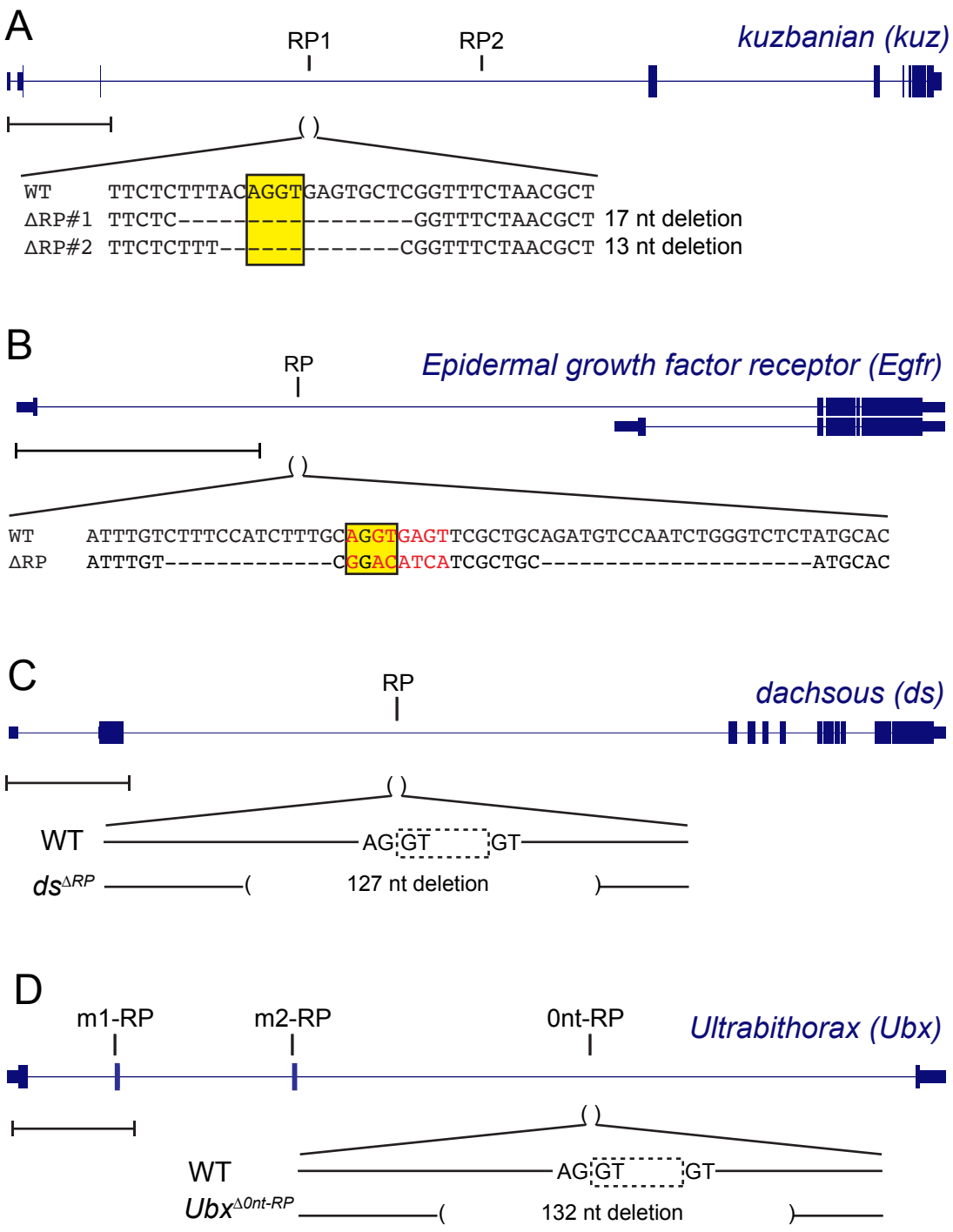


Figure 3.10. RPs and RS-exons deleted using a transgenic CRISPR/Cas9 approach. Gene models of disrupted loci along with the locations of RPs. Scale bar indicates 10 kb. Descriptions of the RP mutations are indicated below for (A) *kuz*, (B) *Egfr*, (C) *ds* and (D) *Ubx*. For *kuz* and *Egfr*, the RP but not the entire cryptic RS-exon was disrupted. However, the entire cryptic exon was removed in *ds* and *Ubx*

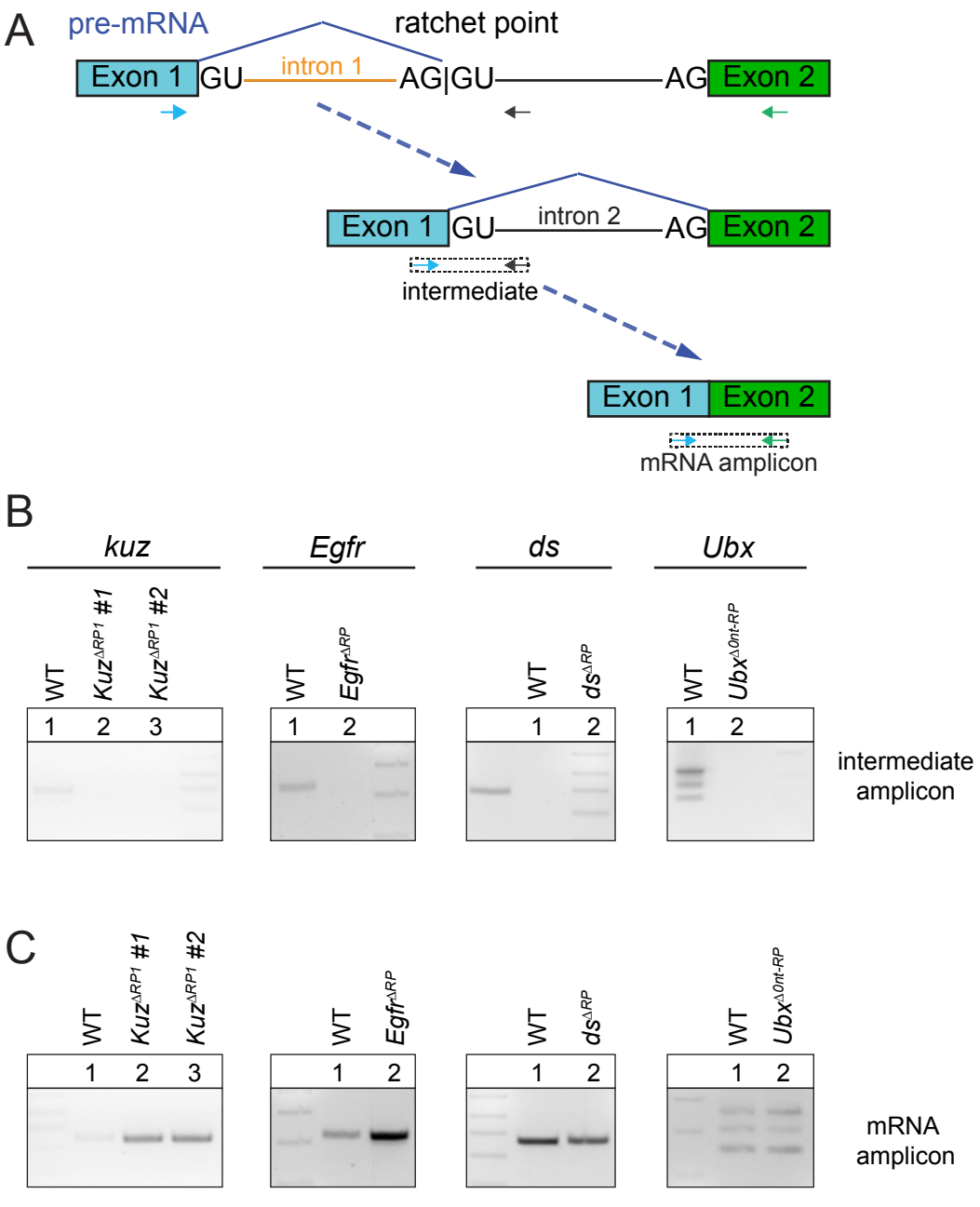


Figure 3.11. Molecular evalutation of RNA splicing in RP deletion mutants.
(A) Model for intronic recursive splicing. PCR amplicons are displayed using dotted boxes and primers as arrows
(B) Wildtype but not ΔRP mutants produced intermediate amplicons, indicating that RP deletions successfully disrupt recursive splicing.
(C) Wildtype and ΔRP mutants produced correctly spliced mRNA amplicons, suggesting that disruption of recursive splicing has no consequence on host gene intron removal.

(**Figure 3.11B**). This indicated that deletion of the RP successfully abolished recursive splicing.

Despite loss of recursive splicing, mature mRNA seemed correctly processed in all four mutants (**Figure 3.11C**). This was found for genes with one (*Egfr* and *ds)* or more RPs (*kuz* and *Ubx).* Moreover, for *Ubx,* which produces three mRNA isoforms, loss of the 0nt-RP did not seem to alter the ratio of alternatively spliced transcripts (**Figure 3.11C**). Of note, these data reflect patterns found in total RNA from adult female flies, so it is still unclear if intron processing remains unaffected under all cellular contexts. I conclude that the deleted RPs appear generally dispensable for host gene intron removal.

Using Cas9-mediated homologous recombination to induce double RP deletions in *mbl*

RPs might appear dispensable for long intron removal if there are other cryptic SS that can compensate for their loss. Since it is not possible to determine all possible cryptic SS without extensive validation, I resorted to making more deletions of annotated recursive splice sites within extreme cases of RS. I selected *muscleblind (mbl)* as a potential target, as the 75070-nt-long intron 2 contains four RPs. *mbl* is an essential gene in *Drosophila* with important roles during development and differentiation. The gene encodes a zinc finger containing RBP that is involved in diverse RNA metabolism (Irion, 2012). Moreover, this locus displays curious alternative splicing, and accumulates the most abundant fruit fly circular RNA from the exon directly upstream of intron 2 (Westholm et al., 2014). The combination of length, ample back-splicing and recursive splice sites suggests that RNA processing at this locus might be tightly regulated by several mechanisms and have important consequences on gene output. Therefore, I

decided to test the necessity of RS for *mbl* function. However, as there are four RPs within the same intron, I redesigned my mutagenesis strategy to allow creation and isolation of *cis*-double deletion mutants.

The inefficient aspect of my mutagenesis approach was that only one deletion could be induced per experiment, and the low throughput PCR screening assay. To improve the latter, I used the same CRISPR/Cas9 mutagenesis, but with homology directed repair (HDR) to replace RPs with fluorescence markers (Gratz et al., 2014), so that transformants could be initially identified through visual screening. The readily available HDR vector pHD-dsRED contains the 3xP3 driven dsRED marker which can be detected in the fly visual system. I had access to another HDR reporter that was recently developed in the Lai laboratory, in which the 3xP3-dsRED cassette was replaced with a ubiquitous GFP marker. Once inserted, the marker cassettes can be excised efficiently as they are flanked by FRT sites. Lastly, the GFP HDR template also contained an attP site for easy future reinsertion of DNA sequences (Bischof et al., 2007; Groth et al., 2004).

I attempted to use these two discernable HDR templates to make double RP deletions in *cis* (see Methods). *nos*>Cas9 expressing embryos were injected with sets of guide RNAs intended to induce deletions at two RPs. The injection mixture also contained dsRED and GFP HDR templates that were directed at the RPs. F0 flies that survived to adulthood were crossed to a second chromosome balancer and progeny were screened for makers. A successful double RP deletion should precisely replace targeted RPs with dsRED and GFP. Therefore, the animals will have both GFP and dsRED fluorescence.

Two injections were performed. In the first, I attempted to remove *mbl* RP1 and RP3, and replace with dsRED and GFP. In the second, I attempted to delete *mbl* RP2 and RP4, and replace with GFP and dsRED. I observed dsRED+ and GFP+ singly

marked animals and these were molecularly verified to precisely replace RPs, indicating that the experimental design can be used to generate RP deletions. However, despite screening over 5000 animals, I did not observe any GFP+, dsRED+ doubly marked animals, suggesting that double RP deletion and cassette replacement was not an efficient process. Sequencing of the unmarked loci indicated *cis* indels, but the RPs were not disrupted or deleted.

Overall, I generated single deletions for three out of four RPs (including the RS-exon) in *mbl* intron 2 (**Figure 3.12A**). RP2 and RP3 were replaced with GFP, whereas RP4 was replaced with dsRED. Known *mbl* loss-of-function mutants have severe phenotypes, including homozygous lethality. However, my intronic mutants did not display and overt defects.

mbl RP mutants exhibit normal mRNA splicing

I characterized the molecular consequences of mutating intronic RPs in *mbl.* Based on the short cryptic exon model of recursive splicing (Joseph et al., 2018), there are four obligate intermediates that will be produced during the removal of the second intron. Each of these intermediates were evaluated by rt-PCR for wildtype animals as well as the three RP deletion mutants (**Figure 3.12B**). The deletion of intronic RPs lead to specific loss of the respective intermediates for all three mutants (**Figure 3.12C**). Interestingly, an unanticipated, longer RP3 intermediate was observed in ΔRP3 animals (**Figure 3.12C**). This was found to be due to splicing of exon 2 to a strong 3'SS (NNSPLICE score of 0.94) on the antisense of the ubiquitin promoter (**Figure 3.13**). Of note, despite GFP cassette insertion in both ΔRP2 and ΔRP3 mutants (**Figure 3.12A**), the unexpected 3'SS was only activated in ΔRP3 animals (**Figure 3.12C**, lane 2 and 3 – RP2 and RP3 intermediates).

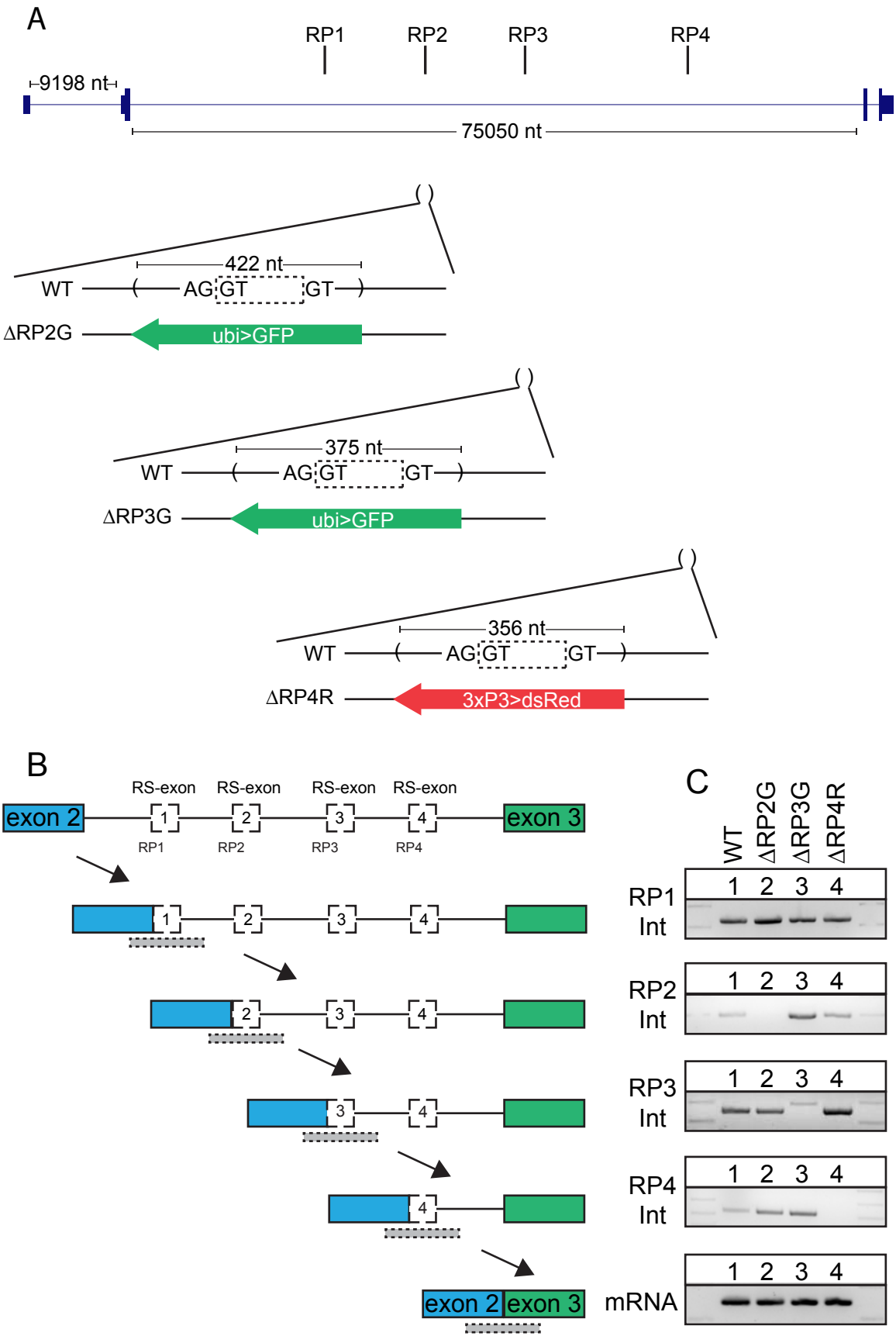

A
RP1
RP2
RP3
RP4
9198 nt
75050 nt
422 nt
WT
AG
GT
GT
ΔRP2G
ubi>GFP
375 nt
WT
AG
GT
GT
ΔRP3G
ubi>GFP
356 nt
WT
AG
GT
GT
ΔRP4R
3xP3>dsRed
B
RS-exon
RS-exon
RS-exon
RS-exon
exon 2
exon 3
1
2
3
4
RP1
RP2
RP3
RP4
exon 2
exon 3
C
WT
ΔRP2G
ΔRP3G
ΔRP4R
RP1 Int
RP2 Int
RP3 Int
RP4 Int
mRNA
1 2 3 4

Figure 3.12. Deletion of *mbl* RP2, RP3 and RP4 does not alter mRNA production.
(A) Top: gene model displaying *mbl* along with the location of four evenly spaced intronic RPs within the ~75 kb intron 2. Below: schematics of RPs replaced with fluorescence markers. 422 nt flanking RP2 was replaced with ubi>GFP in the reverse oreintation. 375 nt flanking RP3 was also replaced with ubi>GFP in the reverse oreintation. Finally, 356 nt RP4 was replaced 3XP3>dsRED in the reverse orientation. “G” in ΔRP2G and ΔRP3G and “R” in ΔRP4R, indicate the GFP and dsRED markers.
(B) A model for *mbl* intronic recursive splicing involving five splicing intermediates. PCR amplicons are displayed using dotted boxes and primers as arrows. These amplicons were tested in the mutants in (C).
(C) rt-PCR analyses of the four recursive intermediates and the mRNA amplicon in wildtype and *mbl* intronic mutants. While the RP1 intermediate is unaffected in all three mutant conditions, the RP2 and RP4 intermediates are only lost in Δ RP2G and ΔRP4R respectively. A longer RP3 intermediate product is observed in ΔRP3G animals and involves the splicing of *mbl* exon 2 to the antisense of ubiquitin promoter (**Figure 3.13**). Despite the various changes in splicing intermediates, canonical mRNA amplicons are produced and indicate that removal of intron 2 is not inhibited by RP loss.

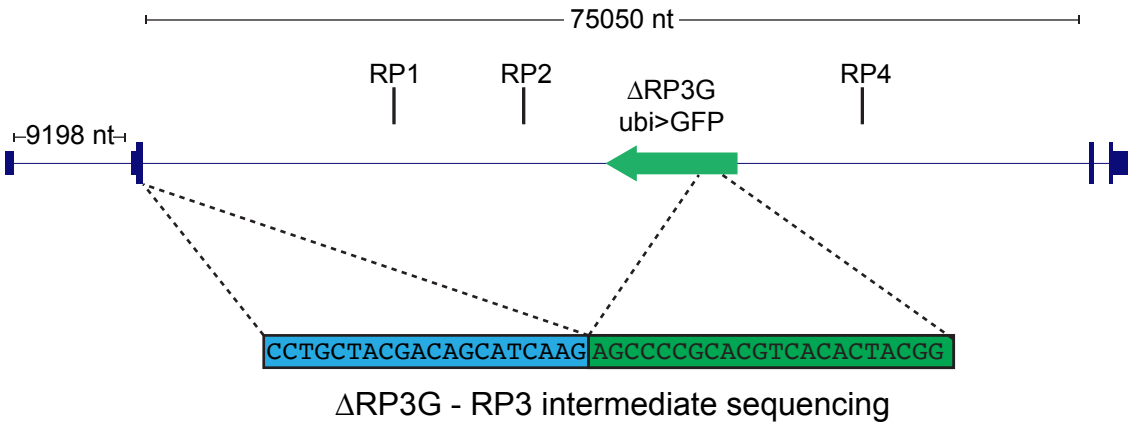


Figure 3.13. Exon 2 ligates to the antisense of the ubiquitin promoter. Alignment of the RP3 intermediate amplicon from ΔRP3G to the mutant *mbl* locus

Consistent with the previous RP deletions, mature mRNA amplicons were unaffected in all three *mbl* RP mutants (**Figure 3.12C**). Even for ΔRP3 which displayed spurious splicing (**Figure 3.12C**, RP3 intermediate), the mature product correctly ligated exons 2 and 3. Overall, these results are consistent with individual RPs being dispensable for accurate pre-mRNA processing.

Ubx$^{\Delta m1}$ and *Ubx*$^{\Delta m2}$ are hypomorphic alleles

Since intronic RPs did not appear to be required for accurate host intron removal, I switched my attention to expressed RS-exons. Typically included in the mature transcript, I wondered if these exons – coding and noncoding – contribute to some combination of RNA splicing and gene function. Several such exons have been previously described by Burnette and colleagues (J M Burnette et al., 2005) and I annotate an expanded list in Chapter 2 (Joseph et al., 2018). The first discovered and best-known examples are two 51 nt microexons (m1 and m2) in the *Drosophila* Hox gene *Ultrabithorax (Ubx)* (**Figure 3.10D**) (Hatton et al., 1998). These microexons, encoding for 17 amino acids each, are typically included in the mature mRNA, but also subject to intense alternative splicing (de Navas et al., 2011) making them attractive candidates for in-depth characterization.

The *Ubx*^{MX17} allele, created using X-ray mutagenesis, represents a key reagent to explore the significance of the *Ubx* microexons (Busturia et al., 1990). This mutant has a large intronic inversion of ~18 kb that includes the m2 microexon, and only expresses one *Ubx* isoform lacking both m1 and m2 microexons (Busturia et al., 1990; de Navas et al., 2011; Subramaniam et al., 1994). Remarkably, despite occurring more than 5 kb away from the m1 microexon, the *Ubx*^{MX17} inversion also causes m1 skipping in mRNA, leading to the speculation that the m2 exon is required for m1 inclusion (Subramaniam et

al., 1994). Loss of *Ubx* microexons also has demonstrable consequences on protein function since animals survive to adulthood but display a number of small changes in phenotype, the most obvious being a partial haltere to wing transformation (Busturia et al., 1990). In addition, flight and behavioral defects have been noted, as well as transformation of positional neuroblast identities in the developing embryonic CNS (Geyer et al., 2015; Subramaniam et al., 1994). Altogether, the mutations in Ubx^{MX17} creates a domino effect that starts from changes in RNA processing, to protein expression and finally animal development. However, as the genetic lesion in Ubx^{MX17} is a large 18 kb inversion, it is generally agreed that the defects observed confound several molecular influences, one of which is m2 loss. In fact, deletion of the m2 exon in a *Ubx* splicing minigene reporter did not lead to complete m1 skipping (Hatton et al., 1998), strengthening the view that other variables may influence RNA processing in Ubx^{MX17}. Hence, I decided to utilize my mutagenesis strategy to precisely delete the m1 and m2 microexons and study consequences on *Ubx* expression and animal development.

I chose to use the $Ubx^{\Delta 0nt\text{-}RP}$ allele to introduce further RS-exon deletions as this would allow me the unique chance to eliminate multiple recursive splice sites from the same host intron. I applied the same CRISPR/Cas9 with HDR approach to induce m1 and m2 deletions in the $Ubx^{\Delta 0nt\text{-}RP}$ background. Originally, both m1 and m2, were replaced with the ubi>GFP (with attP) cassette. As these alleles contained off-target lethalities, I used the GFP marker to backcross to wildtype animals (Canton S) for seven generations. Finally, I excised the marker to generate the Ubx^{13A} and Ubx^{22} alleles (**Figure 3.14A**). Unlike Ubx^{MX17} which is an ~18 kb inversion, the m2 deletion in Ubx^{22} is a drastically smaller 349 nt deletion. Similarly, in Ubx^{13A} a 593 nt deletion removes the m1 microexon.

Both mutants are homozygous lethal. However, this appears the effects of lingering off-target mutations as Ubx^{13A} and Ubx^{22} complement known amorphic *Ubx*

alleles. Nevertheless, the enhancement of the classic haltere to wing transformation was clearly evident for both alleles using the sensitized backgrounds of *Ubx*^{1} and *Ubx*$^{6\text{-}28}$ (**Figure 3.14B**). Interestingly, deletion of m2 evokes a stronger transformation, arguing for functional differences between the two similarly sized microexons. Thus, I conclude that *Ubx*^{13A} and *Ubx*^{22} are hypomorphic alleles.

Next, I aimed to examine the expression of Ubx protein in mutant animals. Similar to **Figure 2.2**, I performed immunostaining on first instar larval CNS to visualize Ubx protein in the double mutants. While the larval CNS may not be directly connected to the adult haltere, the literature on *Ubx* alternative splicing indicates a requirement in the CNS (Geyer et al., 2015; Subramaniam et al., 1994), thus it seemed a reasonable tissue for initial inspection. In wildtype animals Ubx is strongly expressed at the boundary of the thoracic and abdominal segments (**Figure 3.15**). This overall pattern of Ubx protein was found to be preserved in the *Ubx*^{13A} and *Ubx*^{22} double mutant condition, as well as the single mutant *Ubx*$^{\Delta 0nt\text{-}RP}$ larval CNS. Thus, despite the distinct perturbations and phenotypic strengths, the protein expression pattern appears generally unchanged and requires in-depth quantification for additional granularity.

Finally, I sought to examine the consequence of precise microexon deletions on *Ubx* splicing. This was of particular interest as *in vivo* genetic loss of m2 previously correlated with m1 skipping (de Navas et al., 2011; Subramaniam et al., 1994). Due to the two recursively spliced microexons and the cryptic RS-exon (0-nt RP), *Ubx* can be viewed as a four-intron gene, which requires four splicing reactions to remove the intronic fragments (**Figure 3.16A**). I designed primers to detect splicing intermediates that ligated exon 1 to all three RS-exons (**Figure 3.16A**, primers). I was unable to examine RNA processing in adults due to homozygous lethality of *Ubx*$^{\Delta 0nt\text{-}RP}$, *Ubx*^{13A} and *Ubx*^{22} and resorted to inspect first instar larval total RNA.

rt-PCR analyses to detect an RNA intermediate downstream of the m1 exon produced an amplicon in all conditions expect *Ubx*^{13A} (**Figure 13.16B**). This is molecular confirmation that loss of m1 eliminates splicing into this locus. A similar set of results were obtained for the m2 intermediate, which normally yields two amplicons (m1 + m2 or m2 only) (**Figure 13.16C**). While *Ubx*$^{\Delta 0nt\text{-}RP}$ produces both expected isoforms, the longer product containing both m1 and m2 is specifically lost in *Ubx*^{13A}, while both amplicons are lost in *Ubx*^{22}. As all three *Ubx* mutants contain the deletions of the 0nt-RP, no intermediates could be detected downstream of the cryptic RS-exon (**Figure 13.16D**). Finally, to understand changes to the coding sequence of *Ubx,* I inspected the mRNA in mutant animals. Although Ubx has six mRNA isoforms, three ("a" isoforms) are abundantly expressed in wildtype larvae (**Figure 13.16E**). The shortest of these skips both microexons (isoform IVa), the longest includes both (isoform Ia) whereas the intermediate length only includes m2 (isoform IIa). The same three exon combinations are type "b" if exon 1 is 27 nt longer. Consistent with the genetics, *Ubx*$^{\Delta 0nt\text{-}RP}$ yielded all three isoforms at the expected ratios. In *Ubx*^{13A}, only the longest isoform containing m1 and m2 was lost, and there was an overall increase in the intermediate length product containing m2 (isoform IIa). The dominant transcript in *Ubx*^{22} is the microexon skipped mRNA (isoform IVa). Intriguingly, there appears to be a minor level of transcripts with m1. However, these m1 containing transcripts in *Ubx*^{22} are type "b" and have a longer exon 1 (**Figure 13.16E**). Thus, while all three mutants express *Ubx* mRNA, *Ubx*^{13A} and *Ubx*^{22} have altered expression of the complement of *Ubx* isoforms. In the case of *Ubx*^{13A}, I observe the expected loss of the m1 from mRNA. However, in *Ubx*^{22}, consistent with *Ubx*^{MX17}, there is a depletion of m1 containing transcripts. Moreover, the lower levels of transcripts with m1 inclusion appear to have annotated alternative splicing of exon 1. Moreover, while m1 behaves similarly to the other RS-exons mutated in this study, the data supports a hypothesis in which the m2 microexon is required for the inclusion of m1

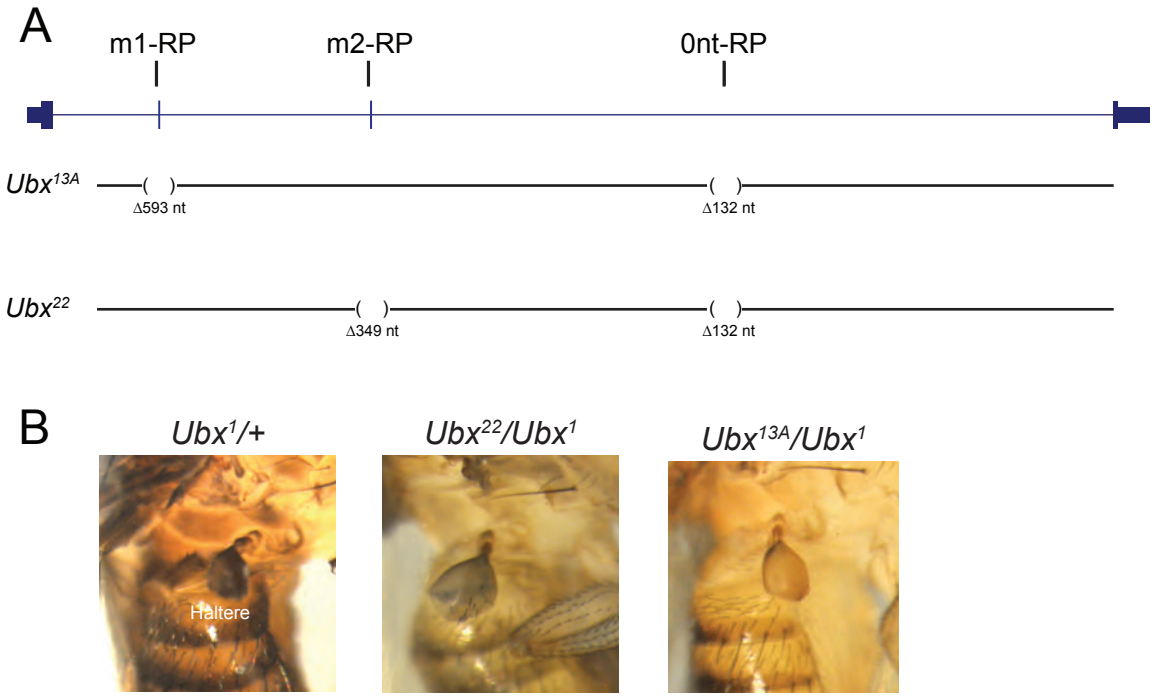


Figure 3.14. Animals with *Ubx* m1 and m2 microexon deletions display loss-of-function phenotype.
(A) Schematics of RS-exon deletions in the *Ubx* locus. Ubx^{13A} is a deletion of 593 nt surrounding the m1 microexon and 132 nt surrounding the 0nt-RP. Conversely, Ubx^{22} is a deletion of 349 nt including the m2 microexon and 132 nt of the 0nt-RP.
(B) Lateral images of adult flies that illustrate haltere to wing transformation *Ubx* animals. The RS-exon deletions enhance the Ubx^{1} phenotype.

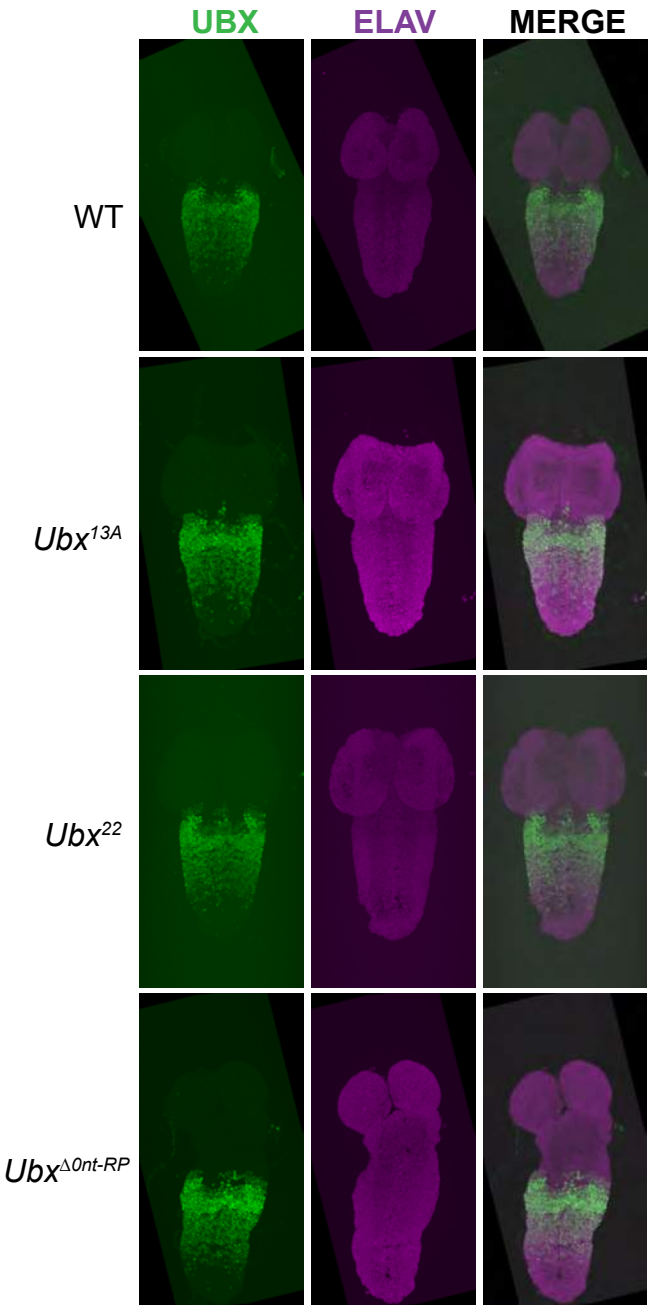


Figure 3.15. Ubx protein is expressed in *Ubx* RS-exon mutants. Immunostaining of first instar larval CNS. WT indicates the normal segmental pattern of Ubx protein (green) in the ventral nerve cord, counterstained with pan-neuronal ELAV (magenta). The same pattern of Ubx expression is observed in the single mutant $Ubx^{\Delta 0nt\text{-}RP}$, as well as double mutants Ubx^{13A} and Ubx^{22}.

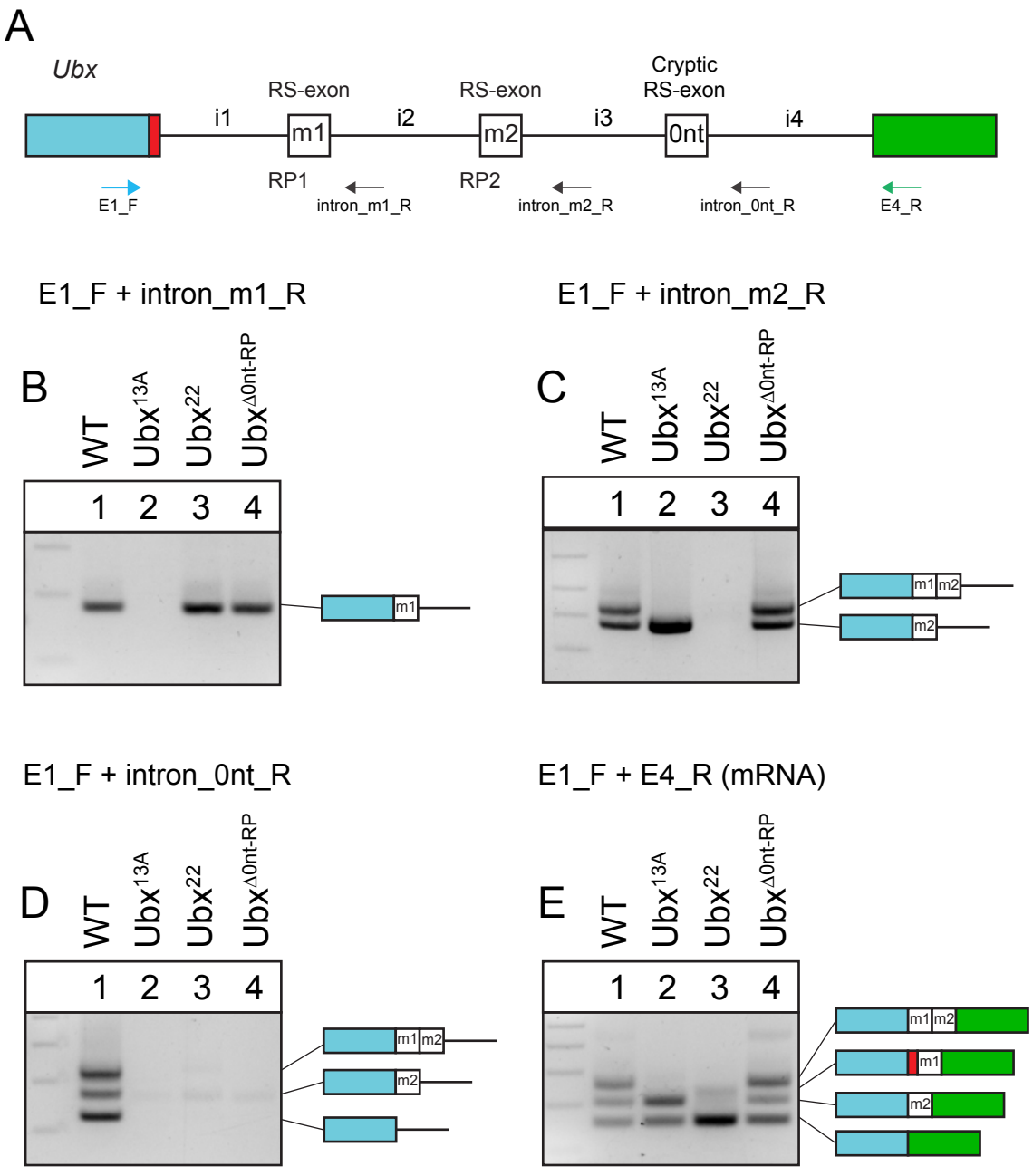

A
Ubx
RS-exon
RS-exon
Cryptic
RS-exon
i1
m1
i2
m2
i3
0nt
i4
RP1
RP2
E1_F
intron_m1_R
intron_m2_R
intron_0nt_R
E4_R
E1_F + intron_m1_R
E1_F + intron_m2_R
B
C
WT
Ubx13A
Ubx22
UbxΔ0nt-RP
1 2 3 4
m1
m1 m2
m2
E1_F + intron_0nt_R
E1_F + E4_R (mRNA)
D
E

Figure 3.16. Molecular characterization of splicing products in *Ubx* mutants

(A) Top: gene model displaying *Ubx* along with the location of three RS-exons within a large ~74 kb intron. Because of three RS-exons, this gene essentially has four intronic fragments (i1-i4). Primers used to examine each of the three splicing intermediates as well as mRNA are indicated along with primer IDs.
(B-E) rt-PCR analyses of the three recursive intermediates and the mRNA amplicon in wildtype and *Ubx* mutants.
(B) Removal of i1 produces intermediate 1. rt-PCR indicates that the intermediate 1 amplicon is only lost in Ubx^{13A} mutants. This is not surprising given that Ubx^{13A} has an m1 deletion.
(C) Removal of intron 1 and intron 2 generates intermediate 2. Wildtype animals express two isoforms of intermediate 2. The longer contains m1 and m2, whereas the shorter contains just m2. Importantly, m1 skipping in the shorter intermediate occurs through recursive splicing. Ubx^{13A} mutants animals only produce the short isoform whereas Ubx^{22} mutants which lack m2 do not produce a second intermediate.
(D) Removal of intron 1, 2 and 3 generates intermediate 3. As all three mutants contain a deletion of the cryptic RS-exon (0nt-RP), intermediate 3 is not produced in all three examined *Ubx* mutant.
(E) mRNA is produced after removal of all four introns. The same complement of mRNA amplicons can be observed for wildtype and the single mutant ($Ubx^{\Delta 0nt\text{-}RP}$). In contrast deletion of the m1 or m2 microexon has mRNA consequences as both exons are included in wildtype mRNA. Only the m1 containing product was lost in Ubx^{13A}. Interestingly, both wildtype m1 and m2 containing isoforms are lost in Ubx^{22}, suggesting the m2 microexon and flanking sequences are required for inclusion of the m1 microexon as well. Moreover, a small amount of m1 inclusion was observed in Ubx^{22}, but found to have annotated alternative splicing in exon 1 as well (red box, **Figure 3.16A**).

in *Ubx* transcripts. Overall my results demonstrate that recursive splicing may contextually aid in the regulation of RNA processing and gene expression.

Discussion

Multiple factors influence choice between RP 5'SS and RS-exon 5'SS

Several factors are known to regulate SS choice that leads to alternative splicing. These include *cis*-elements, *trans*-acting factors, the histone code, RNA modifications, RNAPII regulation, gene architecture and other factors (De Conti et al., 2013; Lee & Rio, 2015). Yet, despite the two decades that have passed since the first discovery of recursive splice sites in introns, relatively little is known regarding the mechanism of RS. In this study, I examine the role of exonic splicing regulatory elements, the EJC, and 5'SS strength as influences on RS-exon inclusion. First, using *in vivo* mutagenesis, I show that decreasing RP 5'SS strength endogenously can convert cryptic RS-exons in expressed RS-exons. Thus, my data provide strong support for the 5'SS competition model (Sibley et al., 2015) using an experimental system with appropriately long introns.

Orthogonally, my RS minigene reporters indicated that 5'SS is not adequate to regulate inclusion, as a few reporters are able to include the RS-exon despite having stronger RP 5'SS and *vice versa*. In this regard, the RS-exon swap experiments hinted that sequence in RS-exons is independently sufficient at instructing expression patterns. Hence swapping in a cryptic RS-exon in place of an expressed exon results in RS-exon skipping. Consistently, the opposite effect is seen for reporters in which an expressed RS-exon replaces a cryptic one. These data hint at the presence of exonic SREs that guide the observed patterns of AS. In general, ESEs are commonly observed within constitutively expressed exons (Z Wang et al., 2004). This is most certainly true for the expressed *Ubx* microexons m1 and m2, which show deep evolutionary conservation

across all 51 nt, including wobble positions (James M. Burnette et al., 1999). However, in the case of cryptic RS-exons, it is unclear if these exons contain ESS sequences, or whether the default state for RS-exons (in the absence of SREs) is to activate the RP 5'SS. The latter seems more likely given that cryptic RS-exons (beyond the RP 5'SS) are poorly conserved and are unlikely to contain important regulatory elements.

Finally, I demonstrate that pre-removal of the upstream intron segment causes RS-exon skipping. This clearly indicates that SS choice is influence by the history of previous splicing. This attribute is characteristic of the EJC deposited upstream of exon junctions during splicing (Boehm & Gehring, 2016). As this function was previously reported in the mammalian system (Blazquez et al., 2018), my experiments demonstrate that the EJC has a conserved function to suppress regenerated splice sites after splicing. Conversely, understanding how cryptic RS-exons (intronic RPs) evade EJC regulation represents a potentially productive future direction.

RPs may be dispensable for long intron removal

To my knowledge, there are no published reports of RP deletions. However, the question of their function remains an important concern for the splicing field. Using CRISPR/Cas9 mutagenesis I delete several intronic RPs and show that mRNA production in these RP deletion mutants is generally unaffected. However, it is quite possible that RS may be required in specific cell types, and under conditions that have not yet been identified. Thus, the qualitative rt-PCR test may not provide enough granularity to make a conclusive statement. Indeed, deeper characterization and generation of more RP mutants will provide a better assessment.

Cryptic splice sites sometimes function as decoys to prevent activation of other detrimental splice sites. Furthermore, they may also be coupled to other kinds of

regulation, for example prevention of pre-mature cleavage and polyadenylation – a process called telescripting (Berg et al., 2012; Oh et al., 2017). It will be interesting to examine the RP mutants within these contexts and explore other functional possibilities.

The *Ubx* m2 microexon regulates m1 inclusion

The Ubx^{MX17} allele is a large ~18 kb inversion of sequence that surrounds the m2 exon. Intriguingly, this genetic lesion also leads to skipping of the m1 exon which lies approximately 5 kb away (de Navas et al., 2011; Subramaniam et al., 1994). The splicing of this mutant has generated the hypothesis that the m2 exon is required for m1 inclusion in *Ubx* mRNA. However, the large inversion confounds several effects, so it is not clear if this attribute is related to m2, or some other defect. Attempts to model this in cell culture using minigene reporters found that m1 could was included in m2 deletion constructs, although at lower levels (Hatton et al., 1998). Thus, given that there was an open RNA processing question and strong likelihood of functional consequences, engineering precise deletions of the *Ubx* microexons seemed an attractive proposition.

Here I report the generation of Ubx^{13A} and Ubx^{22}, mutants that contain double deletions of the *Ubx-0nt-exon,* and either m1 (Ubx^{13A}) or m2 (Ubx^{22}). These mutants appear to be *Ubx* hypomorphs, with Ubx^{22} having the stronger phenotype. While mutant animals express the Ubx protein, molecular characterization indicates difference in mRNA. The m1 exon is lost in mRNA transcripts from Ubx^{13A}, and consistent with the observations for Ubx^{MX17}, Ubx^{22} animals predominantly skip both m1 and m2. Thus, it seems that the m2 exon may regulate inclusion of the upstream m1 exon. As both the m1 and m2 RPs have been replace with attP, it will be possible to study how these microexons contribute to RNA processing and protein function.

Methods

Recursive splice site engineered alleles

All animal mutants reported in this study were generated using CRISPR-Cas9 mutagenesis. The strategy used to manipulate each RSS is detailed below. All targeting guide RNA sequences and primer information can be found in **Tables 3.8-3.11**. All Injections were performed at Bestgene, Inc. (https://www.thebestgene.com/)

muscleblind RP2, *RP3* and *RP4* deletions

The general approach for generating *muscleblind* RSS deletion alleles was to cut and replace the RS-exon and flanking intronic sequences with a fluorescent marker using homology directed repair. To accomplish this, we cloned four targeting guide RNA sequences per RS-exon under the U6 promoter in the gRNA expression vector, pCFD5 (Port & Bullock, 2016). ~1000 nt left and right homology arms for each target were cloned into either pHD-dsRED (*mbl-RP1 and mbl-RP4)* (Gratz et al., 2014) or pHD-attP-ubiGFP (*mbl-RP2 and mbl-RP3)* donor plasmids. Importantly, fluorescent markers on donor plasmids were designed to lie on the opposite strand relative to *mbl.* Combinations of guide RNA expression vectors and donor plasmids were injected into embryos expressing Cas9 and progeny were screened for fluorescent green bodies and red eyes. Precise replacement was then confirmed via PCR and Sanger sequencing. Initially, we aimed to replace multiple RS-exons during a single injection experiment. Injection cocktails are listed below and specify target RS-exons.

1. *RP-1/RP3*: pCFD5-RP1 + pHD-dsRED-RP1-HR + pCFD5-RP3 + pHD-attP-ubiGFP-RP3-HR.
2. *RP-2/RP4*: pCFD5-RP4 + pHD-dsRED-RP4-HR + pCFD5-RP2 + pHD-attP-ubiGFP-RP2-HR.

As evident from the above mixes, our experimental design was to make double RP mutants (*mbl RP1 + mbl RP3 and mbl RP2 + mbl RP4)* simultaneously, and we aimed to identify successfully mutated double mutants using dsRED and GFP as markers. However, as we did not obtain any dual marked animals, we were unable to generate animals with dual replacement. Thus, *mbl RP1* and *mbl RP4* were marked with dsRED and *mbl RP2* and *mbl RP3* were marked with GFP. For animals that had successful HR-based replacement of one RS-exon, we routinely checked the other target RS-exon for lesions that might arise due to non-homologous modes of DNA repair, such as insertions or deletions. On average, about 40 animals per confirmed single mutant were examined. While we occasionally detected cis-indels at the other RS-exon targets, we did not generate any RS-exon manipulations or deletions. All alleles used for downstream experiments were selected and verified to have no mutations at all non-marked RS-exons and flanking sequences.

Egfr$^{\Delta RP}$, *kuz*$^{\Delta RP}$ and *Ubx*$^{\Delta 0nt\text{-}RP}$

These mutants were generated using transgenic CRISPR (Kondo & Ueda, 2013). Briefly, for each target, two targeting guide RNAs were cloned into the gRNA expression vector pCDF4 (Port et al., 2014). Transgenic flies expressing the gRNA expression vectors were generated using the PhiC31 integrase-mediated transgenesis strategy. Next, we combined transgenic Cas9 (*nos>Cas9)* and CRISPR, and crossed these flies

to *yw* flies. Progeny from this cross were balanced and screen via PCR and Sanger sequencing for RP mutagenesis.

in vivo kuz and *Bx* RP 5'SS mutagenesis

These mutants were also generated using the same reagents as listed in Chapter 2.

$ds^{\Delta RP}$

These mutants were also generated using the transgenic CRISPR strategy as described above, with the only difference being that, four targeting sequences were employed using the pCFD5 gRNA expression vector (Port & Bullock, 2016).

Ubx^{-13A} and Ubx^{22}

Ubx ultraconserved microexons (m1 and m2) were deleted in the $Ubx^{\Delta 0nt\text{-}RP}$ background using CRISPR/Cas9 targeted mutagenesis with homology directed repair. To accomplish this, *Ubx*-Δ*RP* was first combined with transgenic Cas9 (*nos>Cas9).* Embryos carrying the RP deletion and transgenic Cas9 were injected with four targeting guide RNA sequences per microexon under the U6 promoter in the gRNA expression vector, pCFD5 (Port & Bullock, 2016). ~1000 nt left and right homology arms for each target were cloned into pHD-attP-ubiGFP donor plasmids. Importantly, the fluorescent marker on pHD-attP-ubiGFP was designed to lie on the opposite strand relative to *Ubx.* guide RNA expression vectors and donor plasmids were injected into the aforementioned embryos and progeny were screened for green bodies. Precise replacement was then confirmed via PCR and Sanger sequencing. Subsequently, the donor plasmid and marker were excised by balancing the third chromosome over *TM6B-hs-Cre* (BDRC #1501)

Immunostaining

To study *Ubx* phenotypes, I used the amorphic alleles Ubx^{1} and $Ubx^{6\text{-}28}$. To stain for Ubx, *Ubx*/TM6B-[ubi-GFP]* or *yw* flies were allowed to lay eggs in cages for 24 hrs at 25°C (* denotes *Ubx* alleles). After sufficient time, GFP-negative first instar larvae were hand-picked under a fluorescence microscope and dissected to obtain CNS. The samples were fixed and incubated with the following primary antibodies: rat anti-Elav (1:100, 7E8A10, DSHB) and mouse anti-Ubx (1:10, FP3.38, DSHB).

Constructs and cell culture

The splicing reporter used in this study is previously reported in Chapter 2 (Joseph et al., 2018). For each cloned RS reporter (**Figure 3.4**), I amplified ~3 kb of intronic sequences containing the RP using PCR. The sequences were cloned into the minigene construct using NotI and EcoRV restriction sites. All RP cloning primers are listed in **Table 3.5.** Mutagenesis of RS reporters to generate RP 5'SS disruptions was performed using site directed mutagenesis. A similar strategy was used to pre-remove intron segment 1 in RS reporters and to swap RS-exons. Primers used are listed in **Tables 3.5-3.7**.

All transfections in this study were performed using S2-R^{+} cells cultured in Schneider *Drosophila* medium with 10% fetal Bovine serum. Cells were seeded in 6-well plates at a density of 1 million/mL and transfected with 200 ng of construct using the Effectene transfection kit [Qiagen]. Cells were harvested following three days of incubation.

rt-PCR of mRNA and recursive intermediates

kuz and *Bx* RP 5'SS mutants (**Figure 3.1-3.2**): First instar larval samples were used for *kuz* mutants, whereas adult female flies for *Bx* mutants. rt-PCR primers listed in Chapter 2 were reused for these experiments. For cell culture tests, rt-PCR primers are as listed in Chapter 2. $kuz^{\Delta RP}$, $ds^{\Delta RP}$, $Egfr^{\Delta RP}$, *mbl* and *Ubx* deletions: Adult females flies were used for $kuz^{\Delta RP}$, $ds^{\Delta RP}$, $Egfr^{\Delta RP}$ and *mbl* mutants, whereas first instar larvae were used for all *Ubx* mutant samples. All rt-PCR primers used in this study are listed in **Table 3.4**. rt-PCRs were done using AccuPrime™ Pfx DNA polymerase [ThermoFisher Scientific] with standard protocol using 32 cycles for mRNA and 34 cycles for intermediates.

RNA isolation and cDNA preparation

S2 cells, mutants and control animals were homogenized and RNA was extracted using the standard Trizol protocol. 5 µg of RNA was treated with Turbo DNase [Ambion] for 45 min before cDNA synthesis using SuperScript III [Life Technology] with random hexamers.

Table 3.1. Sequences of *Bx* RP 5'SS mutants

ID	Sequence
WT	TTGTTTTTCCAGGTAAGTGTCAACACCCACCCAATTGCTA
13	TTGTTTTTCCAGGTgtgtcAAGTGTCAACACCCACCCAtttCTA
20	TTGTTTTTCCAGGTt--TGTCAACACCCACCCAtt--CTA
16	TTGTTTTTCCAGGTA-GTGTCAACACCCACCCAtTT-CTA
21	TTGTTTTTCCAGGT------CAACACCCACCCAtTT-CTA
24	TTGTTTTTCCAGGTcAAGTGTCAACACCCACCCATTtCTA
12	TTGTTTTTCCAGGT---TGTCAACACCCACCCAtTT-CTA
23	TTGTTTTTCCAGGT----GTCAACACCCACCCAATTGCTA

Table 3.2. Sequences of *kuz* RP 5'SS mutants

ID	Sequence
WT	TTCTCTTTACAGGTGAGTGCTCGGTTTCTAACGCT
24	TTCTCTTTACAGGTGAGTGCTCGGTTTCTAACGCT
14	TTCTCTTTACAGGTGAG--CTCGGTTTCTAACGCT
30	TTCTCTTTACAGGTGAa----------AACGCT
26	TTCTCTTTACAGGT-AGTGCTCGGTTTCTAACGCT
15	TTCTCTTTACAGGT----------TTCTAACGCT
16	TTCTCTTTACAGGTcttctctttCTCGGTTTCTAACGCTGAAAATG
31	TTCTCTTTACAGGTttctctCTCGGTTTCTAACGCTGAAAATG

Table 3.3. Sequence of Gdep and FP RS-exons

RS-exon	Sequence
Ubxm1	AGGTAAGATAAGATCTGATTTAACACAATACGGCGGCATATCAACAGACATGGGTAAGA
Gdep	AGGTAAGATAAAACCAATAAATTCCCAATACTATAAAATATCAACAGACATGGGTAAGA
Ubx0nt	AGGTAAGTGTCAAATATTTAATACACCCTTAAACCAAAACAAAAA-CATTGACAAAGTGAGT
FP	AGGTAAGTGTCAAATATTTAATACACCCTTAAACCAAAACAAAAAACATTGACAAAGTGAGT

Table 3.4. All rt-PCR primers used in this study

dachsous (ds)	sequence	amplicon
ds_exon2_fwd	CCTGATCACCACCCGATCG	mRNA
ds_exon3_rvs	GCTACTCCTCCGCTCGAAG	mRNA
ds_exon2_fwd	CCTGATCACCACCCGATCG	intermediate
dsRP_seqR	aacgtcttgacaggcgac	intermediate
Ultrabithorax (Ubx)		
ubx.univ.int.fwd	CCAGCAATCACACATTCTACC	m1 intermediate
Ubx_m1_int_rvs	AGTGGACCTGCTCTACACTC	m1 intermediate
ubx.univ.int.fwd	CCAGCAATCACACATTCTACC	m2 intermediate
Ubx_m2_int_rvs	CTGGACATTTTGGAGTGGACG	m2 intermediate
ubx.univ.int.fwd	CCAGCAATCACACATTCTACC	0nt intermediate
ubx.int.rvs	CTTTGCCCAGCACGCATGAG	0nt intermediate
ubx.univ.int.fwd	CCAGCAATCACACATTCTACC	mRNA
ubx.mRNA.rev	CATCTCGATTCTCCGTCTG	mRNA
Egfr		
Egfr_exon1_fwd	GGACAGCAGCTCCATCTGG	mRNA
Egfr_exon2_rvs	GCTCCAGGTTGCCATCCAC	mRNA
Egfr_exon1_fwd	GGACAGCAGCTCCATCTGG	intermediate
EgfrRP_seqR	AAACCATTGAGACAGTACGC	intermediate
mbl		
q87_mbl_circ2	CCAACGTGGAGGTCCAGAAC	mRNA
q.f.150_mbl	CGGTCAGATAGGGGTTTGTT	mRNA
q87_mbl_circ2	CCAACGTGGAGGTCCAGAAC	intermediate 1
int_mblRI1_R	GCAAACTCGCCTGCATTGAC	intermediate 1
q87_mbl_circ2	CCAACGTGGAGGTCCAGAAC	intermediate 2
int_mblRI2_R	GTCCCTGTCTCTGTCTGCAGT	intermediate 2
q87_mbl_circ2	CCAACGTGGAGGTCCAGAAC	intermediate 3
int_mblRI3_R	TTTTGCCAGTCGCTCAGCTC	intermediate 3
q87_mbl_circ2	CCAACGTGGAGGTCCAGAAC	intermediate 4
int_mblRI4_R	CCCAGCAGCATCCCTCTCTC	intermediate 4

Table 3.5. Primers used to clone 15 RS reporters and those with RP 5'SS disruptions

chinmo_f	CTTGCTGTCTCCCTTTCTCC
chinmo_r	ACAAGCAAGCAGACACAAGC
cut_f	GAAAACAACAAGGGTCAACTGATG
cut_r	TAAAAGTGAGCCACAGAAGCG
fra_f	GCCAGATACTGTTGTCCAC
fra_r	TTTATGGTTTCTGCAGCGAC
hth_f	CCACAACGCAGTTGCTCC
hth_r	CAAACGACGAGCGACAGC
nmo_f	TTGACGCAAGGTGGAGTTTG
nmo_r	TTGTTGCTCAAGATCACACAC
shep_f	AACTGCAGCGACAACAGC
shep_r	GAACCACATATAGGACCACG
sm_f	GATTTCGTCAACTGCTCATACC
sm_r	GTAAGGTTTGTGCGTGGAG
Ubx0nt_f	CAACGATGGCAGTTCAGC
Ubx0nt_r	CTGCATGTAGCAGGGATC
Ubxml_f	TGACTTCTTCTGGCTGCAAC
Ubxml_r	GGTGCTTATCTGTGAGAGTC
heph245_f	AAGCTCAGTGCGAAAGCTCC
heph245_r	GTCATCATGAACCGTCAGTC
heph349_f	TTCGTGCTTTGGCAGGATGG
heph349_r	CTCTCCGAACTTCCAAGACG
msi87_f	TGGCACGTGCATCTCGTCAC
msi87_r	GCATTCCTCAACTGAGCTAC
mub29_f	CTTTCTCGGACTCTCGATCC
mub29_r	GAAAGTTGCAACTGCACCTC
ps67_f	AGCTGCTGCACAGTGTCAAC
ps67_r	CATGTGATACGTTGTCTCGC
egfr_f	ATCGAGCAGGCTTGTTGTC
egfr_r	ATAACTCACCACTAGCTTAGCG

RP 5'SS disruption	
mut_heph245RP_f	CTGCTTTCAGaaAAGTTTGCACTAC
mut_heph245RP_r	AAGAGAGAAACGAAACGTTAAATTTC
mut_mubRP_f	CTCGTTTCAGaaACGTGTCCATG
mut_mubRP_r	AGAAAACAGTGAATTTAATGAAC
mut_ctRP_f	TCTTTTACAGaaATGTTTACATCGAAG
mut_ctRP_r	AGAAGACATGTGTCAATTAG
mut_Ubx0ntRP_f	TCTTTTCTAGaaAAGTGTCAAATATTTAATAC
mut_Ubx0ntRP_r	GAAGAAAATAGTTTGATTAGTATTAG

Table 3.6. Primers used to remove intron segment 1 from RS minigene reporters

Ubx0nt_del_Upint_F	gGTAAGTGTCAAATATTTAATACAC
Ubxm1_del_Upint_F	gGTAAGATAAGATCTGATTTAACAC
smRI_del_Upint_F	gGTAAGTCGCTGTTTTCTATAC
msiRI_del_Upint_F	gGTCAGTATCGGAGATGAAC
heph245_del_Upint_F	gGTAAGTTTGCACTACGAG
kuz_E3_del_Upint_R	tCTTTTAACTCCAGAAATATCTTTTG

Table 3.7. Primers used for RS-exon swaps and RS-exon modifications

These constructs switch Ubx0nt RS-exon with Ubxm1,m2 and chinmo RS-exons

spe_Ubx_0-m1F	TACGGCGGCATATCAACAGACATGGGTGAGTAAATAAGTATAATAATAAAAAG
spe_Ubx_0-m1R	TTGTGTTAAATCAGATCTTATCTTACCTAGAAAAGAGAAGAAAATAGTTTG
spe_Ubx_0-m2F	GGCTCACTTCTACCAGACTGGCTAGGTGAGTAAATAAGTATAATAATAAAAAG
spe_Ubx_0-m2R	CGCAAGAGATTCTGAGTATCTCTTACCTAGAAAAGAGAAGAAAATAGTTTG
spe_Ubx0-chinmoF	ATTAGGCCGTTGTGGTGTATAGCGTAACGTGAGTAAATAAGTATAATAATAAAAAG
spe_Ubx0-chinmoR	TAGTTCGAAAAAAGGCACCAGTGCTTACCTAGAAAAGAGAAGAAAATAGTTTG

These constructs switch Ubxm1 RS-exon with Ubx0nt,m2 and chinmo RS-exons

spe_Ubxm1-0ntF	TAAACCAAAACAAAAACATTGACAAAGTAAGAAAATTTCCACTTTTATTTC
spe_Ubxm1-0ntR	AGGGTGTATTAAATATTTGACACTTACCTGAAAATGCAAGCAAAG
spe_Ubxm1-m2F	GGCTCACTTCTACCAGACTGGCTAGGTAAGAAAATTTCCACTTTTATTTC
spe_Ubxm1-m2R	CGCAAGAGATTCTGAGTATCTCTTACCTGAAAATGCAAGCAAAG
spe_Ubxm1-chinmoF	ATTAGGCCGTTGTGGTGTATAGCGTAACGTAAGAAAATTTCCACTTTTATTTC
spe_Ubxm1-chinmoR	TAGTTCGAAAAAAGGCACCAGTGCTTACCTGAAAATGCAAGCAAAG
Ubxm1_AA_Gdep_F	tactataaaatatcaacagacatggGTAAGAAAATTTCCACTTTTATTTC
Ubxm1_AA_Gdep_R	ttgggaatttattggttttatcttacCTGAAAATGCAAGCAAAG
Ubx0nt_FramePres_F	taaaccaaaaccaaaaacatcgacaaaGTGAGTAAATAAGTATAATAATAAAAAG
Ubx0nt_FramePres_R	agggtgtattaaatatttgacacttacCTAGAAAAGAGAAGAAAATAGTTTG

Table 3.8. guide RNA cloning primers and sequencing primers for *ds*, *Ubx-0nt* and *Egfr*

guide RNA cloning primers	
Egfr-RP	
Egfr.RI-BsaF	CGGTCTCA CTTC G AGCGAACTCACCTGCAAAGA GTTTTAGAGCTAGAAATAGCAAG
Egfr.RI-BsaR	CGGTCTCA AAAC ATCTGGGTCTCTATGCACAT C GAAGTATTGAGGAAAACATACC
Ubx-0nt-RP	
Ubx.RI-BsaF	CGGTCTCA CTTC G AAAATAGTTTGATTAGTATT GTTTTAGAGCTAGAAATAGCAAG
Ubx.RI-BsaR	CGGTCTCA AAAC GTTAAAGCAGCGGTGAGTGG C GAAGTATTGAGGAAAACATACC
ds-RP	
pCFD5-dsP1f	GCGGCCCGGGTTCGATTCCCGGCCGATGCACTTAACTAGGCGTAAGTATTGTTTTA-GAGCTAGAAATAGCAAG
pCFD5-dsP1r	ATACTTACTGCAGATACTTATGCACCAGCCGGGAATCGAACCC
pCFD5-dsP2f	TAAGTATCTGCAGTAAGTATGTTTTAGAGCTAGAAATAGCAAG
pCFD5-dsP2r	GAAATAAAAGATGGCATAAATGCACCAGCCGGGAATCGAACCC
pCFD5-dsP3f	TTTATGCCATCTTTTATTTCGTTTTAGAGCTAGAAATAGCAAG
pCFD5-dsP3r	ATTTTAACTTGCTATTTCTAGCTCTAAAACTACACGCTGTGCGATGAATCTGCACCAGCCG-GGAATCGAACCC
genotyping primers	
Egfr.RI-CHKF	*GCGAAAGTGTTGCAAGTGCTGGGAAAGC*
Egfr.RI-CHKR	*CACGACAACGGAGAGCAGCGTTTTAGCC*
Ubx.RI-CHKF	*CTTTACACCTTTACACGGGCGTATTTTC*
Ubx.RI-CHKR	*GGATGGCAGGGGTGTGTGGGTGCTATG*
dsRP_seqF	*cgagatcaaacgcagagc*
dsRP_seqR	*aacgtcttgacaggcgac*

Table 3.9. Primers used to clone mbl guide RNA constructs and HDR templates

HR primers	backbone	
SapI_mblRI1_HR1_fwd	ATGCAACATGTGCCAGAAGC	pHD-dsRED
SapI_mblRI1_HR1_rev	AAGTTTGGGGACATATTGCAGAG	pHD-dsRED
AarI_mblRI1_HR2_fwd	CTTCTGCAACTTTGCCGTCG	pHD-dsRED
AarI_mblRI1_HR2_rev	CTTTGTACCCACACGTGATGGC	pHD-dsRED
SapI_mblRI2_HR1_fwd	TGCGTTCCCTCATGTGGAAG	pHD-attP-GFP
SapI_mblRI2_HR1_rev	CCCCACAAAACTATGTCGCAC	pHD-attP-GFP
NotI_mblRI2_HR2_fwd	CAGAGCGAGGGTTAAGGCTG	pHD-attP-GFP
EcoRI_mblRI2_HR2_rev	GTGCAGCGGAAGTAGCAGC	pHD-attP-GFP
SapI_mblRI3_HR1_fwd	ATCGGACTTCCCTACTTTGTATGC	pHD-attP-GFP
SapI_mblRI3_HR1_rev	GCTAATTACGCAACAAGGGACATC	pHD-attP-GFP
NotI_mblRI3_HR2_fwd	CGCTGCATTTTATGGCAATGCG	pHD-attP-GFP
EcoRI_mblRI3_HR2_rev	GCTTCTACTATTCAGACCCAAGGAGC	pHD-attP-GFP
SapI_mblRI4_HR1_fwd	GCAAAAGGAGGAGGTAATGACAGG	pHD-dsRED
SapI_mblRI4_HR1_rev	GTGTCGAGCGCTTGCAAC	pHD-dsRED
AarI_mblRI4_HR2_fwd	AGCTGGAACCTCTGACCAACTG	pHD-dsRED
AarI_mblRI4_HR2_rev	CCCTGGCTGAACTAAACCGAAC	pHD-dsRED

guides clones into PCFD5

mbl_1_PCR1fwd GCGGCCCGGGTTCGATTCCCGGCCGATGCAAAAATCGAAGTGTATCTTTGGTTTTA-GAGCTAGAAATAGCAAG
mbl_1_PCR1rev ACAGACACAAGCATTTGATGTGCACCAGCCGGGAATCGAACCC
mbl_1_PCR2fwd CATCAAATGCTTGTGTCTGTGTTTTAGAGCTAGAAATAGCAAG
mbl_1_PCR2rev AACTTAAACTTAGTTTCTATTGCACCAGCCGGGAATCGAACCC
mbl_1_PCR3fwd ATAGAAACTAAGTTTAAGTTGTTTTAGAGCTAGAAATAGCAAG
mbl_1_PCR3rev ATTTTAACTTGCTATTTCTAGCTCTAAAACGTTTACATTCGGGGCAAAGGTGCACCAGCCG-GGAATCGAACCC

mbl_2_PCR1fwd GCGGCCCGGGTTCGATTCCCGGCCGATGCAGGGGAAGTTGCGGTACCATTGTTTTA-GAGCTAGAAATAGCAAG
mbl_2_PCR1rev CCCCGAAGTTTTTTGCACTCTGCACCAGCCGGGAATCGAACCC
mbl_2_PCR2fwd GAGTGCAAAAAACTTCGGGGGTTTTAGAGCTAGAAATAGCAAG
mbl_2_PCR2rev GTTAATAAAGAAGGCTAGCATGCACCAGCCGGGAATCGAACCC
mbl_2_PCR3fwd TGCTAGCCTTCTTTATTAACGTTTTAGAGCTAGAAATAGCAAG
mbl_2_PCR3rev ATTTTAACTTGCTATTTCTAGCTCTAAAACGGAATTGGAGACCACAGAGCTGCACCAGCCG-GGAATCGAACCC

mbl_3_PCR1fwd GCGGCCCGGGTTCGATTCCCGGCCGATGCATCTCTGCTAATTACGCAACAGTTTTA-GAGCTAGAAATAGCAAG
mbl_3_PCR1rev CCTTGACTTTCCTTTTGTCATGCACCAGCCGGGAATCGAACCC
mbl_3_PCR2fwd TGACAAAAGGAAAGTCAAGGGTTTTAGAGCTAGAAATAGCAAG
mbl_3_PCR2rev AATCGTCTTGTTGTCCGTTTTGCACCAGCCGGGAATCGAACCC
mbl_3_PCR3fwd AAACGGACAACAAGACGATTGTTTTAGAGCTAGAAATAGCAAG
mbl_3_PCR3rev ATTTTAACTTGCTATTTCTAGCTCTAAAACCAGCATCGCGTGTCGCATTGTGCACCAGCCG-GGAATCGAACCC

mbl_4_PCR1fwd GCGGCCCGGGTTCGATTCCCGGCCGATGCACTATTCAGAGTTTGGTCTATGTTTTA-GAGCTAGAAATAGCAAG
mbl_4_PCR1rev TTGTGTCAGCAGCACACATGTGCACCAGCCGGGAATCGAACCC
mbl_4_PCR2fwd CATGTGTGCTGCTGACACAAGTTTTAGAGCTAGAAATAGCAAG
mbl_4_PCR2rev TTGAGTAACCCCAACTATTTTGCACCAGCCGGGAATCGAACCC
mbl_4_PCR3fwd AAATAGTTGGGGTTACTCAAGTTTTAGAGCTAGAAATAGCAAG
mbl_4_PCR3rev ATTTTAACTTGCTATTTCTAGCTCTAAAACACTGTTTATGCTGATCATCATGCACCAGCCG-GGAATCGAACCC

Table 3.10. Primers used for Ubx m1 and m2 deletions

HR primers	
XhoI_Ubxm1_HR_fwd	GCATGTAAACAGCACTCAGC
SpeI_Ubxm1_HR_rev	CCGGTTAAGATTTGCCAACC
Not1_Ubxm1_HR2_fwd	TATCGTACCTCGTGCTATCG
EcoRI_Ubxm1_HR2_rev	CGGTGCTTATCTGTGAGAGTC
StuI_Ubxm2_HR1_fwd	TGATACGTTGTTGAGCTCCAG
SpeI_Ubxm2_HR1_rvs	GTCCAGGAACGTCATTATGCC
NotI_Ubxm2_HR2_fwd	CGTCCACTCCAAAATGTCCAG
NheI_Ubxm2_HR2_rev	TGGAAACGTACTTGTTGTTGCC

Table 3.11. Gentyping primers for Ubx

Ubxm1_delCheck_F	TGAAGTGCACTTTGAGTGGC
Ubxm1_delCheck_R	TTCGGCTACTTGATCGTCGG
int_Ubxm2_F	GAAATTCCTCCGGCAGCCTC
int_Ubxm2_R	ACCTCTCGAACTCTGGCAGG

Chapter 4

The Exon Junction Complex and intron removal prevents resplicing of mRNA

Summary

Accurate splice site selection is critical for fruitful gene expression. Recently, the mammalian EJC was shown to repress competing, cryptic, splice sites (SS). However, the evolutionary generality of this remains unclear. Here, I demonstrate the *Drosophila* EJC suppresses hundreds of functional cryptic SS, even though the majority of these bear weak splicing motifs and might appear incompetent. Mechanistically, the EJC directly conceals critical splicing elements by virtue of its position-specific recruitment, preventing SS definition. Unexpectedly, I discover the EJC inhibits scores of regenerated 5' and 3' recursive splice sites on segments that have already undergone splicing, and that loss of EJC regulation triggers faulty resplicing of mRNA. An important corollary is that certain intronless cDNA expression constructs yield high levels of unanticipated, truncated transcripts generated by resplicing. I conclude the EJC has conserved roles to defend transcriptome fidelity by (1) repressing illegitimate splice sites on pre-mRNAs, and (2) preventing inadvertent activation of such sites on spliced segments.

Introduction

Canonical splice sites contain instructive information across the exon/intron boundary. Cryo-EM structures of prespliceosomal complexes show that U1 snRNA establishes base contacts across the -2 to +6 position for a typical 5'SS, AG|GUAAGU (where | marks the exon/intron boundary) (Kondo et al., 2015). Similarly, the U2AF complex shows preference for a AG|GU motifs at 3'SS, which includes two nucletodies into the exon (Kielkopf et al., 2001). Moreover, exonic segments of splice sequences are also utilized during the catalytic stages of splicing, for example the juxtaposition of exon boundaries by U5 snRNA (Newman & Norman, 1992; Sontheimer & Steitz, 1993). Thus, when processed, exon junctions contain remnants of splice sequences. However, the activity of these segments post-splicing remains poorly explored.

It has been observed that exon junction sequences can function as cryptic splice sites (Dibb & Newman, 1989; Sadusky et al., 2004). This has led to one view of intron birth, in which they insert into cryptic or protosplice sites; sequences that are typically inactive but contain the information content required to pair with spliceosomal building blocks, such as U1 snRNP or U2AF (Kielkopf et al., 2001; Zhuang & Weiner, 1986). However, an alternate assessment is that intron removal may regenerate cryptic splice sites. This has been observed at cassette exons in the context of recursive splicing, but the recent discovery of suppressed, 5' recursive splice sites at constitutive exons junctions (Blazquez et al., 2018; Boehm et al., 2018) reignites this discussion by suggesting that even seemingly constitutive exons may regenerate cryptic splice sites at exon junctions. Furthermore, these studies showed that recruitment of the exon junction complex (EJC) silences the activity of cryptic splice sites.

The EJC is a multisubunit conglomerate that is deposited in a sequence-independent fashion ~24 nt upstream of exon-exon junctions (Boehm & Gehring, 2016; Hervé Le Hir et al., 2016). Assembly of its three-member core complex begins during

splicing, and the first step involves the position-specific deposition of the DEAD-box protein eIF4AIII onto RNA by the spliceosome factor CWC22. Next, a heterodimer of MAGOH/Mago Nashi and RBM8A/Y14/Tsunagi binds eIF4AIII, stabilizing the complex on RNA. The core EJC complex interacts with multiple peripheral complexes involved in diverse RNA metabolism pathways (Schlautmann & Gehring, 2020). Accordingly, EJC dysfunction broadly affects development, disease and cancer (Bonnal et al., 2020).

Curiously, while the EJC is well-conserved, the literature indicates fundamental differences in its requirements between invertebrates and vertebrates (Schlautmann & Gehring, 2020). The EJC was first linked to the process of nonsense mediated mRNA decay (V. N. Kim et al., 2001; Lykke-Andersen et al., 2001), a process that exploits deposition of the EJC by the spliceosome (H Le Hir et al., 2000). Translation removes EJCs from the open reading frame, but the presence of premature termination codons cause EJCs to remain within aberrant 3' UTRs, thereby triggering NMD. However, as introns do not inherently elicit NMD in *Drosophila*, its pathway does not appear to involve the EJC (Nicholson & Mühlemann, 2010).

The central connections between the EJC and the spliceosome (Singh et al., 2012) has also warranted attention towards splicing-related functions of the EJC. Here as well there is evidence for functional distinctions. In *Drosophila*, the EJC positively regulates splicing of long introns, such as *mapk* (Ashton-Beaucage et al., 2010; Roignant & Treisman, 2010), and also activates suboptimal splice sites, such as within *piwi* (Hayashi et al., 2014; Malone et al., 2014). By contrast, recent analysis of the mammalian EJC shows that many of its direct splicing targets are instead inhibited (Blazquez et al., 2018; Boehm et al., 2018), indicating a role in cryptic splice site avoidance during pre-mRNA maturation. However, the generality and scope of such a mode of splicing control remains poorly understood as only the latter reports examined *de novo* splicing.

Therefore, I analyzed the effects of the *Drosophila* EJC on splicing in greater detail. Although *Drosophila melanogaster* has one of the best annotated metazoan transcriptomes (Brown et al., 2014; Sanfilippo et al., 2017; Westholm et al., 2014), I unexpectedly detect many hundreds of novel splice junctions upon depletion of core EJC components in a single celltype. As in mammals, *de novo* splicing analysis demonstrates the fly EJC protects neighboring introns from cryptic splice site activation. This function is required under unusual circumstances including out-of-order splicing and appears to rely on occlusion of competing, weak splice sites. Next, I identify scores of splice defects that arise from cryptic splice sites at exon junction sequences. Two key sources of evidence implicate exon junction sequences as sources of cryptic splice sites. First, I validate that cryptic splice donors and acceptors are regenerated at exon junctions. Second, I elucidate that even poor matches to consensus splice motifs can act as functional splice sites at exon junctions. While these sites are suppressed on pre-mRNAs, I find that silencing is also required on mRNAs to prevent further resplicing. My results suggest that exon junction sequences are a source of cryptic 5' and 3' SS, and provides the basis for an intrinsic requirement of the EJC to suppress accidental activation. Overall, my findings broaden a newly appreciated, ancestral function of the EJC, and emphasize that bypass of this regulatory process via cDNA constructs can have unexpected deleterious consequences.

Results

EJC depletion leads to activation of spurious junctions

Recently, Roignant and colleagues reported RNA-seq datasets from S2 cells depleted for core EJC factors *eIF4AIII*, *tsu* (Y14) and *mago* (Akhtar et al., 2019). I re-examined these data for splicing defects, and paid particular attention to spurious splice site usage. I utilized MAJIQ to acquire currently unannotated junctions (3606 novel

splice sites supported by ≥5 split reads in the aggregate data), of which 1677 were >2-fold upregulated in at least one EJC-KD condition. As the three core EJC factors are mutually required for stable EJC association at exon-exon junctions, it is reasonable to expect these to reveal a set of common molecular defects. Indeed, there was both substantial and significant overlap in novel junctions amongst all three conditions (*p-value* < $1x10^{-8}$ for three-way overlap), and 876 junctions were elevated in two out of three EJC-KD datasets (**Figure 4.1A**). To introduce further stringency, I also filtered for >2-fold PSI change in 2/3 EJC depletions, yielding 573 spurious junctions from 386 genes (**Figure 4.1B)**. These genes are diverse, with gene ontology (GO) analysis comprising diverse cellular processes including system development and signaling.

The most frequent spurious junctions involved activation of exonic, alternative 5' or 3' SS, followed by novel alternative splicing and intronic SS activation (**Figure 4.2A**). These are expected to delete exonic sequence (alternative 5' or 3' SS) or insert intronic sequence (intronic SS), relative to canonical mRNA products. I depict *straw* as an example of aberrant splicing occuring at a constitutive exon-exon junction (**Figure 4.2B**). Here, depletion of *eIF4AIII*, *tsu* and *mago*, but not *lacZ* control, all induced high-frequency usage of a novel exonic, alternative 5' SS that joins to the constitutive 3' SS 3248 nt downstream. Importantly, this presumably defective transcript comprises the major isoform in all three core-EJC knockdowns, as it removes 91 nt of coding sequence and is thus out of frame.

I used rt-PCR to validate *de novo* splice isoforms in EJC-depleted S2 cells. I selected transcripts with high activation of exonic 5' and 3' SS (PSI > 0.2), such as *straw, multiple ankyrin repeats single KH domain (mask), baboon* and *eukaryotic translation initiation factor 4G1 (eIF4G1)*, but also evaluated targets with moderate changes (0.01< PSI < 0.05) such as *Crk oncogene and unkempt*. As EJC stabilization during pre-mRNA processing requires *eIF4AIII*, *tsu* and *mago*, but not *btz*, I utilized knockdown of *btz* and

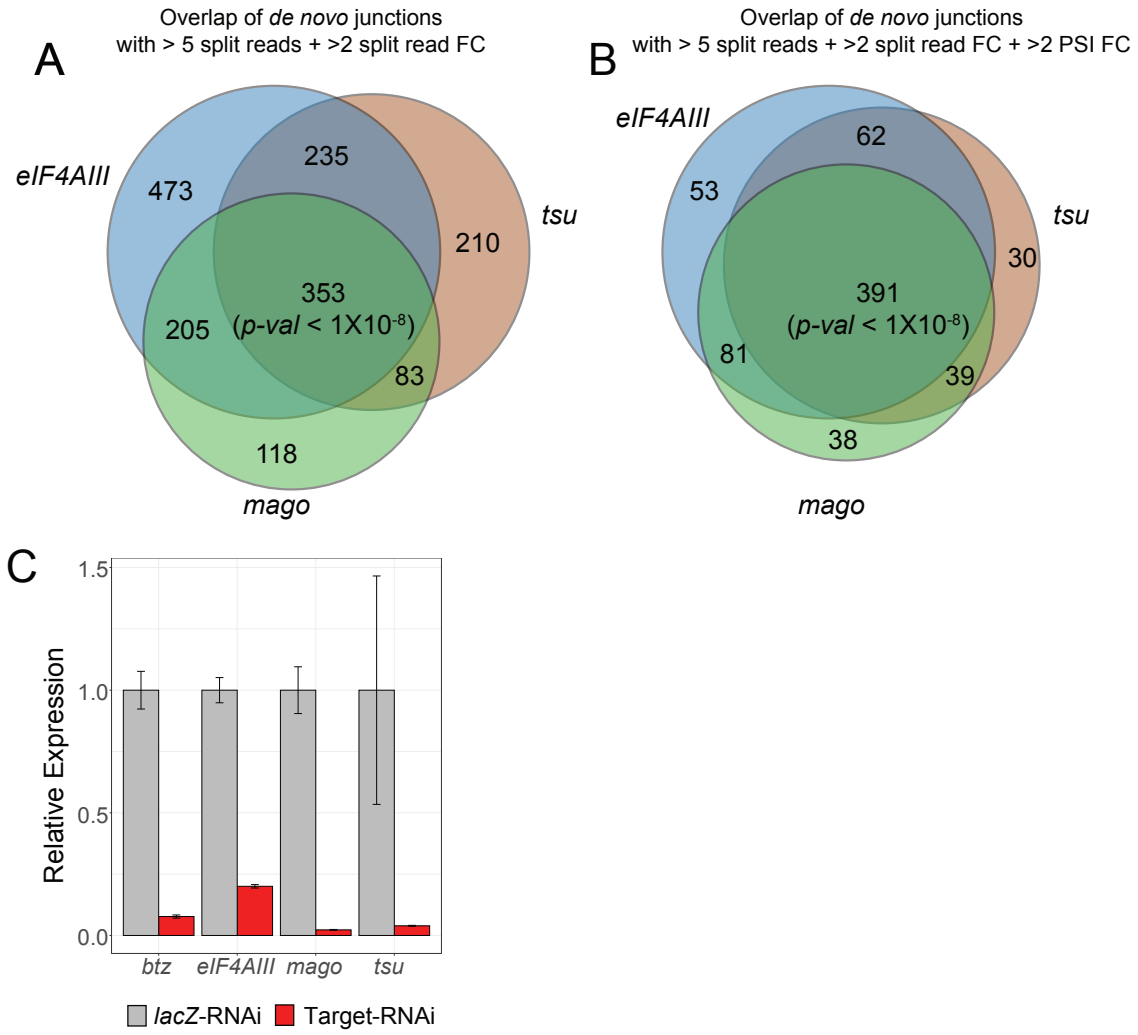


FIgure 4.1. core-EJC depletion yields broad activation of *de novo* splice junctions. (A) Strong overlap of *de novo* splice junctions between core-EJC knockdown conditions based. The Venn diagram depicts which of 1677 junctions with at least 5 split reads had > 2-fold split read changes between treatment and controls.p-value for three-way overlap was calculated using a permutation test with 10^8 tests. (B) Strong overlap of high-confidence *de novo* splice junctions between core-EJC knockdown conditions. The Venn diagram depicts which of 876 junctions with at least 5 split reads and > 2-fold split read changes also show > 2-fold changes in percent selected index (PSI) between treatment and controls. p-value for three-way overlap was calculated using a permutation test with 10^8 tests. (C) Knockdown of EJC factors in S2 cells using dsRNA. quantitative rt-PCR of core-EJC and *btz* transcripts after dsRNA treatment.

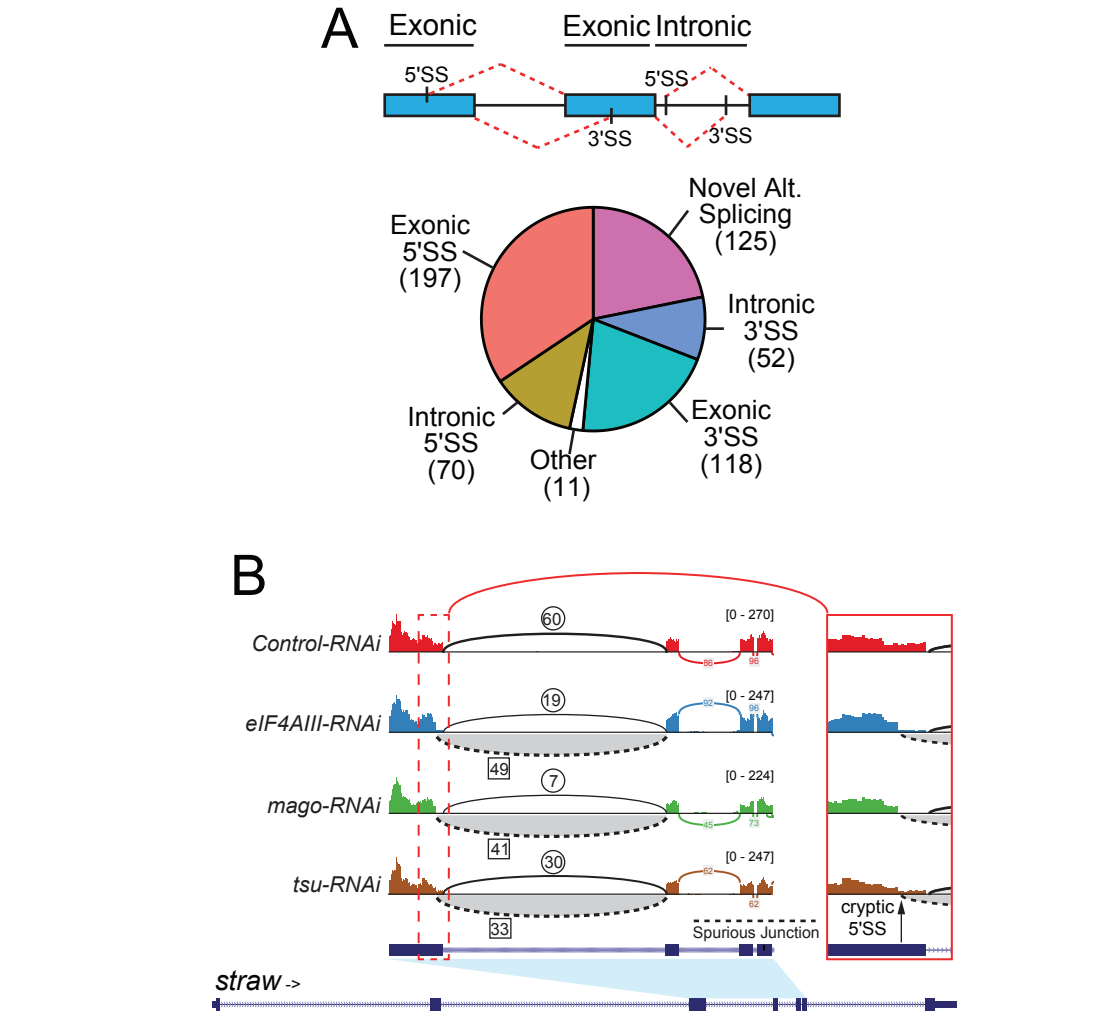

A
Exonic
Exonic Intronic
5'SS
5'SS
3'SS
3'SS
Novel Alt. Splicing (125)
Exonic 5'SS (197)
Intronic 3'SS (52)
Intronic 5'SS (70)
Other (11)
Exonic 3'SS (118)
B
Control-RNAi
eIF4AIII-RNAi
mago-RNAi
tsu-RNAi
[0 - 270]
[0 - 247]
[0 - 224]
[0 - 247]
60
19
49
7
41
30
33
Spurious Junction
cryptic 5'SS
straw ->
C
straw
mask
eIF4G1 (5'UTR)
Haspin
baboon (babo)
eIF4G1 (CDS)
unkempt (unk)
Crk oncogene (Crk)
RNAi:
mago
eIF4AIII
tsu
btz
lacZ

Figure 4.2. Transcriptome-wide *de novo* alternative splicing upon depletion of functional Exon Junction complex
(A) Overview of upregulated *de novo* splice junctions in EJC-depleted cells. Top: schematic of exonic and intronic cryptic 5' and 3' SS. Bottom: Pie chart indicating the distribution of different splice junction classes.
(B) Sashimi plot depicting HISAT2-mapped sequencing coverage along a portion of *straw*, which has defective splicing under core-EJC LOF. The gene model depicts the location of the cryptic 5' SS relative to the annotated 5' SS. Junction spanning read counts mapping to the canonical junction are circled, whereas cryptic junction read counts are squared. Note that spliced reads mapping to the cryptic junction are found in *eIF4AIII*-, *mago*- and *tsu*-KD but not the control comparison. Region containing the cryptic 5'SS has been zoomed on the right.
(C) Validation of *de novo* splicing events in core-EJC depleted cells. EJC core components (*eIF4AIII*, *mago*, *tsu* and *btz*) were knocked down in *Drosophila* S2 cells using dsRNA. After knockdown, eight targets identified in (A) were evaluated using an rt-PCR assay and demonstrated splicing defects (asterisk). Importantly, only core-EJC factor KD produced cryptic bands, but not *btz* or control conditions. Note that several splicing defects are observed for *unkempt (unk).*

lacZ as controls (**Figure 4.1C**). For all eight amplicons tested, I observed splicing defects only under core-EJC (*eIF4AIII*, *tsu* and *mago*) knockdown conditions (**Figure 4.2C**). These data provide stringent validation of my annotation of spurious junctions, and highlight a previously unappreciated quality control function of the *Drosophila* EJC.

The EJC suppresses cryptic exonic 3' SS during pre-mRNA processing

These alterations in transcript processing were reminiscent of how the human EJC, recruited to exon junctions, directly influences the splicing of neighboring introns (Blazquez et al., 2018; Boehm et al., 2018). Accordingly, I examined the mechanism of EJC-regulated splicing defects in *Drosophila*. I began by examining transcripts with spurious exonic 3' SS. These represent a majority of *de novo* events observed in my analysis, and are predicted to cause broad loss of mRNA sequences. Cryptic 3' SS exhibit strong positional bias and cluster specifically around exon junctions **(Figure 4.3A**). However, while cryptic 3' SS contain the invariant 3' AG dinucleotide (**Figure 4.4A**), quantitative assessment of SS strength indicated broad variation (**Figure 4.3A**). In fact, most activated 3' SS in this category are extremely weak and would not normally be considered functionally competent, especially when considering their sheer frequency in the transcriptome at large. Thus, it was important to manipulate these RNA substrates to understand their splicing capacities more directly.

I selected *CG7408* as a paradigm: it reproducibly exhibited defective splice isoforms in all core-EJC knockdowns (**Figure 4.3B**), but its putative 3' SS is extremely weak (NNSPLICE score of 0.29, **Figure 4.3A**) and poorly conserved (**Figure 4.4B**). I used rt-PCR to validate the expected transcript defects in EJC-depleted cells (**Figure 4.3C)**, and confirmed 183 nt exon deletion relative to the canonical splice isoform via Sanger sequencing. The cryptic junction replaces intron 1, where canonical splicing typically utilizes one of three annotated 3' SS, the dominant of which is stereotypically

strong (**Figure 4.4B**, NNSPLICE score of 0.91). I then constructed a minigene bearing exons 1-4 of *CG7408* (**Figure 4.3D**, genomic). When transfected into S2 cells, this reporter recapitulated normal splicing through activation of annotated 3' SS (**Figure 4.3E**, genomic). Importantly, a "fully pre-spliced" reporter lacking all introns, i.e., mimicking an mRNA expression construct, yielded a single normal product (**Figure 4.3D-E**, mRNA). Thus, pre-processed *CG7408* transcripts that cannot recruit EJC, also do not undergo further processing. At face value, this appears consistent with the hypothesis that the EJC regulates splicing of flanking introns.

I explored this further by testing for potentially distinct consequences of EJC recruitment to individual *CG7408* exon junctions, by removing each intron in turn (**Figure 4.3D** - Δi1, Δi2 and Δi3). These manipulations should only abolish EJC recruitment at individual pre-processed exon junctions. Δi1 only produced the dominant canonical isoform and Δi3 produced the two known canonical isoforms at the same proportions as the genomic construct (**Figure 4.3E**). By contrast, pre-removal of intron 2 yielded fully aberrant transcripts (**Figure 4.3E**, Δi2). These tests emphasize the functional requirement of intron 2 for correct processing of *CG7408* and demonstrate that even poor matches to consensus splice sites (i.e., the *CG7408* cryptic 3' SS) can be potently activated in the absence of the EJC.

I emphasize that these data support a mechanism in which intron 2 is excised first, and this order is required for the correct definition of the annotated intron 1 3' SS (**Figure 4.3G**). Out-of-order splicing has been previously observed (Drexler et al., 2020; Khodor et al., 2011; LeMaire & Thummel, 1990; Pandya-Jones & Black, 2009; Takahara et al., 2002), but has mostly been documented only as a phenomenon. It has generally been unclear if out-of-order splicing has functional impact on accurate pre-mRNA maturation. These experiments, along with recent work by Gehring and colleagues

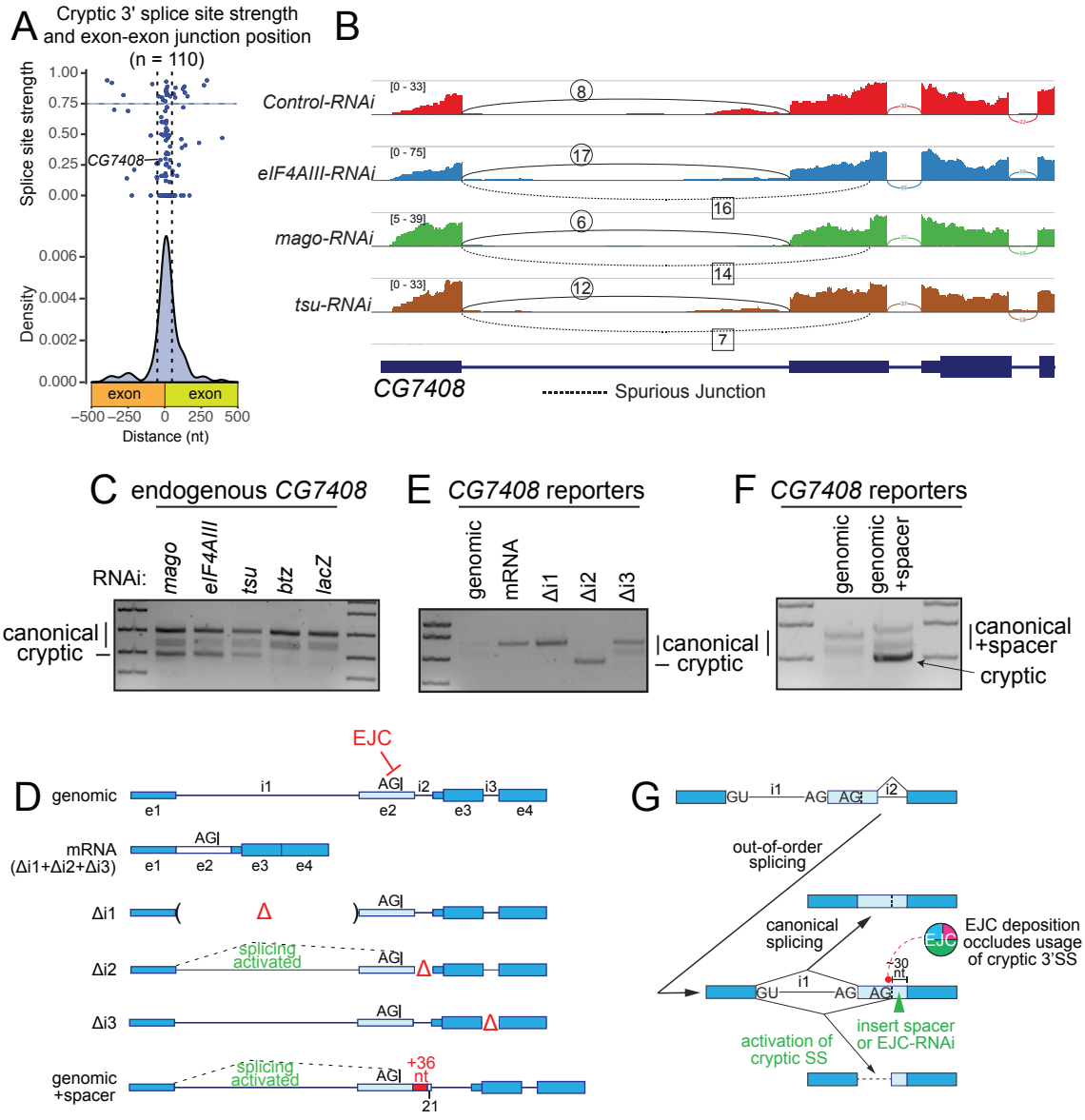

A
Cryptic 3' splice site strength
and exon-exon junction position
(n = 110)
Splice site strength
1.00
0.75
0.50
0.25
0.00
CG7408
Density
0.006
0.004
0.002
0.000
exon
exon
−500 −250 0 250 500
Distance (nt)
B
Control-RNAi
eIF4AIII-RNAi
mago-RNAi
tsu-RNAi
[0 - 33]
[0 - 75]
[5 - 39]
[0 - 33]
8
17
16
6
14
12
7
CG7408
Spurious Junction
C endogenous CG7408
RNAi:
mago
eIF4AIII
tsu
btz
lacZ
canonical
cryptic
E CG7408 reporters
genomic
mRNA
Δi1
Δi2
Δi3
canonical
cryptic
F CG7408 reporters
genomic
genomic +spacer
canonical +spacer
cryptic
D
EJC
genomic
i1
i2
i3
e1
e2
e3
e4
AG
mRNA
(Δi1+Δi2+Δi3)
Δi1
Δi2
Δi3
splicing activated
genomic +spacer
+36 nt
21
G
GU
AG
out-of-order splicing
canonical splicing
EJC deposition occludes usage of cryptic 3'SS
30 nt
activation of cryptic SS
insert spacer or EJC-RNAi

Figure 4.3. EJC-depletion leads to activation of cryptic 3' splice sites
(A) Depiction of 3' SS position of spurious junctions relative to exon-exon boundaries as density and dot plot. The dot plot indicates splice site scores as calculated via NNSPLICE. Horizontal dashed line depicts threshold for strong 3' SS, and vertical dashed lines specify 50 nt flanking exon-exon junctions.
(B) Sashimi plot depicting HISAT2-mapped sequencing coverage along a portion of *CG7408*, which has a cryptic 3' SS that is activated under core-EJC LOF. Junction spanning read counts mapping to the canonical junction are circled, whereas cryptic junction read counts are squared. Note that spliced reads mapping to the cryptic junction are found in *eIF4AIII*, *mago* and *tsu* KD but not the control comparison.
(C) Validation of *CG7408* cryptic 3' SS activation in core-EJC, but not *btz* or *lacZ* KD conditions
(D) Schematic of *CG7408* splicing reporters. Exons 1-4 (introns included) were cloned and subjected to further manipulation. Locations of pre-removed introns (∆), as well as a construct lacking all introns (mRNA) are included. For reference, the position of the cryptic 3'SS is marked on exon 2. genomic+spacer represents a modified version of the genomic splicing reporter with an insertion of 36 nt spacer sequence on exon 2.
(E) rt-PCR of reporter (D) constructs ectopically expressed in S2 cells demonstrates that intron 2 is required for accurate processing of the minigene. Canonical and cryptic products are indicated.
(F) Cryptic splicing is detected with the inclusion of a 36 nt spacer sequence.
(G) Schematic of out-of-order splicing and positional requirement of the core-EJC for accurate 3' SS definition.

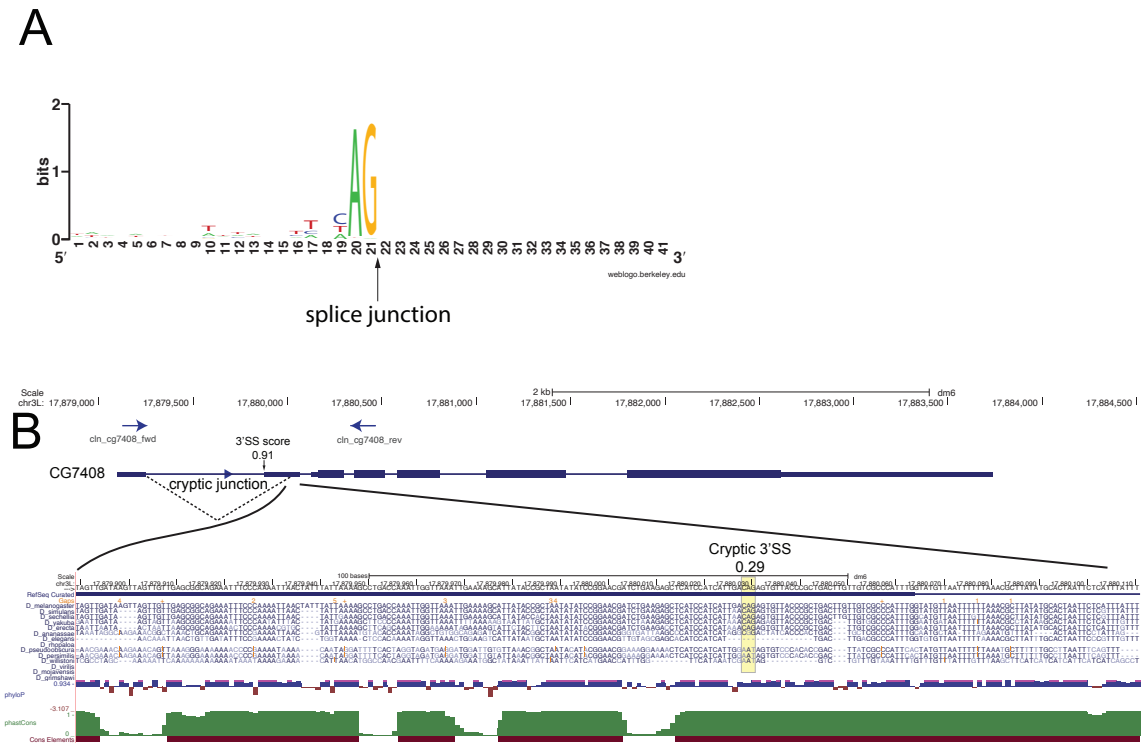


Figure 4.4. A majority of cryptic 3' SS activated under EJC-loss are weak
(A) Nucleotide content of cryptic 3' SS. These sequences, apart from the invariant AG dinucleotide show poor strength.
(B) Example of a weak cryptic 3' SS (NNSPLICE score of 0.29) found on the *CG7408* transcript. Conservation of the weak splice site is depicted using the multiple alignment format on the UCSC genome browser, as well as phyloP and phastCons scores.

(Boehm et al., 2018) indicate a requirement for out-of-order splicing for proper mRNA maturation.

How does the EJC inhibit definition of cryptic exonic 3' SS? In human cells, the EJC can directly mask cryptic 3' SS. Based on the close clustering of these sites around the position of EJC recruitment (**Figure 4.3A**), I reasoned that the EJC may occlude important features of the 3' SS, such as the branchpoint, polypyrimidine tract or 3' intron junction that base-pairs with the U2 snRNP complex. I tested this hypothesis by separating the cryptic 3' SS on my genomic reporter from the site of EJC recruitment, by inserting a 36 nt spacer (**Figure 4.3D**, genomic+Spacer). Unlike the genomic construct, which yields only annotated splice isoforms, the genomic+Spacer variant yielded additional truncated transcripts, consistent with derepression of the cryptic 3' SS (**Figure 4.3F**). Altogether, these data demonstrate the fly EJC aids accurate SS selection during pre-mRNA processing by masking cryptic 3' SS.

The EJC prevents cryptic exonic 5' SS activation during pre-mRNA processing

I next used analogous strategies to study cryptic exonic 5' SS. These sites represent ~35% of novel splice junctions upregulated under EJC-depleted conditions and are expected to be deleterious to mRNA processing fidelity. Bioinformatic analysis indicated that cryptic 5' SS share general structural properties with 3' SS, such as clear preference in the vicinity of exon junctions but distribution across a wide range of strengths (**Figure 4.5A, Figure 4.6A**).

I selected *CG3632* for mechanistic tests, as core-EJC knockdown data showed activation of a poorly conserved, weak cryptic 5' SS (**Figure 4.6B** and **Figure 4.5B-C** – NNSPLICE score of 0.54) on exon 14. Using rt-PCR and Sanger sequencing, I validated that EJC-depletion induces a defective *CG3632* splice isoform lacking 71 nt of coding sequence (**Figure 4.6C**).

I hypothesized that the EJC, recruited to the exon 13/14 junction, suppresses the cryptic 5' SS on exon 14 and activates the canonical 5' SS during removal of intron 14. I tested this using a minigene reporter consisting of exon 14 (containing the cryptic 5' SS) and its immediately flanking introns and exons (**Figure 4.6D**, genomic). Expression of this reporter in S2 cells predominantly resulted in the canonical product, but I also observed a minor amount of cryptic 5' SS activation (**Figure 4.6E**, genomic). As a negative control, I generated a version lacking both introns (**Figure 4.6D**, Δi13+14), which produced the expected mRNA (**Figure 4.6E**, Δi13+14). Notably, removal of intron 13 alone (**Figure 4.6D**, Δi13), mimicking loss of EJC recruitment at the exon 13/14 junction, yielded high levels of cryptic 5' SS activation (**Figure 4.6E**, Δi13) that were fully suppressed by mutation of the cryptic 5' SS in the Δi13 reporter (**Figure 4.6D-E**, Δi13+SD mut). Altogether, these data support that deposition of the EJC during pre-mRNA processing suppresses cryptic 5' SS during subsequent intron removal.

The EJC suppresses recursive splice sites

Given that the EJC suppresses both 5' and 3' SS, a potentially more complex scenario might exist if both types of cryptic splice sites were to be activated in the vicinity of each other. I inspected my catalog of spurious junctions for this possibility, and considered that even modest matches to consensus splice sites (**Figure 4.3A and Figure 4.6A**) might serve as viable candidates for further evaluation. Interestingly, many sequences at exon junctions were potentially able to regenerate weak splice sites after intron removal, reminiscent of the process of recursive splicing (RS) (Burnette et al., 2005).

I first investigated a spurious junction within *Casein kinase IIß* (*CkIIß*), where core-EJC LOF led to loss of 54 nt of canonical mRNA sequence (**Figure 4.7A-B**). Assessment of the novel 3' SS on exon 3 revealed that it lacks a polypyrimidine tract and

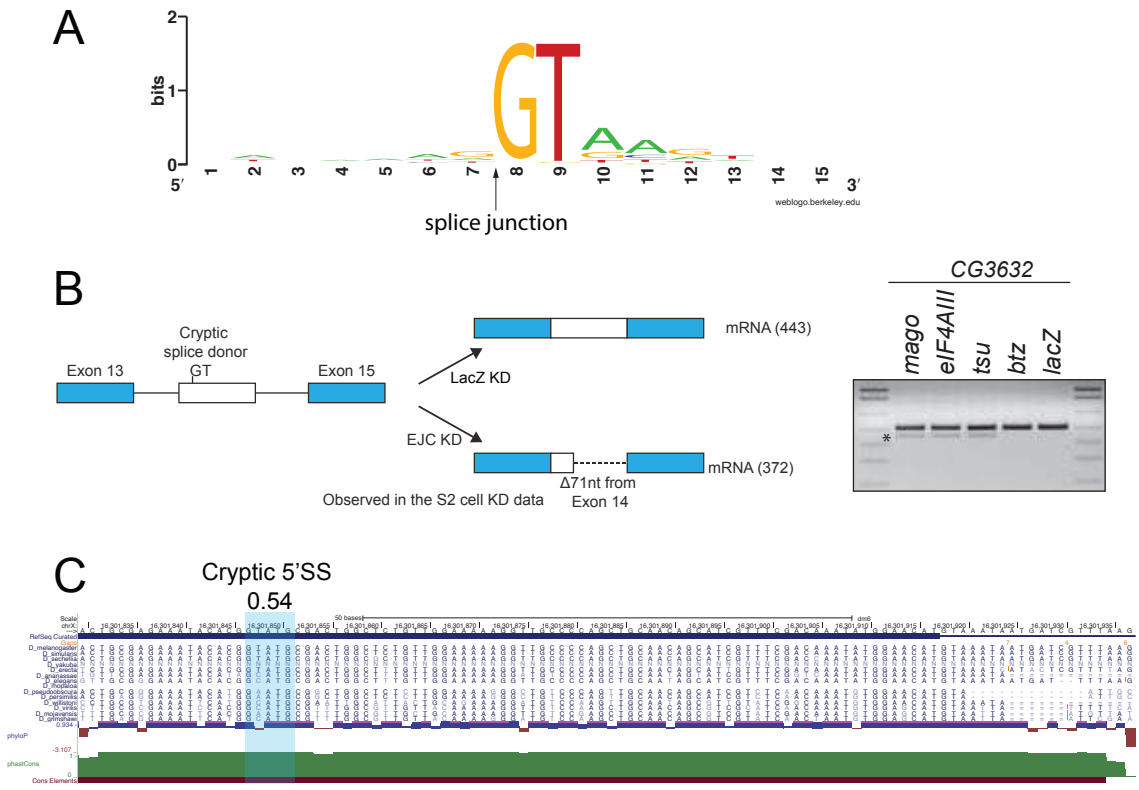


Figure 4.5. A majority of cryptic 5' SS activated under EJC-loss are weak
(A) Nucleotide content of cryptic 5' SS.
(B) Schematic of a de novo splicing event detected on the *CG3632* transcript. Validation of splicing defects shown on the right.
(C) Cryptic 5' SS (NNSPLICE score of 0.54) found on the *CG3632* transcript. Conservation of the weak splice site is depicted using the multiple alignment format on the UCSC genome browser, as well as phyloP and phastCons scores.

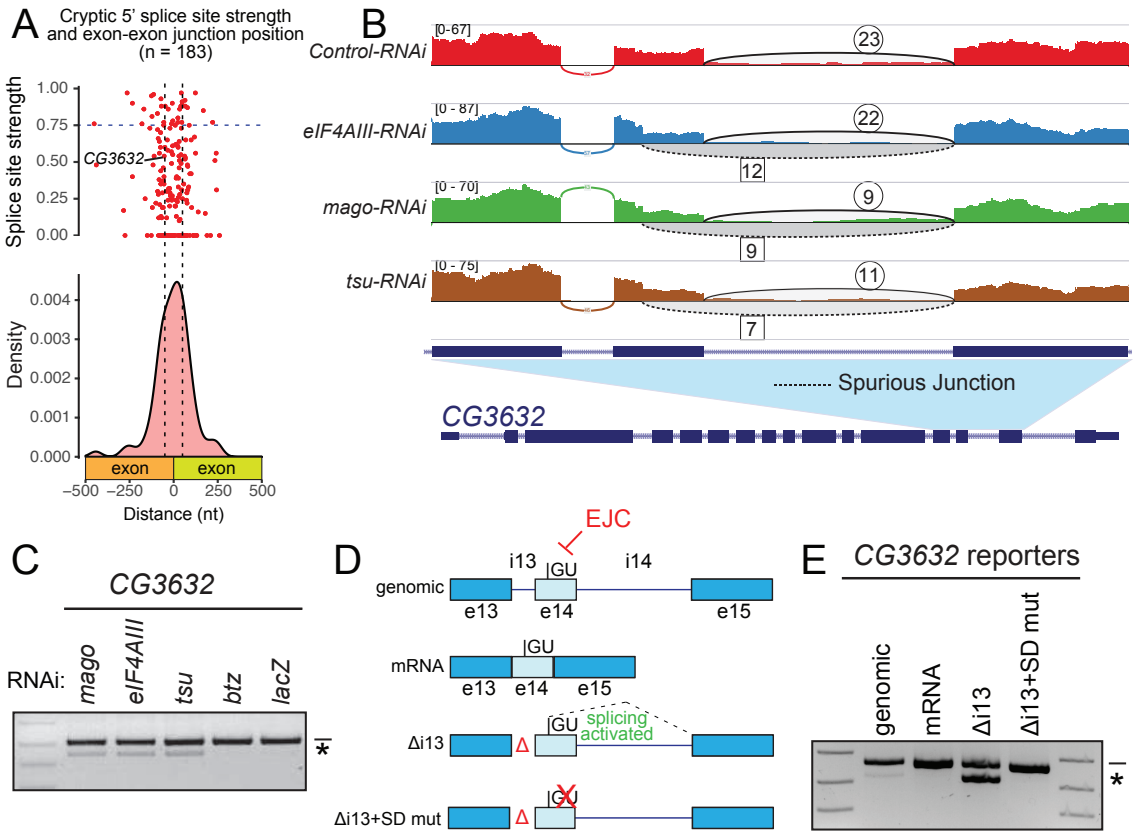


Figure 4.6. EJC-depletion leads to activation of cryptic 5' splice sites.
(A) Metagene of cryptic 5' SS position relative to exon-exon boundaries as density and dot plot. The dot plot indicates splice site scores as calculated via NNSPLICE (see Methods). Horizontal solid line depicts a threshold for strong 5' SS.
(B) Sashimi plot depicting HISAT2-mapped sequencing coverage along a portion of CG3632, which has a cryptic 5' SS that is activated under core-EJC LOF. Junction spanning read counts mapping to the canonical junction are circled, whereas cryptic junction read counts are squared. Note that spliced reads mapping to the cryptic junction are found ineIF4AIII, mago and tsu KD but not the control comparison.
(C) Validation of CG3632 cryptic 5' SS activation (asterisk) in core-EJC, but not btz or lacZ KD conditions.
(D) Schematic of CG3632 splicing reporters. Exons 13-15 (introns included) were cloned and subjected to further manipulation. Locations of pre-removed introns (Δ), as well as a construct lacking all introns (mRNA) are included. The position of the cryptic 5'SS is marked on exon 14, and was mutated in Δi3+SD mut.
(E) rt-PCR of reporter (D) constructs expressed in S2 cells demonstrates that intron 13 is required for accurate processing of the minigene. Canonical products are indicated by the line and cryptic products by an asterisk.

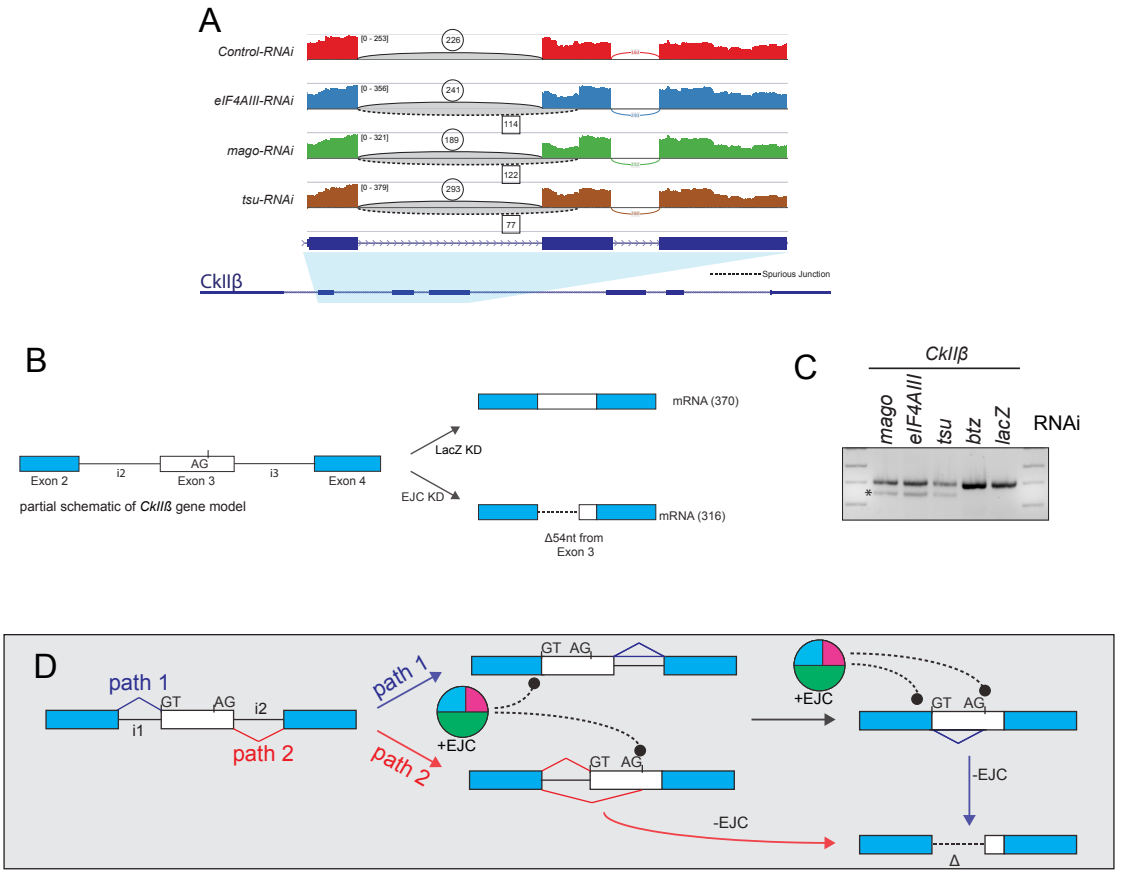


Figure 4.7. de novo splicing on *CkIIß* is a result of dual cryptic splice site activation.

(A) Sashimi plot depicting HISAT2-mapped sequencing coverage along a portion of *CkIIß*, which has a cryptic 3' SS that is activated under core-EJC LOF. Junction spanning read counts mapping to the canonical junction are circled, whereas cryptic junction read counts are squared. Note that spliced reads mapping to the cryptic junction are found in *eIF4AIII*, *mago* and *tsu* but not the control comparison.

(B) Schematic of a de novo splicing event detected on the *CkIIß* transcript.

(C) Validation of *CkIIß* cryptic 3' SS activation in core-EJC, but not *btz* or *lacZ* KD conditions

(D) Models that explain the *CkIIß* splicing defects. Path 1 and 2 reflect alternate orders of intron removal. Crucially, path 1 leads to EJC-suppressed cryptic splicing on mRNAs using the indicated 5' recursive splice site and a cryptic 3'SS, whereas path 2 can also produce a splice defect after removal of intron 2.

is a poor match to the consensus (**Figure 4.8A**). On the surface, the mechanism of cryptic 3' SS activation on *CkIIβ* might appear similar to that of *CG7408* (**Figure 4.3F, Figure 4.7D, path 2**). However, upon examining *CkIIβ* for splice sites, I found an additional poor recursive 5' SS at the beginning of exon 3 (**Figure 4.8A**). Therefore, I imagined an alternate scenario, whereby dual cryptic 5' and 3' SS might be derepressed upon EJC loss, leading to resplicing (**Figure 4.7D, path 1**). Crucially, whether one-step splicing (via alternative splicing) or resplicing (via recursive splicing event), the resulting mRNA products are indistinguishable (**Figure 4.8A**). Therefore, I devised reporter tests to clarify the underlying mechanism.

I first used rt-PCR to validate that core-EJC knockdown resulted in substantial activation of a truncated *CkIIß* splice isoform corresponding to RNA-seq data (**Figure 4.7C**). I then analyzed a series of splicing minigenes (**Figure 4.8B**). Expression of *CkIIβ* exons 2-4 with all introns present produced a single product with the expected introns spliced out (**Figure 4.8C**, genomic). I precisely tested the positional necessity of the EJC at each exon junction by pre-removing each intron (**Figure 4.8B**, Δi2 and Δi3). These reporters also underwent normal splicing (**Figure 4.8C**, Δi2 and Δi3), demonstrating that *CkIIβ* processing defects were in fact mechanistically distinct from those determined for *CG7408.* Strikingly, upon testing a construct with both introns pre-removed, I observed a switch to truncated product output, corresponding to activation of the unannotated recursive 5' SS and 3' SS (**Figure 4.8B-C**, mRNA). This supports a model where the EJC is required at multiple positions to repress spurious 5' and 3' SS simultaneously (**Figure 4.7D, path 1**).

I characterized another instance of dual cryptic splice site within *CG31156*, albeit of a different flavor. Here, sashimi plots indicate activation of an exonic 5' SS within exon 2 (**Figure 4.9A-B**) and I validated this 110 nt deletion isoform using rt-PCR (**Figure 4.8D**). Importantly, based on these data alone, it would be reasonable to predict this as a

case of alternative cryptic 5' SS activation. However, I noticed that removal of the canonical intron 2 regenerates a putative recursive 3' SS at the exon 2/3 boundary (**Figure 4.8E, Figure 4.9C**). Therefore, I examined reporters to examine the mechanism underlying this unwanted splicing pattern. Expression of the genomic reporter that required intron removal yielded the expected mRNA product (**Figure 4.8F**, genomic). Conversely, pre-removal of the intron and expression of the mRNA resulted in the truncated re-spliced product (**Figure 4.8F**, mRNA). Accordingly, these data again indicate that the EJC represses dual cryptic splice sites during mRNA processing (**Figure 4.9D**).

Cryptic recursive splice sites suppressed by the EJC exhibit atypical properties

I emphasize that these instances of recursive splice sites (RSS) are quite distinct from those studied previously in *Drosophila*. Fly genomes are known to contain hundreds of RSS for which the hybrid 5'/3' splice sites are highly conserved, flanked by short cryptic downstream exons, and highly biased to reside in long introns (mean length ~50 kb) (Duff et al., 2015; Joseph et al., 2018; Pai et al., 2018). It has been suggested that recursive splicing aids processing of long introns; however, it is also conceivable that it is easier to capture RS intermediates within long introns. The examples of cryptic RSS on the *CkIIβ* and *CG31156* transcripts clearly deviate from canonical RSS architectural properties, i.e., they are hosted in short introns and exhibit modest to poor conservation. Moreover, the example of a recursive 3' SS in *CG31156* is to my knowledge the first validated instance, and represents a conceptually novel RSS location. Importantly, the relevant AG dinucleotide in the *CG31156* 3' recursive splice site is not preserved beyond the closest species in the melanogaster subgroup (**Figure 4.9C**), and the amino acids encoded by the functional 5' RSS in *CkIIβ* diverge with clear

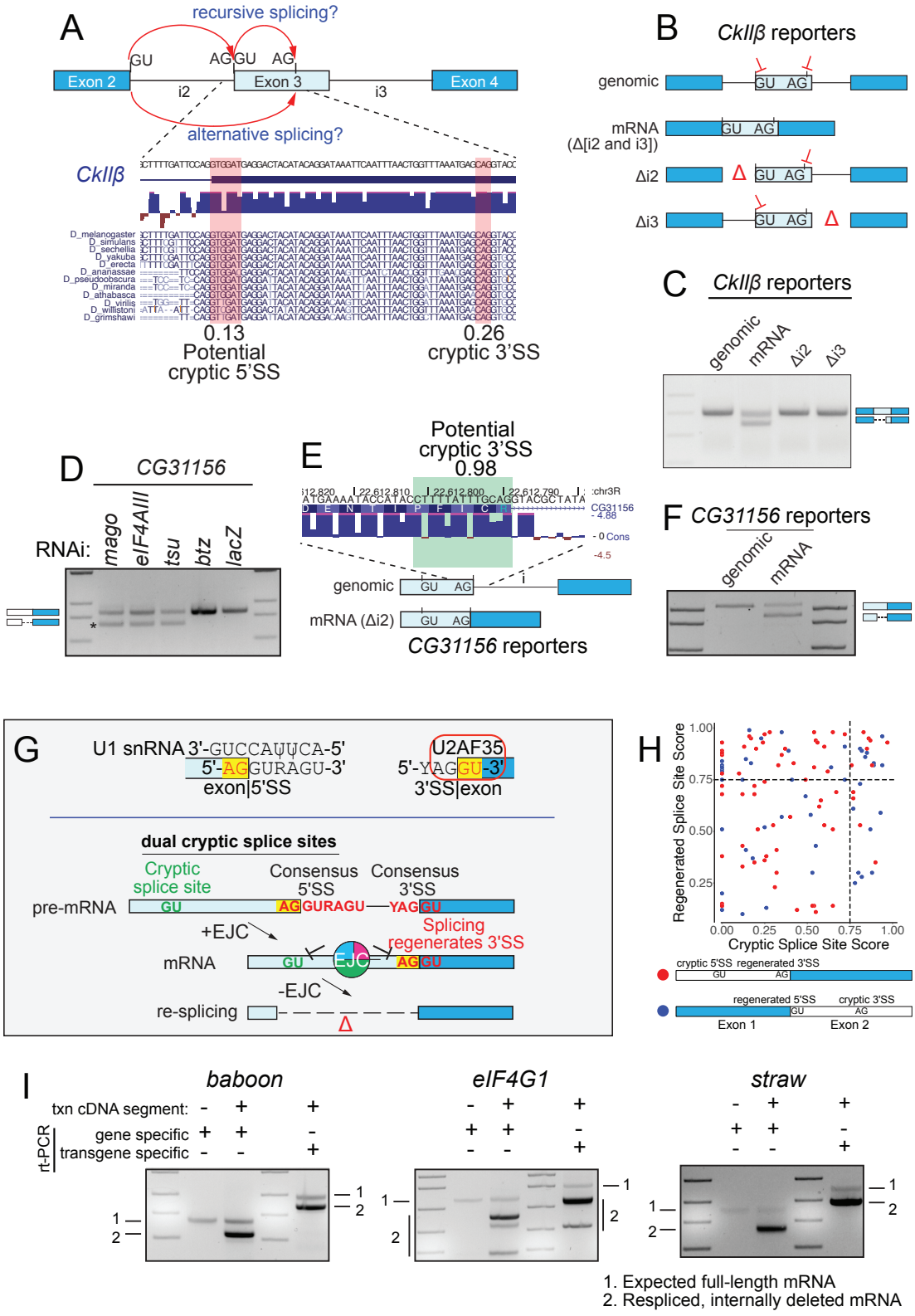

A
recursive splicing?
GU
AG GU
AG
Exon 2
Exon 3
Exon 4
i2
i3
alternative splicing?
CkIIβ
0.13
Potential
cryptic 5'SS
0.26
cryptic 3'SS
B
CkIIβ reporters
genomic
GU AG
mRNA
(Δ[i2 and i3])
Δi2
Δ
Δi3
C
CkIIβ reporters
genomic
mRNA
Δi2
Δi3
D
CG31156
RNAi:
mago
eIF4AIII
tsu
btz
lacZ
E
Potential
cryptic 3'SS
0.98
CG31156
genomic
mRNA (Δi2)
CG31156 reporters
F
CG31156 reporters
genomic
mRNA
G
U1 snRNA 3'-GUCCAUUCA-5'
5'-AGGURAGU-3'
exon|5'SS
U2AF35
5'-YAGGU-3'
3'SS|exon
dual cryptic splice sites
Cryptic
splice site
Consensus
5'SS
Consensus
3'SS
pre-mRNA
GU
AG GURAGU
YAG GU
+EJC
Splicing
regenerates 3'SS
mRNA
EJC
-EJC
re-splicing
Δ
H
Regenerated Splice Site Score
Cryptic Splice Site Score
cryptic 5'SS
regenerated 3'SS
regenerated 5'SS
cryptic 3'SS
Exon 1
Exon 2
I
baboon
eIF4G1
straw
txn cDNA segment:
rt-PCR
gene specific
transgene specific
1. Expected full-length mRNA
2. Respliced, internally deleted mRNA

Figure 4.8. EJC-depletion leads to activation of dual cryptic splice sites and resplicing of mRNA

(A) Above: Schematic of resplicing splicing versus alternative resplicing, both of which would yield the same aberrant mRNA product. Below: Sequence of *CkIIβ* transcript lost due to cryptic splicing. Cryptic 3' SS activated is highlighted in red, as well as a potential regenerated 5' SS. Scores listed are generated by NNSPLICE. Conservation across Drosophilid family is shown.

(B) Schematic of *CkIIβ* splicing reporters. Exons 2-4 (introns included) were cloned and subjected to further manipulation. Locations of pre-removed introns (Δ), as well as a construct lacking all introns (mRNA) are included. For reference, the position of the cryptic 3' SS and potential 5' recursive splice sites is marked on exon 3.

(C) rt-PCR of *CkIIβ* reporter constructs in S2 cells demonstrates that introns are required for accurate processing of the minigene. Canonical and cryptic products are indicated.

(D) Validation of *CG31156* cryptic 5' SS activation in core-EJC, but not *btz* or *lacZ* KD conditions

(E) Schematic of *CG31156* splicing reporters with and without introns. Location of potential 3’ recursive splice site on exon 2 is indicated along with conservation scores.

(F) rt-PCR of reporter constructs in S2 cells demonstrates that introns are required for accurate processing of the minigene. Canonical and cryptic products are indicated.

(G) Model for mRNA re-splicing. Top, Binding sites of U1 snRNA and U2AF35 define the 5' SS and 3' SS, respectively, but also impose constraints on flanking exonic sequences that intrinsically regenerate splice site mimics in a recursive fashion. (Bottom) When located in proximity to another cryptic splice site, these can lead to mRNA resplicing in the absence of the EJC. An example of dual cryptic splice sites with a regenerated 3' SS is shown, but this can also occur with a regenerated 5' SS.

(H) Comparison of splice site strengths for cases of dual cryptic splice site activation. Cases that contain regenerated 3' and 5' splice sites at exon junctions and their structures are schematized and distinguished by red and blue dot. Dashed lines mark thresholds for reasonably strong splice sites.

(I) Re-splicing on cDNAs. Constructs bearing cDNA segments of *baboon, eIF4G1* and *straw* were expressed in S2 cells and yielded re-spliced amplicons. Gene specific primers that amplify endogenous and ectopic products only show re-splicing from the intron-less reporter. Transgene-specific primers demonstrate mostly re-spliced products.

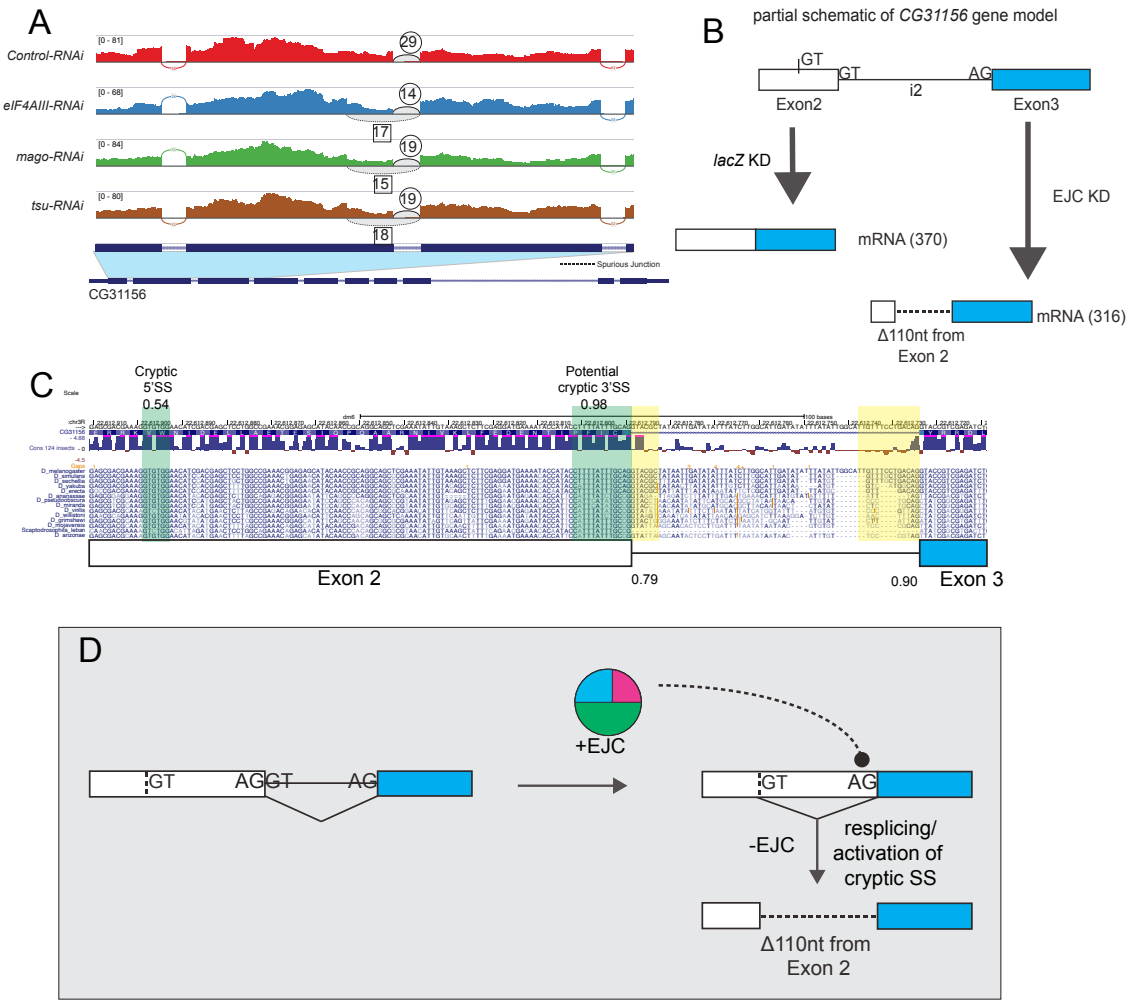


Figure 4.9. de novo splicing on *CG31156* is a result of dual cryptic splice site activation.
(A) Sashimi plot depicting HISAT2-mapped sequencing coverage along a portion of *CG31156*, which has a cryptic 5' SS that is activated under core-EJC LOF. Junction spanning read counts mapping to the canonical junction are circled, whereas cryptic junction read counts are squared. Note that spliced reads mapping to the cryptic junction are found in *eIF4AIII*, *mago* and *tsu* but not the control comparison.
(B) Schematic of a de novo splicing event detected on the *CG31156* transcript.
(C) Conservation of the cryptic 5' SS (NNSPLICE score of 0.54) and a potential 3' recursive splice site (NNSPLICE score of 0.98) found on the *CG31156* transcript highlighted in green, relative to the gene model. Conservation of the splice site is depicted using the multiple alignment format on the UCSC genome browser, as well as phyloP scores. Canonical splice sites are highlighted in yellow.
(D) Model of activation of dual cryptic splice sites on the *CG31156* transcript. Activation of the cryptic 5' SS with an additional cryptic 3' recursive splice sites leads to deletion of 110 nt of mRNA.

wobble patterns (**Figure 4.8A**). Thus, these examples of cryptic exonic recursive splicing are functional, but evolutionarily fortuitous.

The EJC protects spliced mRNAs from resplicing

Since many genes span large genomic regions, cDNA constructs have been a mainstay of directed expression strategies. It is generally expected that these should be effective at inducing gain-of-function conditions, yet cDNA constructs are not typically vetted for proper processing. My finding of dual cryptic splice sites on transcripts was alarming because in both cases, I observed resplicing on mRNA constructs (**Figure 4.8C and Figure 4.8F**). To reiterate, the EJC prevents dual cryptic SS from resplicing on transcript segments that have already undergone intron removal, but such protection will be missing from intronless cDNA copies.

I was keen to assess the breadth of this concept. To do so, I examined the sequence of mRNAs bearing EJC-suppressed cryptic SS, and looked for additional unidentified, complementary SS. Notably, since resplicing would have to map to a canonical junction, I looked for regenerated SS at exon junction sequences. An initial survey for SS invariant dinucleotide signatures (AG for 3' SS and GT for 5' SS) indicated that 64/118 junctions with cryptic 3' SS and 104/183 junctions with cryptic 5' SS were compatible with resplicing. The fact that over half of both classes of cryptic splicing events were potentially compatible with resplicing might at first glance seem like a tremendous enrichment. However, it does in fact reflect fundamental features of extended consensus splice sequences that basepair with the spliceosome, namely the U1 snRNP and U2AF35 binding sites, respectively (**Figure 4.8G**-top). Quantification of these sequences indicated a range of regenerated 5' and 3' SS at exon junctions, with at least 59 junctions resembling strong SS (**Figure 4.8H**, NNSPLICE>0.75). However, as several cryptic 5' and 3' SS amongst my validated loci (**Figures 4.1- Figure 4.8**) were

extremely poor, with functional dual cryptic splice sites in *CkIIβ* scoring at only 0.13 and 0.26 (**Figure 4.8A**), the functional breadth of this phenomenon is undoubtedly broader. Therefore, I imagined a scenario where a core function of the EJC is to repress splice sites that were regenerated at exon junctions as a consequence of intron removal using canonical splice sites (**Figure 4.8G**-bottom).

Nevertheless, as this model cannot be explicitly distinguished from alternative splicing without experimental testing, I selected additional loci for analysis. Therefore, I constructed partial cDNA constructs for three genes, encompassing regions I had validated as subject to EJC-suppression of cryptic splicing (**Figure 4.2C**), and selected targets that survey a range of regenerated SS strengths. These include *straw*, which yields a strong 3' RSS (NNSPLICE score of 0.98) after removal of intron 3; *eIF4G1*, which regenerates a moderate 5' RSS (NNSPLICE score of 0.64) after processing of intron 10; and *baboon,* which produces an exceptionally poor 3' RSS (NNSPLICE score of 0) after removal of intron 4, bearing only the AG dinucleotide.

In contrast to the endogenous genes which produced a single amplicon, expression of all three cDNA constructs yielded substantial re-spliced products, supporting my view that the EJC prevents activation of dual cryptic SS on mRNAs, including cryptic SS at exon junction sequences (**Figure 4.8I**). Unexpectedly, SS strength did not correlate with levels of re-splicing. Indeed, the majority of transcripts from all three reporters were truncated, including from *baboon*. Furthermore, the *eIF4G1* reporter yielded three truncated products, suggesting that other sequences may also serve as cryptic SS. As these examples of re-splicing occur on coding regions of the transcript, all of them either delete amino acids or generate frameshifts. I conclude that many cDNA constructs are potentially prone to resplicing due to loss of protection afforded by the EJC.

Moreover, I propose there are distinct classes of cryptic SS within exon junction sequences. The first, examples of which were documented previously, and extended in this study, comprise strong, autonomous splice donors that occur at the 5' ends of exons and are involved in recursive splicing (Blazquez et al., 2018; Boehm et al., 2018; Burnette et al., 2005; Duff et al., 2015; Hatton et al., 1998; Pai et al., 2018). The second class, which I discover in this study, includes the auxiliary, exonic remnants of canonical splice sites subsequent to intron removal. Crucially, these are weak and are not expected to function as autonomous splice sites, but they can nevertheless become substantially activated under EJC loss-of-function conditions.

Discussion

Conserved role for the EJC to repress cryptic splicing and its regulatory implications

Although introns are not essential for gene expression, they play important facilitatory roles by enhancing export and translation in part through recruitment of the EJC during splicing. Subsequently, it was recognized that once deposited, the EJC also promotes accurate gene expression by regulating processing of neighboring introns. Recently, in the mammalian setting, the role of the EJC during pre-mRNA splicing was extended to include suppression of cryptic splice sites (Blazquez et al., 2018; Boehm et al., 2018).

Here I reveal that the fly EJC similarly plays a broad role in direct suppression of cryptic exonic splice sites, owing to its characteristic deposition 20-24 nt upstream of exon-exon junctions. Thus, I highlight that concealment and suppression of cryptic splice sites is a conserved EJC activity (Boehm et al., 2018). Importantly, the positional recruitment of the EJC during splicing is conserved and sequence-independent (H Le Hir et al., 2000). Thus, I infer this function should also be independent of splice site divergence between phyla, as well as splice site strength, and should not require

accessory components. In contrast, non-conserved roles of the EJC appear to rely on integration within and diversification of distinct functional networks. For example, while the Upf (*Up-f*rameshift) proteins coordinate NMD across eukaryotes (He & Jacobson, 2015), the mechanisms differ. In mammals, NMD is coordinated with intron removal through direct interactions between the EJC and Upf3 (V. N. Kim et al., 2001; H Le Hir et al., 2001; Lykke-Andersen et al., 2001). However, these interactions are not found in invertebrates, and consequently the invertebrate EJC is not involved in NMD (Nicholson & Mühlemann, 2010).

In addition to pre-mRNAs, I show that the EJC also suppresses cryptic splice sites within spliced mRNAs. Although this mechanism cannot be distinguished from alternative splicing (**Figure 4.8A**) without further experimentation, I readily detect re-splicing on all cDNA constructs tested. Unexpectedly, while these junctions appear to contain just one cryptic SS, my data indicates that these transcripts contain secondary cryptic splice sites that mediate resplicing. Importantly, I validate that even poor matches to SS consensus motifs are competent for re-splicing. Curiously, as all of my demonstrated examples involve a recursive event at either the 5' or 3' cryptic SS, my findings broaden a phenomenon that was previously described within long introns (Duff et al., 2015; Joseph et al., 2018; Pai et al., 2018). Furthermore, canonical SS sequences that undergo base pairing interactions with U1 snRNA (5' SS) and U2AF35 (3' SS) have motifs AG|GURAGU and YAG|GU (Kielkopf et al., 2001; Kondo et al., 2015). It is noteworthy that core splice site signals contain bases that are compatible with regeneration of splice sites and that these naturally occur proximal to EJC recruitment sites. Accordingly, I propose that an ancestral function for the conserved position of EJC deposition may be to prevent accidental activation of regenerated splice sites.

Finally, my observations of re-splicing on cDNAs reflect an essential function for introns in protecting mRNA fidelity. For all tested cases of cDNA resplicing on coding

sequences, I note deletions of peptide segments or truncations with loss of domains required for protein function (**Figure 4.10A-C**). Importantly, these affected targets include essential genes, such as *eIF4G1* and activin receptor *baboon.* In the case of *baboon,* the 54 nt splicing defect leads to a deletion of 18 amino acids (195-212, **Figure 4.10A**). For *eIF4G1,* re-splicing removes 131 nt of mRNA sequence, alters the open reading frame and leads to protein truncation with loss of the MI and W2 domains (**Figure 4.10B**). Finally, re-splicing on *straw* transcripts also alters reading frame by removing 91 nt of mRNA, and is predicted to remove 2/3 Plastocyanin-like domains (**Figure 4.10C**). Thus, my findings have serious implications for functional genomics as well as community genetic studies (Wei et al., 2020; Yu et al., 2011), where cDNA expression constructs and collections are often employed with little attention paid to mRNA processing. Altogether, my work uncovers an important co-transcriptional function of intron removal and the role of the EJC to protect the transcriptome from unwanted re-splicing.

Materials and Methods

Bioinformatic analysis

The core-EJC knockdown RNA-sequencing datasets were previously reported (Akhtar et al., 2019) and obtained from the NCBI Gene Expression Omnibus (GSE92389). Raw sequencing data was mapped to the *Drosophila* reference genome sequence (BDGP Release 6/dm6) using HISAT2 (D. Kim et al., 2015) under the default settings. Splice junctions were mapped using the MAJIQ algorithm (2.0) under default conditions (Vaquero-Garcia et al., 2016). Splice graphs and known/novel local splice variants were defined with the MAJIQ Builder using annotations of known genes and splice junctions from Ensembl release 95 and all BAM files. The MAJIQ Quantifier was

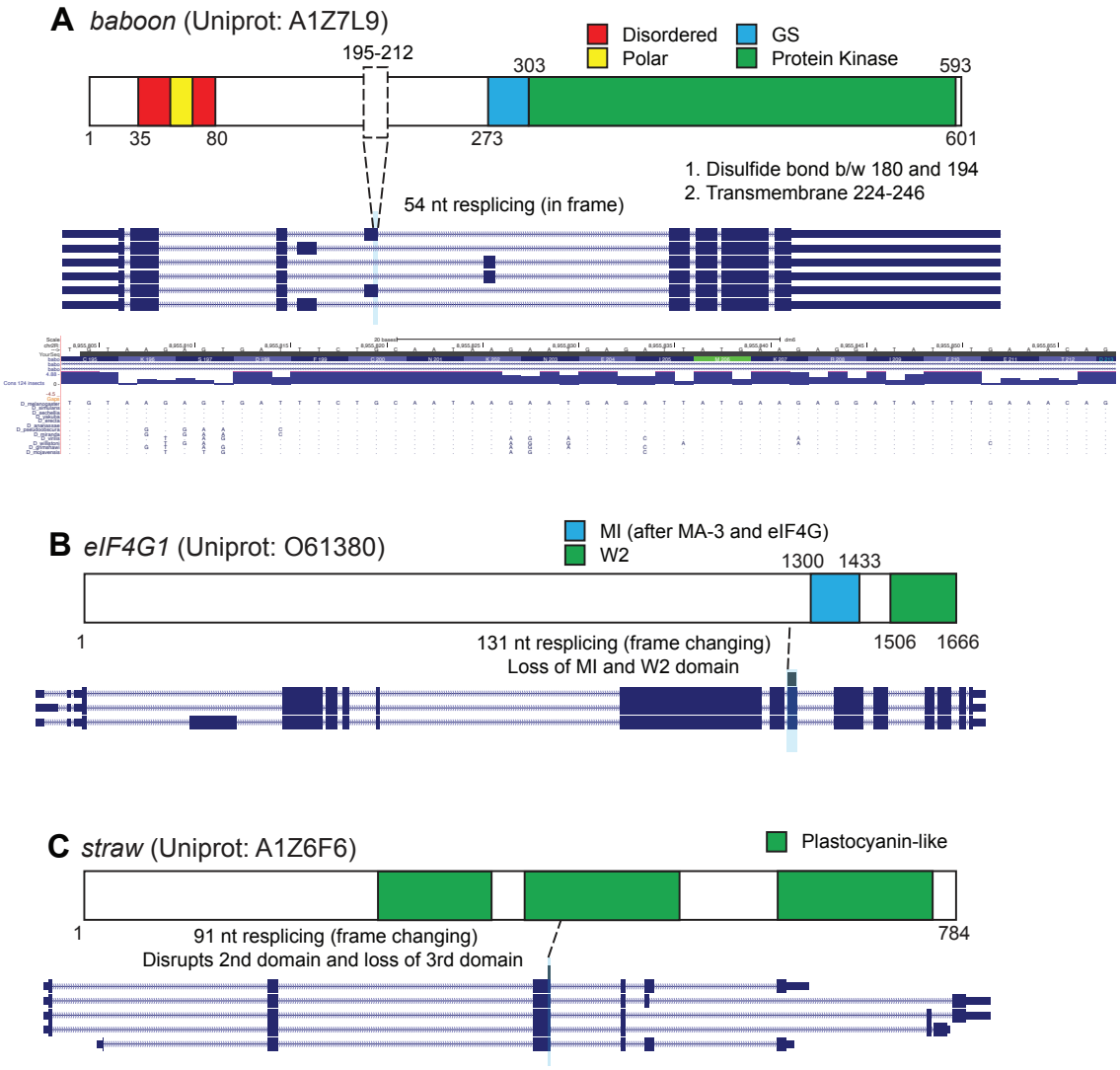


Figure 4.10. Re-splicing on mRNAs alters translated proteins
(A-C) Protein and transcript structures are schematized and the location of cryptic resplicing highlighted in blue.
(A) Re-splicing on *baboon* leads to a 54 nt deletion of the mRNA and an 18 amino acid deletion. The deletion does not overlap known domains. Conservation plots for deleted 54 nt region is included.
(B) Re-splicing on *eIF4G1* leads to a 131 nt deletion, leading to a change in reading frame and truncation of the C terminal domains of eIF4G1. Importantly, critical domains required for eIF4G1 function are lost due to re-splicing.
(C) Re-splicing on *straw* leads to a 91 nt deletion, leading to a change in reading frame and truncation of the protein. Importantly, 2 of 3 Plastocyanin-like domains are lost due to transcript defects.

used to calculate relative abundances (percent selected index - PSI) for all defined junctions. The resulting data was output into tabular format using the Voila function.

A custom R script was written to process all MAJIQ-defined novel junctions relative to the Ensembl gene annotations and identify *de novo* EJC-suppressed junctions. First, I quantified usage of all novel junctions by mining mapped libraries (BAM files) for high quality junction spanning reads with at least 8 nt of overhang and no mismatches. These counts were normalized to sequencing depth per library. To identify *de novo* junctions that may be upregulated, I first selected junctions with at least 5 split reads. In order to enrich for *de novo* junctions that are suppressed by the EJC pathway, I looked for those with > 2 fold difference in at least 2/3 core-EJC RNAi conditions relative to the *lacZ* control. To apply further stringency, I also required that the PSI measurements reflect sufficient change between treatment and control conditions. Therefore, I applied an additional filter of PSI fold change > 2 in at least 2/3 core-EJC RNAi conditions. These criteria produced a total of 573 novel junctions.

The 5' and 3' ends of these junctions were compared against known gene annotations to characterize splice sites. Exonic 5' and 3' SS reflect sites that mapped on exons while the other end mapped to a canonical splice junction, and the same process was used to define intronic 5' and 3' SS. *de novo* cases of alternative splicing reflect junctions that utilized annotated splice sites but represented novel connectivity. Sashimi plots were generated using features available on the Integrative Genomics Viewer (IGV) (Robinson et al., 2011).

I generated a custom pipeline to assess recursive splicing potential (Figure 4.8). Briefly, I identified transcripts that contained cryptic exonic 5' and 3' splice sites. For these transcripts, I mapped the position of all splice junctions on the mRNA, which could in theory generate the observed splicing defects. I examined sequences directly

downstream of relevant splice junctions to identify potential 5' recursive splicing and those directly upstream to identify potential 3' recursive splicing.

I calculated splice site strengths using NNSPLICE (https://www.fruitfly.org/seq_tools/splice.html) (Reese et al., 1997). The sequences used for these analyses were obtained from mRNA rather than the genomic context, which may contain intronic sequences as well. To generate nucleotide content plots, splice sites and their indicated flanking sequences were obtained from mRNAs and fed to WebLogo version 2.8.2 (Crooks et al., 2004). The splice sites are centered in these plots.

All custom scripts used in this study are reported on the Lai lab GitHub page.

Constructs and cell culture

All splicing reporters were cloned into pAC-5.1-V5-His (ThermoFisher Scientific) using compatible restriction sites. I used PCR to amplify minigene splicing reporters from Drosophila genomic DNA, and used site directed mutagenesis to remove specified introns. I used cDNAs to amplify reporters lacking introns. For genes with multiple isoforms (such as *CG7408*), I cloned the dominant fragment. All primers used for generating constructs and mutagenesis have been summarized in **Table 4.1**.

Transfections were performed using S2-R+ cells cultured in Schneider Drosophila medium with 10% FBS. Cells were seeded in 6-well plates at a density of $1x10^6$ cells/mL and transfected with 200 ng of plasmid using the Effectene transfection kit (Qiagen). Cells were harvested following 3 days of incubation.

Knockdown of EJC factors in S2 cells

The indicated EJC components were knocked down via RNAi (dsRNA-mediated interference) in S2-R+ cells. The MEGAscript™ RNAi kit (ThermoFisher Scientific) was used to produce dsRNAs required for this experiment. Briefly, DNA templates containing promoter sequences on either 5' end were produced through PCR with T7-promoter-fused primers. 2 µg of DNA template was transcribed *in vitro* for 4 hours as recommended by the manufacturer. The products were incubated at 75º C for 5 minutes and brought to room temperature to enhance dsRNA formation. A cocktail of DNAseI and RNAse removed DNA and ssRNAs, and the remaining dsRNA was purified using the provided reagents. All dsRNA reagents were verified by running on a 1% agarose gel and quantified by measuring absorbance at 260 nm using a NanoDrop™ (ThermoFisher Scientific).

For knockdown, $3x10^6$ S2-R+ cells in 1 mL serum free medium were incubated with 15 µg of dsRNA for 1 hour at room temperature. Then, 1 mL of medium containing 20% FBS was added to the cells and the whole mixture was moved to a 6 well plate. Cells were collected after 4 days of incubation.

RT-PCR

After transfection or RNAi treatment, cells were washed in ice cold PBS and pelleted using centrifugation. RNA was collected using the TRIzol reagent (Invitrogen) under the recommended conditions. 5 µg of RNA was treated with Turbo DNase (Ambion) for 45 min before cDNA synthesis using SuperScript III (Life Technology) with random hexamers. RT-PCR was performed using AccuPrime Pfx DNA polymerase (ThermoFisher Scientific) with standard protocol using 26 cycles and primers that were specific to each minigene construct. All primers are listed and described in **Table 4.1**.

Table 4.1 List of Primers (Page 1 of 2)

cloning primers	sequence
cln_cg7408_fwd	ACTGTACTGAATTCTGCTCTCCGCACTTCGAGTC
cln_cg7408_rev	ACTGTACTGCGGCCGCCTGGGTGTGCACATTGGAGC
cln_CG3632_f	ACTGTACTGAATTCGCACACCAATTTCCGTCGTC
cln_CG3632_r	ACTGTACTGCGGCCGCCTTCGTCCAAGCCAGAACTCAG
cln_CkIIbeta_f	ACTGTACTGAATTCGCAGCAAAATGAGCAGCTCC
cln_CkIIbeta_r	ACTGTACTGCGGCCGCCAATGGCAGCATGGGCTGAC
cln_CG31156_f	ACTGTACTGAATTCGACCAAACGCCAGCGGTTC
cln_CG31156_r	ACTGTACTGCGGCCGCCAGACATCACAAGTGCCTCCG
cln_laccase2_fwd	ACTGTACTGAATTCCCAGTTTCGTGACCCGAAC
cln_laccase2_rev	ACTGTACTGCGGCCGCCGGTTACATTGGGCGTCC
cln_eif4g_cds_fwd	ACTGTACTGAATTCGCTCTCTAGCGGTTGACAGC
cln_eif4g_cds_rev	ACTGTACTGCGGCCGCACCGTAGTTTCTCTGGTACTCG
cln_babo_fwd	ACTGTACTGGTACCGCAACGGAGTAAGCCCTTCG
cln_babo_rev	ACTGTACTGCGGCCGCCGAACCAGAGGTGGTCATCTC

removing introns via SDM

primers	sequence
CG3632_sdm_F	TACGTGAGGAaATGCTAC
CG3632_sdm_R	gCTTGTGTTCTTGTACGATTTG
CG31156_sdm_F	TACCGTCGAGaTCTGGTG
CG31156_sdm_R	cCTGCAAATAAAAGGTATGGTATTTTC
Ckbeta_sdm2_F	GGACGAGCTCgAGGACAA
Ckbeta_sdm2_R	tCCGGTTCCAAGTCCAAGATC
Ckbeta_sdm1_F	TGGATGAGGAcTACATAC
Ckbeta_sdm1_R	cCTCGCAGAAGAACTCATTG
CG7408_int1delF	CTAGATGTTATTTAGTTGATAAGTTAG
CG7408_int1delR	CGAAATCAAATAAAAGGAATCC
CG7408_int2delF	CTGTCCCACATCATCCTC
CG7408_int2delR	CAAATGGGCGACAACAAG
CG7408_int3delF	GGTTTTGACGACGTTAGC
CG7408_int3delR	CAGATCATCTGCCATAATAATG
CG3632_mut_donorF	AAATACACGGcgaGCGACTGGCTC
CG3632_mut_donorR	CTCGCAGTAGCATTTCCTC
CG7408_3xFlagF	ccacgacatcgactacaaggacgacgacgacaagTGACTTGTTGTCGCCCATTTG
CG7408_3xFlagR	tccttgtagtcaccgtcgtggtccttgtagtccatGCGGGTAACACTCTGTCAATG

Table 4.1 List of Primers (Page 2 of 2)

RNAi template primers	sequence
tsu_RNAiF	TTAATACGACTCACTATAGGGGAGACGATGTGTTGGACATTGACA
tsu_RNAiR	TTAATACGACTCACTATAGGGGAGAGACGCTTTTCGGACTTTTT
mago_RNAiF	TTAATACGACTCACTATAGGGGAGACACGGAGGACTTTTACCTAC
mago_RNAiR	TTAATACGACTCACTATAGGGGAGAATATGGGCTTGATCTTGAAATG
eIF4AIII_RNAiF	TTAATACGACTCACTATAGGGGAGAGACGAATTGACACTGGAAGG
eIF4AIII_RNAiR	TTAATACGACTCACTATAGGGGAGAAGAATATTAGTTTAGATCAAGTCAG
btz_RNAiF	TTAATACGACTCACTATAGGGGAGACCGAAGTGGAGAAACCAACG
btz_RNAiR	TTAATACGACTCACTATAGGGGAGAGATGCCTGTGAGATCTGTGG

rtPCR primer	sequence
cg7408_fwd	TGCTCTCCGCACTTCGAGTC
cg7408_rev	CTGGGTGTGCACATTGGAGC
CG3632_f	GCACACCAATTTCCGTCGTC
CG3632_r	CTTCGTCCAAGCCAGAACTCAG
CkIIbeta_f	GCAGCAAAATGAGCAGCTCC
CkIIbeta_r	CAATGGCAGCATGGGCTGAC
CG31156_f	GACCAAACGCCAGCGGTTC
CG31156_r	CAGACATCACAAGTGCCTCCG
laccase2_fwd	CCAGTTTCGTGACCCGAAC
laccase2_rev	CGGTTACATTGGGCGTCC
eif4g_cds_fwd	GCTCTCTAGCGGTTGACAGC
eif4g_cds_rev	ACCGTAGTTTCTCTGGTACTCG
babo_fwd	GCAACGGAGTAAGCCCTTCG
babo_rev	CGAACCAGAGGTGGTCATCTC
Haspin_e1_fwd	GGAAGGTAGATGGAAGGATCCG
Haspin_e5_rvs	CCTGTGAACTTTCGTATTGATGC
mask_e9_fwd	CGACAGCACTGGACAATAGC
mask_e11_rvs	ACATGCCACGGAAATCGTCC
Crk_e1_fwd	CGTTTCTGATAGGAACAGCTGG
Crk_e3_rvs	AGTCCACCATTGATCCTCGTC
eif4g1_5utr_fwd	TGAACAGAACACATTGCATGTGG
eif4g1_5utr_rvs	CTCTGTAGGAAATCGCCAAACG

Chapter 5

Conclusions and Perspectives

Reconception of zero-nucleotide exons as short cryptic exons and implications for noncanonical splicing

Precision RNA processing pathways demonstrate at least two important attributes. First, the enzymatic machinery has to identify the correct substrates, and second, it needs to execute catalytic activity with nucleotide precision. These features are clearly evident for the splicing reaction as mRNA from split genes display exact intron removal and loss of fidelity can typically lead to deleterious consequences. But while there is a great deal of scholarship on the assembly and activity of the spliceosome, early steps regarding splice site definition remain mysterious. This is particularly apparent for the processing of genes with long introns. Longer introns can contain many instances of SS consensus sequences, just by mere chance, and it is remarkable indeed, that splicing can occur with high accuracy at such loci.

A particular solution to this conundrum came from through the appreciation of how the spliceosome interprets gene and intron/exon architecture. Briefly, the data suggests that the spliceosome operationally defines short blocks of sequences rather than individual elements. Thus, for genes with long introns, exons tend to be smaller (50-250 nt), hence, the splice sites on either side of an exon stimulate the recognition of each other (Berget, 1995; De Conti et al., 2013).

The initial discovery of an intronic recursive splice site in the *Ubx* locus and the detection of partially processed *Ubx* pre-mRNA (J M Burnette et al., 2005) suggested that an exon of zero-nucleotide length could also function as a “short definable block” during exon definition. However, the concept of a zero-nucleotide exon can easily be refuted on the mechanistic basis that simultaneous 5’SS and 3’SS definition (as required

for exon definition) cannot occur due to the steric challenges of U1 snRNP and the U2AF binding the same splice sequences simultaneously. The steric interference argument is supported by observations that reducing the length of an internal exon below 50 nt results in exon skipping (Dominski & Kole, 1992). Orthogonally, Douglas Black has also reported that extending a short cassette exon from 18 nt to 109 nt produces efficient exon inclusion in mRNA (Black, 1991). While these results are undoubtedly confounded by the effects of SREs, it is evident that separation of splice sites is preferable to overlap.

So if not zero-nucleotide exon definition, how are intronic recursive splice sites in *Drosophila* recognized? Attempts to identify additional regulatory elements based on phylogenetics did not prove useful, as there were no local peaks in sequence conservation besides the RSS (Duff et al., 2015). In this document, using complementary approaches, I demonstrate that i. recursive splice sites are consistently paired with downstream 5'SS, and ii. downstream 5'SS and recursive 3'SS are required for cryptic RS-exon definition (**Figure 2.1C**). The proposed mechanism breaks recursive splicing into two phases. During the first, the 3' recursive SS and the downstream 5'SS participate in exon definition. This causes activation of the 3' recursive SS, removal of the upstream intron fragment and regenerates the 5' recursive SS. During the second phase, the 5' recursive SS is activated to remove the remaining intronic sequence. Therefore, the downstream RS-exon splice donor is a silent partner in this process (Joseph et al., 2018).

An intriguing takeaway from these experiments is the idea that exon definition is only essential for one of the two SS found on exons (for RS-exons, definition is only required for activation of the 3' recursive SS), and activation of the other SS occurs separately. This could explain the processing of two other classes of poorly understood, suboptimal intron/exon architectures: i. short microexons flanked by long introns and ii.

long exons (> 300 nt) flanked by long introns (Figure 5.1). 3-30 nt microexons have been detected in the transcriptomes of vertebrates and invertebrates with a tendency for inclusion in the nervous system (Ustianenko et al., 2017). Such small internal microexons (~ 3nt) that are flanked by long introns may encounter similar challenges as zero-nucleotide exons. Therefore, it is reasonable to consider that weak, cryptic downstream splice sites may also facilitate the first step of microexon definition.

Conversely, long exons flanked by long introns may encounter unique challenges as neither exon nor intron is optimal for early spliceosome assembly. In *Drosphila,* when flanked by introns > 10000 nt, exons tend to be ~50 nt in length, displaying preference for an optimal size that overlaps the average RS-exon size (**Figure 2.13E**). However, several members of this class also have long exons. *mbl* exon 2 (654 nt) is an archetype for this class, sandwiched by introns of lengths 9198 nt and 75050 nt. As both exon and intron lengths are suboptimal, it is worth exploring if cryptic splice donors can shorten long exons into an optimally sized exon for initial activation of the canonical 3'SS. Similar to RS, the canonical 5'SS should be able to outcompete the cryptic exonic splice donor during removal of the downstream intron. In support of such a model, chapter 4 provides examples of scores of exonic cryptic 5'SS that map close to the 3'SS of exons. While these were activated under EJC LOF and resulted in unwanted resplicing, it is possible that these cryptic SS may serve a useful purpose during other stages of pre-mRNA processing, such as exon definition. Mutation of cryptic splice donors for long exon/long intron instances will be a worthwhile test of this hypothesis.

Regulation of RS-exon inclusion

Alternative splicing allows production of distinct mRNA species from the same gene. This phenomenon provides fine control over gene output and vastly diversifies the proteome. Moreover, it is quite apparent that AS is dynamic and has important

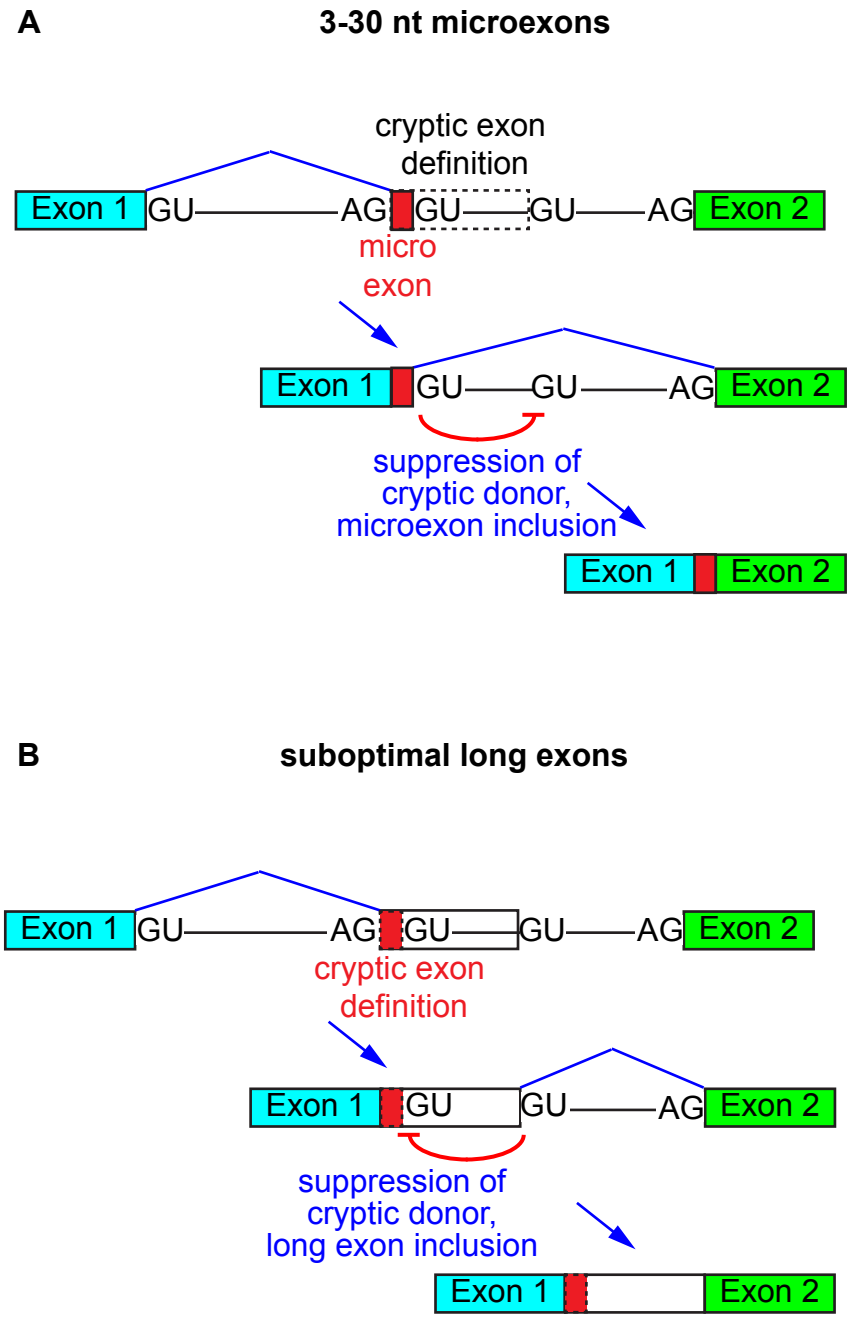


Figure 5.1. Cryptic 5'SS may assist during exon definition of other suboptimal substrates within long introns. (A) microexon and (B) long exons.

physiological functions in executing the biology of eukaryotic cells. Therefore, a lot of effort has gone into understanding mechanisms that regulate AS (Baralle & Giudice, 2017; Lee & Rio, 2015) and reflect useful paradigms to explore the regulation of RS-exon inclusion.

RS-exon splicing is often schematized as a cassette exon, but due to the intermediate pre-mRNA, it is useful to consider RS-exon regulation from the perspective of alternative 5'SS choice. In this context, I explored 5'SS strengths, presence of ESEs and EJC recruitment as factors that regulate RS-exon inclusion. My *in vivo* mutagenesis of the RP 5'SS provide a valuable *in vivo* resource to examine the 5'SS competition model. Moreover, the results indicate that 5'SS strength is critical for RS-exon skipping. Comparative genomics provides further support for this view, as RP 5'SS appear a more conserved element than cognate RS-exon 5'SS.

Using RS-exon swap reporters in cell culture tests, I document that RS-exons contain sufficient information – by themselves – to mirror host intron processing. Again, motivated by the comparative genomics of cryptic and expressed RS-exons, I argued that expressed exons contain ESEs that promote inclusion, whereas poorly conserved, cryptic RS-exons are unlikely to contain important *cis*-regulatory elements. Consistent with the former, several potential splicing factors have been identified (James M. Burnette et al., 1999), and the future use of a massively parallel reporter assay (Litterman et al., 2019) may prove insightful. In this context, it is worth revisiting what is known about regulation through *cis/trans* factors. The current view is that RBPs cooperatively bind conserved, multivalent binding sites within or proximal to AS exons (Ule & Blencowe, 2019). While my reporters indicate the presence of SREs, there are obvious space constraints within these short RS-exons. The modes in which information is contained and communicated will be an exciting direction for this project.

The nebulous function of RPs during pre-mRNA processing

The question of biological significance is critical, yet unanswered in the recursive splicing field. However, with the invention of CRISPR/Cas9 genome engineering tools, manipulating short sequences within long introns with high precision is now an achievable task. Consequently, I report the generation of 9 RP mutants from 5 genes. Deletion of intronic RPs had no observable consequences on mRNA production via rt-PCR assays. This suggests RPs are dispensable for intron processing, but there is plenty to examine, still. A careful study of nascent RNA splicing and other cotranscriptional RNA processing will provide further clarity on the function of these sites. Similarly, it is possible that I have not looked in the right place or setting, so exploring distinct cell types, under different environmental contexts is a prudent future direction. While this first effort at deleting RPs did not yield profound defects, it is worth noting that attempts to understand the function of deeply conserved cassette exons in mammalian cells also found that few, but not all exons had discernable cellular requirements (Thomas et al., 2020). Therefore, another reasonable direction is to continue to make more RP deletions.

Exon junction sequences as deleterious cryptic splice sites

The Ule and Gehring labs recently identified regenerated 5'SS on constitutive exon, broadly expanding the scope of recursive splicing (Blazquez et al., 2018; Boehm et al., 2018). However, it is important to note that these regenerated 5'SS match consensus splice motifs. In contrast, I demonstrate that sequences that do not match splice motifs can also get activated under loss of the EJC. Critically, the data suggests these sequences typically occur at exon junction sequences. Exon junction sequences, in theory, contain the remnant of canonical splice sites after intron removal, but these elements have never been shown to have independent activity of their own. Remarkably,

I find that exon junction sequences can be reactivated as both 5' and 3'SS and the EJC is required for suppression of these during pre-mRNA splicing. Importantly, as EJC recruitment proximal to exon junction sequences is a conserved property, I argue that this is an ancient function for the EJC.

Implicit in this discussion is the notion that not all exon junction sequences possess activatable SS. If this were true, all cDNA expression constructs from multiexon genes would display unexpected resplicing in cell culture. As this is not the case, it appears that a select class of exon junction sequences possess this capacity. Elucidating factors that enhance this attribute will be a significant step towards appreciating the roles of cryptic splicing during pre-mRNA processing.

Overall, I study the landscape, mechanism and function of recursive splicing to discover that cryptic splice sites can be contextually essential or detrimental to accurate gene expression.

Bibliography

Achsel, T., Brahms, H., Kastner, B., Bachi, A., Wilm, M., & Lührmann, R. (1999). A doughnut-shaped heteromer of human Sm-like proteins binds to the 3'-end of U6 snRNA, thereby facilitating U4/U6 duplex formation in vitro. *The EMBO Journal*, *18*(20), 5789–5802. https://doi.org/10.1093/emboj/18.20.5789

Akhtar, J., Kreim, N., Marini, F., Mohana, G., Brüne, D., Binder, H., & Roignant, J.-Y. (2019). Promoter-proximal pausing mediated by the exon junction complex regulates splicing. *Nature Communications*, *10*(1), 521. https://doi.org/10.1038/s41467-019-08381-0

Alt, F. W., Bothwell, A. L., Knapp, M., Siden, E., Mather, E., Koshland, M., & Baltimore, D. (1980). Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. *Cell*, *20*(2), 293–301. https://doi.org/10.1016/0092-8674(80)90615-7

Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., & Komorowski, J. (2009). Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Research*, *19*(10), 1732–1741. https://doi.org/10.1101/gr.092353.109

Anna, A., & Monika, G. (2018). Splicing mutations in human genetic disorders: examples, detection, and confirmation. *Journal of Applied Genetics*, *59*(3), 253–268. https://doi.org/10.1007/s13353-018-0444-7

Antoine Cléry and Frédéric H.-T. Allain. (2011). FROM STRUCTURE TO FUNCTION OF RNA BINDING DOMAINS. In *Madame Curie Bioscience Database [Internet]*. https://www.ncbi.nlm.nih.gov/books/NBK63528/?report=printable

Arnold, E. S., Ling, S.-C., Huelga, S. C., Lagier-Tourenne, C., Polymenidou, M., Ditsworth, D., Kordasiewicz, H. B., McAlonis-Downes, M., Platoshyn, O., Parone, P. A., Da Cruz, S., Clutario, K. M., Swing, D., Tessarollo, L., Marsala, M., Shaw, C. E., Yeo, G. W., & Cleveland, D. W. (2013). ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(8), E736-45. https://doi.org/10.1073/pnas.1222809110

Artero, R. D., Akam, M., & Pérez-Alonso, M. (1992). Oligonucleotide probes detect splicing variants in situ in Drosophila embryos. *Nucleic Acids Research*, *20*(21), 5687–5690. https://doi.org/10.1093/nar/20.21.5687

Ashton-Beaucage, D, Udell, C. M., Lavoie, H., Baril, C., Lefrancois, M., Chagnon, P., Gendron, P., Caron-Lizotte, O., Bonneil, E., Thibault, P., & Therrien, M. (2010). The exon junction complex controls the splicing of MAPK and other long intron-containing transcripts in Drosophila. *Cell*, *143*(2), 251–262. https://doi.org/10.1016/j.cell.2010.09.014

Ashton-Beaucage, Dariel, Udell, C. M., Lavoie, H., Baril, C., Lefrançois, M., Chagnon, P., Gendron, P., Caron-Lizotte, O., Bonneil, E., Thibault, P., & Therrien, M. (2010). The exon junction complex controls the splicing of MAPK and other long intron-containing transcripts in Drosophila. *Cell*, *143*(2), 251–262. https://doi.org/10.1016/j.cell.2010.09.014

Attig, J., Agostini, F., Gooding, C., Chakrabarti, A. M., Singh, A., Haberman, N., Zagalak, J. A., Emmett, W., Smith, C. W. J., Luscombe, N. M., & Ule, J. (2018). Heteromeric RNP Assembly at LINEs Controls Lineage-Specific RNA Processing. *Cell*, *174*(5), 1067-1081.e17. https://doi.org/10.1016/j.cell.2018.07.001

Baralle, F. E., & Giudice, J. (2017a). Alternative splicing as a regulator of development and tissue identity. *Nature Reviews. Molecular Cell Biology*, *18*(7), 437–451.

https://doi.org/10.1038/nrm.2017.27
Baralle, F. E., & Giudice, J. (2017b). Alternative splicing as a regulator of development and tissue identity. *Nature Reviews. Molecular Cell Biology*, *18*(7), 437–451. https://doi.org/10.1038/nrm.2017.27
Barbosa, I., Haque, N., Fiorini, F., Barrandon, C., Tomasetto, C., Blanchette, M., & Le Hir, H. (2012). Human CWC22 escorts the helicase eIF4AIII to spliceosomes and promotes exon junction complex assembly. *Nature Structural & Molecular Biology*, *19*(10), 983–990. https://doi.org/10.1038/nsmb.2380
Bender, W., Akam, M., Karch, F., Beachy, P. A., Peifer, M., Spierer, P., Lewis, E. B., & Hogness, D. S. (1983). Molecular Genetics of the Bithorax Complex in Drosophila melanogaster. *Science (New York, N.Y.)*, *221*(4605), 23–29. https://doi.org/10.1126/science.221.4605.23
Berg, M. G., Singh, L. N., Younis, I., Liu, Q., Pinto, A. M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L., & Dreyfuss, G. (2012). U1 snRNP determines mRNA length and regulates isoform expression. *Cell*, *150*(1), 53–64. https://doi.org/10.1016/j.cell.2012.05.029
Berget, S. M. (1995). Exon recognition in vertebrate splicing. *The Journal of Biological Chemistry*, *270*(6), 2411–2414.
Berget, S. M., Moore, C., & Sharp, P. A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(8), 3171–3175. https://doi.org/10.1073/pnas.74.8.3171
Berglund, J. A., Abovich, N., & Rosbash, M. (1998). A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. *Genes & Development*, *12*(6), 858–867. https://doi.org/10.1101/gad.12.6.858
Berk, A. J. (2016). Discovery of RNA splicing and genes in pieces. *Proceedings of the National Academy of Sciences*, *113*(4), 801–805. https://doi.org/10.1073/pnas.1525084113
Bertram, K., Agafonov, D. E., Liu, W.-T., Dybkov, O., Will, C. L., Hartmuth, K., Urlaub, H., Kastner, B., Stark, H., & Lührmann, R. (2017). Cryo-EM structure of a human spliceosome activated for step 2 of splicing. *Nature*, *542*(7641), 318–323. https://doi.org/10.1038/nature21079
Bischof, J., Maeda, R. K., Hediger, M., Karch, F., & Basler, K. (2007). An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(9), 3312–3317. https://doi.org/10.1073/pnas.0611511104
Black, D. L. (1991). Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non-neuronal cells? *Genes and Development*, *5*(3), 389–402. https://doi.org/10.1101/gad.5.3.389
Blazquez, L., Emmett, W., Faraway, R., Pineda, J. M. B., Bajew, S., Gohr, A., Haberman, N., Sibley, C. R., Bradley, R. K., Irimia, M., & Ule, J. (2018). Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing. *Molecular Cell*, *72*(3), 496-509.e9. https://doi.org/S1097-2765(18)30832-3 [pii]
Boehm, V., Britto-Borges, T., Steckelberg, A.-L., Singh, K. K., Gerbracht, J. V, Gueney, E., Blazquez, L., Altmüller, J., Dieterich, C., & Gehring, N. H. (2018). Exon Junction Complexes Suppress Spurious Splice Sites to Safeguard Transcriptome Integrity. *Molecular Cell*, *72*(3), 482-495.e7. https://doi.org/10.1016/j.molcel.2018.08.030
Boehm, V., & Gehring, N. H. (2016). Exon Junction Complexes: Supervising the Gene Expression Assembly Line. *Trends in Genetics : TIG*, *32*(11), 724–735. https://doi.org/10.1016/j.tig.2016.09.003
Bomze, H. M., & López, A. J. (1994). Evolutionary conservation of the structure and

expression of alternatively spliced Ultrabithorax isoforms from Drosophila. *Genetics*, *136*(3), 965–977. http://www.ncbi.nlm.nih.gov/pubmed/7911773
Bonnal, S. C., López-Oreja, I., & Valcárcel, J. (2020). Roles and mechanisms of alternative splicing in cancer - implications for care. *Nature Reviews. Clinical Oncology*, *17*(8), 457–474. https://doi.org/10.1038/s41571-020-0350-x
Bradley, T., Cook, M. E., & Blanchette, M. (2015). SR proteins control a complex network of RNA-processing events. *RNA (New York, N.Y.)*, *21*(1), 75–92. https://doi.org/10.1261/rna.043893.113
BRENNER, S., JACOB, F., & MESELSON, M. (1961). An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, *190*, 576–581. https://doi.org/10.1038/190576a0
Bringmann, P., & Lührmann, R. (1986). Purification of the individual snRNPs U1, U2, U5 and U4/U6 from HeLa cells and characterization of their protein constituents. *The EMBO Journal*, *5*(13), 3509–3516. http://www.ncbi.nlm.nih.gov/pubmed/2951249
Brody, E., & Abelson, J. (1985). The “spliceosome”: yeast pre-messenger RNA associates with a 40S complex in a splicing-dependent reaction. *Science (New York, N.Y.)*, *228*(4702), 963–967. https://doi.org/10.1126/science.3890181
Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., Booth, B. W., Wen, J., Park, S., Suzuki, A. M., Wan, K. H., Yu, C., Zhang, D., Carlson, J. W., Cherbas, L., Eads, B. D., Miller, D., Mockaitis, K., Roberts, J., … Celniker, S. E. (2014). Diversity and dynamics of the Drosophila transcriptome. *Nature*, *512*(7515), 393–399. https://doi.org/10.1038/nature12962
Burnette, J M, Miyamoto-Sato, E., Schaub, M. A., Conklin, J., & Lopez, A. J. (2005). Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. *Genetics*, *170*(2), 661–674. https://doi.org/genetics.104.039701 [pii]
Burnette, James M., Hatton, A. R., & Lopez, A. J. (1999). Trans-acting factors required for inclusion of regulated exons in the Ultrabithorax mRNAs of Drosophila melanogaster. *Genetics*, *151*(4), 1517–1529.
Busslinger, M., Moschonas, N., & Flavell, R. A. (1981). Beta + thalassemia: aberrant splicing results from a single point mutation in an intron. *Cell*, *27*(2 Pt 1), 289–298. https://doi.org/10.1016/0092-8674(81)90412-8
Busturia, A., Vernos, I., Macias, A., Casanova, J., & Morata, G. (1990). Different forms of Ultrabithorax proteins generated by alternative splicing are functionally equivalent. *The EMBO Journal*, *9*(11), 3551–3555. http://www.ncbi.nlm.nih.gov/pubmed/1976510
Chan, S.-P., Kao, D.-I., Tsai, W.-Y., & Cheng, S.-C. (2003). The Prp19p-associated complex in spliceosome activation. *Science (New York, N.Y.)*, *302*(5643), 279–282. https://doi.org/10.1126/science.1086602
Charenton, C., Wilkinson, M. E., & Nagai, K. (2019). Mechanism of 5’ splice site transfer for human spliceosome activation. *Science (New York, N.Y.)*, *364*(6438), 362–367. https://doi.org/10.1126/science.aax3289
Chen, Y.-C. A., Stuwe, E., Luo, Y., Ninova, M., Le Thomas, A., Rozhavskaya, E., Li, S., Vempati, S., Laver, J. D., Patel, D. J., Smibert, C. A., Lipshitz, H. D., Toth, K. F., & Aravin, A. A. (2016). Cutoff Suppresses RNA Polymerase II Termination to Ensure Expression of piRNA Precursors. *Molecular Cell*, *63*(1), 97–109. https://doi.org/10.1016/j.molcel.2016.05.010
Chow, L. T., Gelinas, R. E., Broker, T. R., & Roberts, R. J. (1977). An amazing sequence arrangement at the 5’ ends of adenovirus 2 messenger RNA. *Cell*, *12*(1), 1–8. https://doi.org/10.1016/0092-8674(77)90180-5
Cobb, M. (2015). Who discovered messenger RNA? *Current Biology*, *25*(13), R526–R532. https://doi.org/10.1016/j.cub.2015.05.032

Combs, D. J., Nagel, R. J., Ares, M., & Stevens, S. W. (2006). Prp43p is a DEAH-box spliceosome disassembly factor essential for ribosome biogenesis. *Molecular and Cellular Biology*, *26*(2), 523–534. https://doi.org/10.1128/MCB.26.2.523-534.2006

Company, M., Arenas, J., & Abelson, J. (1991). Requirement of the RNA helicase-like protein PRP22 for release of messenger RNA from spliceosomes. *Nature*, *349*(6309), 487–493. https://doi.org/10.1038/349487a0

Conklin, J. F., Goldman, A., & Lopez, A. J. (2005). Stabilization and analysis of intron lariats in vivo. *Methods*, *37*(4), 368–375. https://doi.org/10.1016/j.ymeth.2005.08.002

Cook-Andersen, H., & Wilkinson, M. F. (2015). Molecular biology: Splicing does the two-step. *Nature*, *521*(7552), 300–301. https://doi.org/10.1038/nature14524

Cordin, O., Hahn, D., & Beggs, J. D. (2012). Structure, function and regulation of spliceosomal RNA helicases. *Current Opinion in Cell Biology*, *24*(3), 431–438. https://doi.org/10.1016/j.ceb.2012.03.004

Coté, J., Dupuis, S., Jiang, Z., & Wu, J. Y. (2001). Caspase-2 pre-mRNA alternative splicing: Identification of an intronic element containing a decoy 3' acceptor site. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(3), 938–943. https://doi.org/10.1073/pnas.031564098

Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Research*, *14*(6), 1188–1190. https://doi.org/10.1101/gr.849004

Darnell, J. E., Philipson, L., Wall, R., & Adesnik, M. (1971). Polyadenylic acid sequences: role in conversion of nuclear RNA into messenger RNA. *Science (New York, N.Y.)*, *174*(4008), 507–510. https://doi.org/10.1126/science.174.4008.507

De Conti, L, Baralle, M., & Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA*, *4*(1), 49–60. https://doi.org/10.1002/wrna.1140

De Conti, Laura, Baralle, M., & Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. *Wiley Interdisciplinary Reviews. RNA*, *4*(1), 49–60. https://doi.org/10.1002/wrna.1140

de la Mata, M., Alonso, C. R., Kadener, S., Fededa, J. P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D., & Kornblihtt, A. R. (2003). A slow RNA polymerase II affects alternative splicing in vivo. *Molecular Cell*, *12*(2), 525–532. https://doi.org/10.1016/j.molcel.2003.08.001

de Navas, L. F., Reed, H., Akam, M., Barrio, R., Alonso, C. R., & Sanchez-Herrero, E. (2011). Integration of RNA processing and expression level control modulates the function of the Drosophila Hox gene Ultrabithorax during adult development. *Development*, *138*(1), 107–116. https://doi.org/10.1242/dev.051409

Dibb, N. J., & Newman, A. J. (1989). Evidence that introns arose at proto-splice sites. *The EMBO Journal*, *8*(7), 2015–2021. http://www.ncbi.nlm.nih.gov/pubmed/2792080

Domdey, H., Apostol, B., Lin, R. J., Newman, A., Brody, E., & Abelson, J. (1984). Lariat structures are in vivo intermediates in yeast pre-mRNA splicing. *Cell*, *39*(3 Pt 2), 611–621. https://doi.org/10.1016/0092-8674(84)90468-9

Dominski, Z., & Kole, R. (1992). Cooperation of pre-mRNA sequence elements in splice site selection. *Molecular and Cellular Biology*, *12*(5), 2108–2114. https://doi.org/10.1128/mcb.12.5.2108

Drexler, H. L., Choquet, K., & Churchman, L. S. (2020). Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Mol Cell*, *77*(5), 985-998 e8. https://doi.org/10.1016/j.molcel.2019.11.017

Duff, M. O., Olson, S., Wei, X., Garrett, S. C., Osman, A., Bolisetty, M., Plocik, A., Celniker, S. E., & Graveley, B. R. (2015). Genome-wide identification of zero

nucleotide recursive splicing in Drosophila. *Nature*, *521*(7552), 376–379. https://doi.org/10.1038/nature14475

Dujardin, G., Lafaille, C., Petrillo, E., Buggiano, V., Gómez Acuña, L. I., Fiszbein, A., Godoy Herz, M. A., Nieto Moreno, N., Muñoz, M. J., Alló, M., Schor, I. E., & Kornblihtt, A. R. (2013). Transcriptional elongation and alternative splicing. *Biochimica et Biophysica Acta*, *1829*(1), 134–140. https://doi.org/10.1016/j.bbagrm.2012.08.005

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R., & Hood, L. (1980). Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell*, *20*(2), 313–319. https://doi.org/10.1016/0092-8674(80)90617-0

Erkelenz, S., Mueller, W. F., Evans, M. S., Busch, A., Schöneweis, K., Hertel, K. J., & Schaal, H. (2013). Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA (New York, N.Y.)*, *19*(1), 96–102. https://doi.org/10.1261/rna.037044.112

Fabrizio, P., Dannenberg, J., Dube, P., Kastner, B., Stark, H., Urlaub, H., & Lührmann, R. (2009). The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. *Molecular Cell*, *36*(4), 593–608. https://doi.org/10.1016/j.molcel.2009.09.040

Fambrough, D., Pan, D., Rubin, G. M., & Goodman, C. S. (1996). The cell surface metalloprotease/disintegrin Kuzbanian is required for axonal extension in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(23), 13233–13238. https://doi.org/10.1073/pnas.93.23.13233

Ferrari, F., Plachetka, A., Alekseyenko, A. A., Jung, Y. L., Ozsolak, F., Kharchenko, P. V, Park, P. J., & Kuroda, M. I. (2013). “Jump start and gain” model for dosage compensation in Drosophila based on direct sequencing of nascent transcripts. *Cell Reports*, *5*(3), 629–636. https://doi.org/10.1016/j.celrep.2013.09.037

Fica, S. M., Oubridge, C., Galej, W. P., Wilkinson, M. E., Bai, X.-C., Newman, A. J., & Nagai, K. (2017). Structure of a spliceosome remodelled for exon ligation. *Nature*, *542*(7641), 377–380. https://doi.org/10.1038/nature21078

Fica, S. M., Tuttle, N., Novak, T., Li, N.-S., Lu, J., Koodathingal, P., Dai, Q., Staley, J. P., & Piccirilli, J. A. (2013). RNA catalyses nuclear pre-mRNA splicing. *Nature*, *503*(7475), 229–234. https://doi.org/10.1038/nature12734

Fourmann, J.-B., Schmitzová, J., Christian, H., Urlaub, H., Ficner, R., Boon, K.-L., Fabrizio, P., & Lührmann, R. (2013). Dissection of the factor requirements for spliceosome disassembly and the elucidation of its dissociation products using a purified splicing system. *Genes & Development*, *27*(4), 413–428. https://doi.org/10.1101/gad.207779.112

Fukaya, T., Lim, B., & Levine, M. (2017). Rapid Rates of Pol II Elongation in the Drosophila Embryo. *Current Biology : CB*, *27*(9), 1387–1391. https://doi.org/10.1016/j.cub.2017.03.069

Fukumura, K., Wakabayashi, S., Kataoka, N., Sakamoto, H., Suzuki, Y., Nakai, K., Mayeda, A., & Inoue, K. (2016). The Exon Junction Complex Controls the Efficient and Faithful Splicing of a Subset of Transcripts Involved in Mitotic Cell-Cycle Progression. *International Journal of Molecular Sciences*, *17*(8). https://doi.org/10.3390/ijms17081153

Furth, P. A., Choe, W. T., Rex, J. H., Byrne, J. C., & Baker, C. C. (1994). Sequences homologous to 5’ splice sites are required for the inhibitory activity of papillomavirus late 3’ untranslated regions. *Molecular and Cellular Biology*, *14*(8), 5278–5289. https://doi.org/10.1128/mcb.14.8.5278

Galej, W. P., Wilkinson, M. E., Fica, S. M., Oubridge, C., Newman, A. J., & Nagai, K.

(2016). Cryo-EM structure of the spliceosome immediately after branching. *Nature*, *537*(7619), 197–201. https://doi.org/10.1038/nature19316
Georgomanolis, T., Sofiadis, K., & Papantonis, A. (2016). Cutting a long intron short: Recursive splicing and its implications. *Frontiers in Physiology*, *7*(NOV), 1–5. https://doi.org/10.3389/fphys.2016.00598
Geyer, A., Koltsaki, I., Hessinger, C., Renner, S., & Rogulja-Ortmann, A. (2015). Impact of Ultrabithorax alternative splicing on Drosophila embryonic nervous system development. *Mechanisms of Development*, *138 Pt 2*, 177–189. https://doi.org/10.1016/j.mod.2015.08.007
Gilbert, W. (1978a). Why genes in pieces? *Nature*, *271*(5645), 501. https://doi.org/10.1038/271501a0
Gilbert, W. (1978b). Why genes in pieces? *Nature*, *271*(5645), 501. https://doi.org/10.1038/271501a0
González-Morales, N., Géminard, C., Lebreton, G., Cerezo, D., Coutelis, J.-B., & Noselli, S. (2015). The Atypical Cadherin Dachsous Controls Left-Right Asymmetry in Drosophila. *Developmental Cell*, *33*(6), 675–689. https://doi.org/10.1016/j.devcel.2015.04.026
Gratz, S. J., Ukken, F. P., Rubinstein, C. D., Thiede, G., Donohue, L. K., Cummings, A. M., & O'Connor-Giles, K. M. (2014). Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in Drosophila. *Genetics*, *196*(4), 961–971. https://doi.org/10.1534/genetics.113.160713
Grau-Bové, X., Ruiz-Trillo, I., & Irimia, M. (2018). Origin of exon skipping-rich transcriptomes in animals driven by evolution of gene architecture. *Genome Biology*, *19*(1), 135. https://doi.org/10.1186/s13059-018-1499-9
GROS, F., HIATT, H., GILBERT, W., KURLAND, C. G., RISEBROUGH, R. W., & WATSON, J. D. (1961). Unstable ribonucleic acid revealed by pulse labelling of Escherichia coli. *Nature*, *190*, 581–585. https://doi.org/10.1038/190581a0
Groth, A. C., Fish, M., Nusse, R., & Calos, M. P. (2004). Construction of transgenic Drosophila by using the site-specific integrase from phage phiC31. *Genetics*, *166*(4), 1775–1782. https://doi.org/10.1534/genetics.166.4.1775
Guan, F., Caratozzolo, R. M., Goraczniak, R., Ho, E. S., & Gunderson, S. I. (2007). A bipartite U1 site represses U1A expression by synergizing with PIE to inhibit nuclear polyadenylation. *RNA (New York, N.Y.)*, *13*(12), 2129–2140. https://doi.org/10.1261/rna.756707
Hardy, S. F., Grabowski, P. J., Padgett, R. A., & Sharp, P. A. (1984). Cofactor requirements of splicing of purified messenger RNA precursors. *Nature*, *308*(5957), 375–377. https://doi.org/10.1038/308375a0
HARRIS, H. (1959). Turnover of nuclear and cytoplasmic ribonucleic acid in two types of animal cell, with some further observations on the nucleolus. *The Biochemical Journal*, *73*, 362–369. https://doi.org/10.1042/bj0730362
HARRIS, H., & WATTS, J. W. (1962). The relationship between nuclear and cytoplasmic ribonucleic acid. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *156*, 109–121. https://doi.org/10.1098/rspb.1962.0031
Haselbach, D., Komarov, I., Agafonov, D. E., Hartmuth, K., Graf, B., Dybkov, O., Urlaub, H., Kastner, B., Lührmann, R., & Stark, H. (2018). Structure and Conformational Dynamics of the Human Spliceosomal Bact Complex. *Cell*, *172*(3), 454-464.e11. https://doi.org/10.1016/j.cell.2018.01.010
Hastings, K. E. M. (2005). SL trans-splicing: easy come or easy go? *Trends in Genetics : TIG*, *21*(4), 240–247. https://doi.org/10.1016/j.tig.2005.02.005
Hatton, A. R., Subramaniam, V., & Lopez, A. J. (1998a). Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-

exon junctions. *Molecular Cell*, *2*(6), 787–796. https://doi.org/10.1016/s1097-2765(00)80293-2
Hatton, A. R., Subramaniam, V., & Lopez, A. J. (1998b). Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. *Molecular Cell*, *2*(6), 787–796. https://doi.org/S1097-2765(00)80293-2 [pii]
Hausner, T. P., Giglio, L. M., & Weiner, A. M. (1990). Evidence for base-pairing between mammalian U2 and U6 small nuclear ribonucleoprotein particles. *Genes & Development*, *4*(12A), 2146–2156. https://doi.org/10.1101/gad.4.12a.2146
Haussmann, I. U., Bodi, Z., Sanchez-Moran, E., Mongan, N. P., Archer, N., Fray, R. G., & Soller, M. (2016). m(6)A potentiates Sxl alternative pre-mRNA splicing for robust Drosophila sex determination. *Nature*, *540*(7632), 301–304. https://doi.org/10.1038/nature20577
Hayashi, R, Handler, D., Ish-Horowicz, D., & Brennecke, J. (2014). The exon junction complex is required for definition and excision of neighboring introns in Drosophila. *Genes Dev*, *28*(16), 1772–1785. https://doi.org/10.1101/gad.245738.114
Hayashi, Rippei, Handler, D., Ish-Horowicz, D., & Brennecke, J. (2014). The exon junction complex is required for definition and excision of neighboring introns in Drosophila. *Genes & Development*, *28*(16), 1772–1785. https://doi.org/10.1101/gad.245738.114
He, F., & Jacobson, A. (2015). Nonsense-Mediated mRNA Decay: Degradation of Defective Transcripts Is Only Part of the Story. *Annual Review of Genetics*, *49*, 339–366. https://doi.org/10.1146/annurev-genet-112414-054639
Heath, C. G., Viphakone, N., & Wilson, S. A. (2016). The role of TREX in gene expression and disease. *The Biochemical Journal*, *473*(19), 2911–2935. https://doi.org/10.1042/BCJ20160010
Hernandez, N., & Keller, W. (1983). Splicing of in vitro synthesized messenger RNA precursors in HeLa cell extracts. *Cell*, *35*(1), 89–99. https://doi.org/10.1016/0092-8674(83)90211-8
Hollander, D., Naftelberg, S., Lev-Maor, G., Kornblihtt, A. R., & Ast, G. (2016). How Are Short Exons Flanked by Long Introns Defined and Committed to Splicing? *Trends in Genetics*, *32*(10), 596–606. https://doi.org/10.1016/j.tig.2016.07.003
Horowitz, D. S. (2012). The mechanism of the second step of pre-mRNA splicing. *Wiley Interdisciplinary Reviews. RNA*, *3*(3), 331–350. https://doi.org/10.1002/wrna.112
Huff, J. T., Plocik, A. M., Guthrie, C., & Yamamoto, K. R. (2010). Reciprocal intronic and exonic histone modification regions in humans. *Nature Structural & Molecular Biology*, *17*(12), 1495–1499. https://doi.org/10.1038/nsmb.1924
Irion, U. (2012). Drosophila muscleblind Codes for Proteins with One and Two Tandem Zinc Finger Motifs. *PLoS ONE*, *7*(3), e34248. https://doi.org/10.1371/journal.pone.0034248
James, S.-A., Turner, W., & Schwer, B. (2002). How Slu7 and Prp18 cooperate in the second step of yeast pre-mRNA splicing. *RNA (New York, N.Y.)*, *8*(8), 1068–1077. https://doi.org/10.1017/s1355838202022033
Jonkers, I., & Lis, J. T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. *Nature Reviews. Molecular Cell Biology*, *16*(3), 167–177. https://doi.org/10.1038/nrm3953
Joseph, B., Kondo, S., & Lai, E. C. (2018). Short cryptic exons mediate recursive splicing in Drosophila. *Nature Structural & Molecular Biology*, *25*(5), 365–371. https://doi.org/10.1038/s41594-018-0052-6
Kahles, A., Lehmann, K.-V., Toussaint, N. C., Hüser, M., Stark, S. G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., Cancer Genome Atlas Research Network,

& Rätsch, G. (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell*, *34*(2), 211-224.e6. https://doi.org/10.1016/j.ccell.2018.07.001

Kan, L., Grozhik, A. V, Vedanayagam, J., Patil, D. P., Pang, N., Lim, K. S., Huang, Y. C., Joseph, B., Lin, C. J., Despic, V., Guo, J., Yan, D., Kondo, S., Deng, W. M., Dedon, P. C., Jaffrey, S. R., & Lai, E. C. (2017). The m(6)A pathway facilitates sex determination in Drosophila. *Nat Commun*, *8*, 15737. https://doi.org/10.1038/ncomms15737

Kastner, B., Will, C. L., Stark, H., & Lührmann, R. (2019). Structural Insights into Nuclear pre-mRNA Splicing in Higher Eukaryotes. *Cold Spring Harbor Perspectives in Biology*, *11*(11). https://doi.org/10.1101/cshperspect.a032417

Kelly, S., Georgomanolis, T., Zirkel, A., Diermeier, S., O'Reilly, D., Murphy, S., Längst, G., Cook, P. R., & Papantonis, A. (2015). Splicing of many human genes involves sites embedded within introns. *Nucleic Acids Research*, *43*(9), 4721–4732. https://doi.org/10.1093/nar/gkv386

Keren, H., Lev-Maor, G., & Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews. Genetics*, *11*(5), 345–355. https://doi.org/10.1038/nrg2776

Khodor, Y. L., Rodriguez, J., Abruzzi, K. C., Tang, C.-H. A., Marr, M. T., & Rosbash, M. (2011). Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes & Development*, *25*(23), 2502–2512. https://doi.org/10.1101/gad.178962.111

Kielkopf, C. L., Rodionova, N. A., Green, M. R., & Burley, S. K. (2001). A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. *Cell*, *106*(5), 595–605. https://doi.org/10.1016/s0092-8674(01)00480-9

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. https://doi.org/10.1186/gb-2013-14-4-r36

Kim, E., Magen, A., & Ast, G. (2007). Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, *35*(1), 125–131. https://doi.org/10.1093/nar/gkl924

Kim, V. N., Kataoka, N., & Dreyfuss, G. (2001). Role of the nonsense-mediated decay factor hUpf3 in the splicing-dependent exon-exon junction complex. *Science (New York, N.Y.)*, *293*(5536), 1832–1836. https://doi.org/10.1126/science.1062829

Kishor, A., Fritz, S. E., & Hogg, J. R. (2019). Nonsense-mediated mRNA decay: The challenge of telling right from wrong in a complex transcriptome. *Wiley Interdisciplinary Reviews. RNA*, *10*(6), e1548. https://doi.org/10.1002/wrna.1548

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X. S., & Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature Genetics*, *41*(3), 376–381. https://doi.org/10.1038/ng.322

Kondo, S., & Ueda, R. (2013). Highly improved gene targeting by germline-specific Cas9 expression in Drosophila. *Genetics*, *195*(3), 715–721. https://doi.org/10.1534/genetics.113.156737

Kondo, S., Vedanayagam, J., Mohammed, J., Eizadshenass, S., Kan, L., Pang, N., Aradhya, R., Siepel, A., Steinhauer, J., & Lai, E. C. (2017). New genes often acquire male-specific functions but rarely become essential in Drosophila. *Genes & Development*, *31*(18), 1841–1846. https://doi.org/10.1101/gad.303131.117

Kondo, Y., Oubridge, C., van Roon, A. M. M., & Nagai, K. (2015). Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *ELife*, *4*, 1–19. https://doi.org/10.7554/eLife.04986

Koodathingal, P., Novak, T., Piccirilli, J. A., & Staley, J. P. (2010). The DEAH box ATPases Prp16 and Prp43 cooperate to proofread 5' splice site cleavage during pre-mRNA splicing. *Molecular Cell*, *39*(3), 385–395. https://doi.org/10.1016/j.molcel.2010.07.014

Lai, E. C., Woo, S. L., Dugaiczyk, A., Catterall, J. F., & O'Malley, B. W. (1978). The ovalbumin gene: structural sequences in native chicken DNA are not contiguous. *Proceedings of the National Academy of Sciences of the United States of America*, *75*(5), 2205–2209. https://doi.org/10.1073/pnas.75.5.2205

Lawrence, M., Daujat, S., & Schneider, R. (2016). Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics : TIG*, *32*(1), 42–56. https://doi.org/10.1016/j.tig.2015.10.007

Le Hir, H, Gatfield, D., Izaurralde, E., & Moore, M. J. (2001). The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *The EMBO Journal*, *20*(17), 4987–4997. https://doi.org/10.1093/emboj/20.17.4987

Le Hir, H, Izaurralde, E., Maquat, L. E., & Moore, M. J. (2000). The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. *The EMBO Journal*, *19*(24), 6860–6869. https://doi.org/10.1093/emboj/19.24.6860

Le Hir, Hervé, Saulière, J., & Wang, Z. (2016). The exon junction complex as a node of post-transcriptional networks. *Nature Reviews. Molecular Cell Biology*, *17*(1), 41–54. https://doi.org/10.1038/nrm.2015.7

Lee, Y., & Rio, D. C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry*, *84*, 291–323. https://doi.org/10.1146/annurev-biochem-060614-034316

Leeds, N. B., Small, E. C., Hiley, S. L., Hughes, T. R., & Staley, J. P. (2006). The splicing factor Prp43p, a DEAH box ATPase, functions in ribosome biogenesis. *Molecular and Cellular Biology*, *26*(2), 513–522. https://doi.org/10.1128/MCB.26.2.513-522.2006

Lei, Q., Li, C., Zuo, Z., Huang, C., Cheng, H., & Zhou, R. (2016). Evolutionary Insights into RNA trans-Splicing in Vertebrates. *Genome Biology and Evolution*, *8*(3), 562–577. https://doi.org/10.1093/gbe/evw025

LeMaire, M. F., & Thummel, C. S. (1990). Splicing precedes polyadenylation during Drosophila E74A transcription. *Molecular and Cellular Biology*, *10*(11), 6059–6063. https://doi.org/10.1128/mcb.10.11.6059

Lence, T., Akhtar, J., Bayer, M., Schmid, K., Spindler, L., Ho, C. H., Kreim, N., Andrade-Navarro, M. A., Poeck, B., Helm, M., & Roignant, J. Y. (2016). m(6)A modulates neuronal functions and sex determination in Drosophila. *Nature*, *540*(7632), 242–247. https://doi.org/10.1038/nature20568

Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L., & Steitz, J. A. (1980). Are snRNPs involved in splicing? *Nature*, *283*(5743), 220–224. https://doi.org/10.1038/283220a0

Lerner, M. R., & Steitz, J. A. (1979). Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(11), 5495–5499. https://doi.org/10.1073/pnas.76.11.5495

Leung, C. S., Douglass, S. M., Morselli, M., Obusan, M. B., Pavlyukov, M. S., Pellegrini, M., & Johnson, T. L. (2019). H3K36 Methylation and the Chromodomain Protein

Eaf3 Are Required for Proper Cotranscriptional Spliceosome Assembly. *Cell Reports*, *27*(13), 3760-3769.e4. https://doi.org/10.1016/j.celrep.2019.05.100
Li, S., & Mason, C. E. (2014). The pivotal regulatory landscape of RNA modifications. *Annual Review of Genomics and Human Genetics*, *15*, 127–150. https://doi.org/10.1146/annurev-genom-090413-025405
Litterman, A. J., Kageyama, R., Le Tonqueze, O., Zhao, W., Gagnon, J. D., Goodarzi, H., Erle, D. J., & Ansel, K. M. (2019). A massively parallel 3′ UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. *Genome Research*, *29*(6), 896–906. https://doi.org/10.1101/gr.242552.118
Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., Jia, G., Yu, M., Lu, Z., Deng, X., Dai, Q., Chen, W., & He, C. (2014). A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nature Chemical Biology*, *10*(2), 93–95. https://doi.org/10.1038/nchembio.1432
Liu, S., Li, X., Zhang, L., Jiang, J., Hill, R. C., Cui, Y., Hansen, K. C., Zhou, Z. H., & Zhao, R. (2017). Structure of the yeast spliceosomal postcatalytic P complex. *Science (New York, N.Y.)*, *358*(6368), 1278–1283. https://doi.org/10.1126/science.aar3462
Long, M., Rosenberg, C., & Gilbert, W. (1995a). Intron phase correlations and the evolution of the intron/exon structure of genes. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(26), 12495–12499. https://doi.org/10.1073/pnas.92.26.12495
Long, M., Rosenberg, C., & Gilbert, W. (1995b). Intron phase correlations and the evolution of the intron/exon structure of genes. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(26), 12495–12499. https://doi.org/10.1073/pnas.92.26.12495
Luco, R. F., & Misteli, T. (2011). More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. *Current Opinion in Genetics & Development*, *21*(4), 366–372. https://doi.org/10.1016/j.gde.2011.03.004
Luco, R. F., Pan, Q., Tominaga, K., Blencowe, B. J., Pereira-Smith, O. M., & Misteli, T. (2010). Regulation of alternative splicing by histone modifications. *Science (New York, N.Y.)*, *327*(5968), 996–1000. https://doi.org/10.1126/science.1184208
Lykke-Andersen, J., Shu, M. D., & Steitz, J. A. (2001). Communication of the position of exon-exon junctions to the mRNA surveillance machinery by the protein RNPS1. *Science (New York, N.Y.)*, *293*(5536), 1836–1839. https://doi.org/10.1126/science.1062786
Ma, X. M., Yoon, S.-O., Richardson, C. J., Jülich, K., & Blenis, J. (2008). SKAR links pre-mRNA splicing to mTOR/S6K1-mediated enhanced translation efficiency of spliced mRNAs. *Cell*, *133*(2), 303–313. https://doi.org/10.1016/j.cell.2008.02.031
Madhani, H. D., & Guthrie, C. (1992). A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome. *Cell*, *71*(5), 803–817. https://doi.org/10.1016/0092-8674(92)90556-r
Malartre, M. (2016). Regulatory mechanisms of EGFR signalling during Drosophila eye development. *Cellular and Molecular Life Sciences : CMLS*, *73*(9), 1825–1843. https://doi.org/10.1007/s00018-016-2153-x
Malone, C D, Mestdagh, C., Akhtar, J., Kreim, N., Deinhard, P., Sachidanandam, R., Treisman, J., & Roignant, J. Y. (2014). The exon junction complex controls transposable element activity by ensuring faithful splicing of the piwi transcript. *Genes & Development*, *28*(16), 1786–1799. https://doi.org/10.1101/gad.245829.114 [doi]

Malone, Colin D, Mestdagh, C., Akhtar, J., Kreim, N., Deinhard, P., Sachidanandam, R., Treisman, J., & Roignant, J.-Y. (2014). The exon junction complex controls transposable element activity by ensuring faithful splicing of the piwi transcript. *Genes & Development*, *28*(16), 1786–1799. https://doi.org/10.1101/gad.245829.114

Maquat, L. E., Kinniburgh, A. J., Beach, L. R., Honig, G. R., Lazerson, J., Ershler, W. B., & Ross, J. (1980). Processing of human beta-globin mRNA precursor to mRNA is defective in three patients with beta+-thalassemia. *Proceedings of the National Academy of Sciences of the United States of America*, *77*(7), 4287–4291. https://doi.org/10.1073/pnas.77.7.4287

McCracken, S., Fong, N., Rosonina, E., Yankulov, K., Brothers, G., Siderovski, D., Hessel, A., Foster, S., Shuman, S., & Bentley, D. L. (1997). 5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Genes & Development*, *11*(24), 3306–3318. https://doi.org/10.1101/gad.11.24.3306

McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S. D., Wickens, M., & Bentley, D. L. (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature*, *385*(6614), 357–361. https://doi.org/10.1038/385357a0

McMahon, A. C., Rahman, R., Jin, H., Shen, J. L., Fieldsend, A., Luo, W., & Rosbash, M. (2016). TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. *Cell*, *165*(3), 742–753. https://doi.org/10.1016/j.cell.2016.03.007

Milligan, L., Sayou, C., Tuck, A., Auchynnikava, T., Reid, J. E., Alexander, R., Alves, F. de L., Allshire, R., Spanos, C., Rappsilber, J., Beggs, J. D., Kudla, G., & Tollervey, D. (2017). RNA polymerase II stalling at pre-mRNA splice sites is enforced by ubiquitination of the catalytic subunit. *ELife*, *6*. https://doi.org/10.7554/eLife.27082

Mohn, F., Sienski, G., Handler, D., & Brennecke, J. (2014). The rhino-deadlock-cutoff complex licenses noncanonical transcription of dual-strand piRNA clusters in Drosophila. *Cell*, *157*(6), 1364–1379. https://doi.org/10.1016/j.cell.2014.04.031

Mortillaro, M. J., Blencowe, B. J., Wei, X., Nakayasu, H., Du, L., Warren, S. L., Sharp, P. A., & Berezney, R. (1996). A hyperphosphorylated form of the large subunit of RNA polymerase II is associated with splicing complexes and the nuclear matrix. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(16), 8253–8257. https://doi.org/10.1073/pnas.93.16.8253

Newman, A. J., & Norman, C. (1992). U5 snRNA interacts with exon sequences at 5' and 3' splice sites. *Cell*, *68*(4), 743–754. https://doi.org/10.1016/0092-8674(92)90149-7

Nguyen, T. H. D., Galej, W. P., Bai, X., Savva, C. G., Newman, A. J., Scheres, S. H. W., & Nagai, K. (2015). The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature*, *523*(7558), 47–52. https://doi.org/10.1038/nature14548

Nicholson, P., & Mühlemann, O. (2010). Cutting the nonsense: the degradation of PTC-containing mRNAs. *Biochemical Society Transactions*, *38*(6), 1615–1620. https://doi.org/10.1042/BST0381615

Oh, J.-M., Di, C., Venters, C. C., Guo, J., Arai, C., So, B. R., Pinto, A. M., Zhang, Z., Wan, L., Younis, I., & Dreyfuss, G. (2017). U1 snRNP telescripting regulates a size-function-stratified human genome. *Nature Structural & Molecular Biology*, *24*(11), 993–999. https://doi.org/10.1038/nsmb.3473

Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S., & Sharp, P. A. (1986). Splicing of Messenger RNA Precursors. *Annual Review of Biochemistry*, *55*(1), 1119–1150. https://doi.org/10.1146/annurev.bi.55.070186.005351

Padgett, R. A., Hardy, S. F., & Sharp, P. A. (1983). Splicing of adenovirus RNA in a cell-

free transcription system. *Proceedings of the National Academy of Sciences of the United States of America*, *80*(17), 5230–5234. https://doi.org/10.1073/pnas.80.17.5230
Padgett, R. A., Konarska, M. M., Grabowski, P. J., Hardy, S. F., & Sharp, P. A. (1984). Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. *Science (New York, N.Y.)*, *225*(4665), 898–903. https://doi.org/10.1126/science.6206566
Padgett, R. A., Mount, S. M., Steitz, J. A., & Sharp, P. A. (1983). Splicing of messenger RNA precursors is inhibited by antisera to small nuclear ribonucleoprotein. *Cell*, *35*(1), 101–107. https://doi.org/10.1016/0092-8674(83)90212-x
Pai, A. A., Paggi, J. M., Yan, P., Adelman, K., & Burge, C. B. (2018). Numerous recursive sites contribute to accuracy of splicing in long introns in flies. *PLoS Genetics*, *14*(8), e1007588. https://doi.org/10.1371/journal.pgen.1007588
Palmiter, R. D., Sandgren, E. P., Avarbock, M. R., Allen, D. D., & Brinster, R. L. (1991). Heterologous introns can enhance expression of transgenes in mice. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(2), 478–482. https://doi.org/10.1073/pnas.88.2.478
Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, *40*(12), 1413–1415. https://doi.org/10.1038/ng.259
Pandya-Jones, A., & Black, D. L. (2009). Co-transcriptional splicing of constitutive and alternative exons. *RNA (New York, N.Y.)*, *15*(10), 1896–1908. https://doi.org/10.1261/rna.1714509
Parker, R., & Siliciano, P. G. (1993). Evidence for an essential non-Watson-Crick interaction between the first and last nucleotides of a nuclear pre-mRNA intron. *Nature*, *361*(6413), 660–662. https://doi.org/10.1038/361660a0
Parker, R., Siliciano, P. G., & Guthrie, C. (1987). Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA. *Cell*, *49*(2), 229–239. https://doi.org/10.1016/0092-8674(87)90564-2
Ping, X.-L., Sun, B.-F., Wang, L., Xiao, W., Yang, X., Wang, W.-J., Adhikari, S., Shi, Y., Lv, Y., Chen, Y.-S., Zhao, X., Li, A., Yang, Y., Dahal, U., Lou, X.-M., Liu, X., Huang, J., Yuan, W.-P., Zhu, X.-F., … Yang, Y.-G. (2014). Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. *Cell Research*, *24*(2), 177–189. https://doi.org/10.1038/cr.2014.3
Port, F., & Bullock, S. L. (2016). Augmenting CRISPR applications in Drosophila with tRNA-flanked sgRNAs. *Nature Methods*, *13*(10), 852–854. https://doi.org/10.1038/nmeth.3972
Port, F., Chen, H.-M., Lee, T., & Bullock, S. L. (2014). Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(29), E2967-76. https://doi.org/10.1073/pnas.1405500111
Pyle, A. M. (2008). Translocation and unwinding mechanisms of RNA and DNA helicases. *Annual Review of Biophysics*, *37*, 317–336. https://doi.org/10.1146/annurev.biophys.37.032807.125908
Rappsilber, J., Ryder, U., Lamond, A. I., & Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. *Genome Research*, *12*(8), 1231–1245. https://doi.org/10.1101/gr.473902
Reese, M. G., Eeckman, F. H., Kulp, D., & Haussler, D. (1997). Improved splice site detection in Genie. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *4*(3), 311–323. https://doi.org/10.1089/cmb.1997.4.311
Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G.,

& Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754
Roca, X., Krainer, A. R., & Eperon, I. C. (2013). Pick one, but be quick: 5′ splice sites and the problems of too many choices. *Genes and Development*, *27*(2), 129–144. https://doi.org/10.1101/gad.209759.112
Rodriguez, J., Menet, J. S., & Rosbash, M. (2012). Nascent-seq indicates widespread cotranscriptional RNA editing in Drosophila. *Molecular Cell*, *47*(1), 27–37. https://doi.org/10.1016/j.molcel.2012.05.002
Rodriguez, J. R., Pikielny, C. W., & Rosbash, M. (1984). In vivo characterization of yeast mRNA processing intermediates. *Cell*, *39*(3 Pt 2), 603–610. https://doi.org/10.1016/0092-8674(84)90467-7
Roignant, J.-Y., & Treisman, J. E. (2010). Exon junction complex subunits are required to splice Drosophila MAP kinase, a large heterochromatic gene. *Cell*, *143*(2), 238–250. https://doi.org/10.1016/j.cell.2010.09.036
Rottman, F., Shatkin, A. J., & Perry, R. P. (1974). Sequences containing methylated nucleotides at the 5’ termini of messenger RNAs: possible implications for processing. *Cell*, *3*(3), 197–199. https://doi.org/10.1016/0092-8674(74)90131-7
Roundtree, I. A., Evans, M. E., Pan, T., & He, C. (2017). Dynamic RNA Modifications in Gene Expression Regulation. *Cell*, *169*(7), 1187–1200. https://doi.org/10.1016/j.cell.2017.05.045
Rozhkov, N. V, Hammell, M., & Hannon, G. J. (2013). Multiple roles for Piwi in silencing Drosophila transposons. *Genes & Development*, *27*(4), 400–412. https://doi.org/10.1101/gad.209767.112
Ruskin, B., Krainer, A. R., Maniatis, T., & Green, M. R. (1984). Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. *Cell*, *38*(1), 317–331. https://doi.org/10.1016/0092-8674(84)90553-1
Saavedra, P., Brittle, A., Palacios, I. M., Strutt, D., Casal, J., & Lawrence, P. A. (2016). Planar cell polarity: the Dachsous/Fat system contributes differently to the embryonic and larval stages of Drosophila. *Biology Open*, *5*(4), 397–408. https://doi.org/10.1242/bio.017152
Sadusky, T., Newman, A. J., & Dibb, N. J. (2004). Exon junction sequences as cryptic splice sites: implications for intron origin. *Current Biology : CB*, *14*(6), 505–509. https://doi.org/10.1016/j.cub.2004.02.063
Saldi, T., Cortazar, M. A., Sheridan, R. M., & Bentley, D. L. (2016). Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *Journal of Molecular Biology*, *428*(12), 2623–2635. https://doi.org/10.1016/j.jmb.2016.04.017
Salzman, J., Chen, R. E., Olsen, M. N., Wang, P. L., & Brown, P. O. (2013). Cell-type specific features of circular RNA expression. *PLoS Genetics*, *9*(9), e1003777. https://doi.org/10.1371/journal.pgen.1003777 [doi]
Salzman, J., Gawad, C., Wang, P. L., Lacayo, N., & Brown, P. O. (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PloS One*, *7*(2), e30733. https://doi.org/10.1371/journal.pone.0030733 [doi]
Sanfilippo, P., Wen, J., & Lai, E. C. (2017). Landscape and evolution of tissue-specific alternative polyadenylation across Drosophila species. *Genome Biology*, *18*(1), 229. https://doi.org/10.1186/s13059-017-1358-0
SCHERRER, K., LATHAM, H., & DARNELL, J. E. (1963). Demonstration of an unstable RNA and of a precursor to ribosomal RNA in HeLa cells. *Proceedings of the National Academy of Sciences of the United States of America*, *49*, 240–248. https://doi.org/10.1073/pnas.49.2.240
Schlautmann, L. P., & Gehring, N. H. (2020). A Day in the Life of the Exon Junction

Complex. *Biomolecules*, *10*(6). https://doi.org/10.3390/biom10060866
Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E., & Zipursky, S. L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, *101*(6), 671–684. https://doi.org/10.1016/s0092-8674(00)80878-8
Schwer, B, & Guthrie, C. (1992). A conformational rearrangement in the spliceosome is dependent on PRP16 and ATP hydrolysis. *The EMBO Journal*, *11*(13), 5033–5039. http://www.ncbi.nlm.nih.gov/pubmed/1464325
Schwer, Beate. (2008). A conformational rearrangement in the spliceosome sets the stage for Prp22-dependent mRNA release. *Molecular Cell*, *30*(6), 743–754. https://doi.org/10.1016/j.molcel.2008.05.003
Scotti, M. M., & Swanson, M. S. (2016). RNA mis-splicing in disease. *Nature Reviews. Genetics*, *17*(1), 19–32. https://doi.org/10.1038/nrg.2015.3
Séraphin, B. (1995). Sm and Sm-like proteins belong to a large family: identification of proteins of the U6 as well as the U1, U2, U4 and U5 snRNPs. *The EMBO Journal*, *14*(9), 2089–2098. http://www.ncbi.nlm.nih.gov/pubmed/7744014
Sharma, S., Maris, C., Allain, F. H.-T., & Black, D. L. (2011). U1 snRNA directly interacts with polypyrimidine tract-binding protein during splicing repression. *Molecular Cell*, *41*(5), 579–588. https://doi.org/10.1016/j.molcel.2011.02.012
Shen, H., & Green, M. R. (2006). RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans. *Genes & Development*, *20*(13), 1755–1765. https://doi.org/10.1101/gad.1422106
Shepard, S., McCreary, M., & Fedorov, A. (2009). The peculiarities of large intron splicing in animals. *PloS One*, *4*(11), e7853. https://doi.org/10.1371/journal.pone.0007853
Sibley, C. R., Blazquez, L., & Ule, J. (2016). Lessons from non-canonical splicing. *Nature Reviews. Genetics*, *17*(7), 407–421. https://doi.org/10.1038/nrg.2016.46
Sibley, C. R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M. E., Hardy, J., Modic, M., Curk, T., Wilson, S. W., Plagnol, V., & Ule, J. (2015a). Recursive splicing in long vertebrate genes. *Nature*, *521*(7552), 371–375. https://doi.org/10.1038/nature14466
Sibley, C. R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M. E., Hardy, J., Modic, M., Curk, T., Wilson, S. W., Plagnol, V., & Ule, J. (2015b). Recursive splicing in long vertebrate genes. *Nature*, *521*(7552), 371–375. https://doi.org/10.1038/nature14466
Sienski, G., Dönertas, D., & Brennecke, J. (2012). Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell*, *151*(5), 964–980. https://doi.org/10.1016/j.cell.2012.10.040
Sims, R. J., Millhouse, S., Chen, C.-F., Lewis, B. A., Erdjument-Bromage, H., Tempst, P., Manley, J. L., & Reinberg, D. (2007). Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. *Molecular Cell*, *28*(4), 665–676. https://doi.org/10.1016/j.molcel.2007.11.010
Singh, G., Kucukural, A., Cenik, C., Leszyk, J. D., Shaffer, S. A., Weng, Z., & Moore, M. J. (2012). The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. *Cell*, *151*(4), 750–764. https://doi.org/10.1016/j.cell.2012.10.007
Singh, J., & Padgett, R. A. (2009). Rates of in situ transcription and splicing in large human genes. *Nature Structural and Molecular Biology*, *16*(11), 1128–1133. https://doi.org/10.1038/nsmb.1666
Soeiro, R., Birnboim, H. C., & Darnell, J. E. (1966). Rapidly labeled HeLa cell nuclear

RNA. II. Base composition and cellular localization of a heterogeneous RNA fraction. *Journal of Molecular Biology*, *19*(2), 362–372. https://doi.org/10.1016/s0022-2836(66)80010-4
Soeiro, R., Vaughan, M. H., Warner, J. R., & Darnell, J. E. (1968). The turnover of nuclear DNA-like RNA in HeLa cells. *The Journal of Cell Biology*, *39*(1), 112–118. https://doi.org/10.1083/jcb.39.1.112
Sontheimer, E. J., & Steitz, J. A. (1993). The U5 and U6 small nuclear RNAs as active site components of the spliceosome. *Science (New York, N.Y.)*, *262*(5142), 1989–1996. https://doi.org/10.1126/science.8266094
Sontheimer, E. J., Sun, S., & Piccirilli, J. A. (1997). Metal ion catalysis during splicing of premessenger RNA. *Nature*, *388*(6644), 801–805. https://doi.org/10.1038/42068
Spritz, R. A., Jagadeeswaran, P., Choudary, P. V, Biro, P. A., Elder, J. T., DeRiel, J. K., Manley, J. L., Gefter, M. L., Forget, B. G., & Weissman, S. M. (1981). Base substitution in an intervening sequence of a beta+-thalassemic human globin gene. *Proceedings of the National Academy of Sciences of the United States of America*, *78*(4), 2455–2459. https://doi.org/10.1073/pnas.78.4.2455
Steckelberg, A.-L., Boehm, V., Gromadzka, A. M., & Gehring, N. H. (2012). CWC22 connects pre-mRNA splicing and exon junction complex assembly. *Cell Reports*, *2*(3), 454–461. https://doi.org/10.1016/j.celrep.2012.08.017
Steitz, T. A., & Steitz, J. A. (1993). A general two-metal-ion mechanism for catalytic RNA. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(14), 6498–6502. https://doi.org/10.1073/pnas.90.14.6498
Subramaniam, V., Bomze, H. M., & López, A. J. (1994). Functional differences between Ultrabithorax protein isoforms in Drosophila melanogaster: evidence from elimination, substitution and ectopic expression of specific isoforms. *Genetics*, *136*(3), 979–991. http://www.ncbi.nlm.nih.gov/pubmed/7911774
Sun, S., Ling, S.-C., Qiu, J., Albuquerque, C. P., Zhou, Y., Tokunaga, S., Li, H., Qiu, H., Bui, A., Yeo, G. W., Huang, E. J., Eggan, K., Zhou, H., Fu, X.-D., Lagier-Tourenne, C., & Cleveland, D. W. (2015). ALS-causative mutations in FUS/TLS confer gain and loss of function by altered association with SMN and U1-snRNP. *Nature Communications*, *6*, 6171. https://doi.org/10.1038/ncomms7171
Taggart, A. J., Lin, C.-L., Shrestha, B., Heintzelman, C., Kim, S., & Fairbrother, W. G. (2017). Large-scale analysis of branchpoint usage across species and cell lines. *Genome Research*, *27*(4), 639–649. https://doi.org/10.1101/gr.202820.115
Takahara, K., Schwarze, U., Imamura, Y., Hoffman, G. G., Toriello, H., Smith, L. T., Byers, P. H., & Greenspan, D. S. (2002). Order of intron removal influences multiple splice outcomes, including a two-exon skip, in a COL5A1 acceptor-site mutation that results in abnormal pro-alpha1(V) N-propeptides and Ehlers-Danlos syndrome type I. *American Journal of Human Genetics*, *71*(3), 451–465. https://doi.org/10.1086/342099
Tarn, W. Y., Hsu, C. H., Huang, K. T., Chen, H. R., Kao, H. Y., Lee, K. R., & Cheng, S. C. (1994). Functional association of essential splicing factor(s) with PRP19 in a protein complex. *The EMBO Journal*, *13*(10), 2421–2431. http://www.ncbi.nlm.nih.gov/pubmed/8194532
Thomas, J. D., Polaski, J. T., Feng, Q., De Neef, E. J., Hoppe, E. R., McSharry, M. V, Pangallo, J., Gabel, A. M., Belleville, A. E., Watson, J., Nkinsi, N. T., Berger, A. H., & Bradley, R. K. (2020). RNA isoform screens uncover the essentiality and tumor-suppressor activity of ultraconserved poison exons. *Nature Genetics*, *52*(1), 84–94. https://doi.org/10.1038/s41588-019-0555-z
Tian, M., & Maniatis, T. (1993). A splicing enhancer complex controls alternative splicing of doublesex pre-mRNA. *Cell*, *74*(1), 105–114. https://doi.org/10.1016/0092-

8674(93)90298-5
Tilghman, S. M., Curtis, P. J., Tiemeier, D. C., Leder, P., & Weissmann, C. (1978). The intervening sequence of a mouse beta-globin gene is transcribed within the 15S beta-globin mRNA precursor. *Proceedings of the National Academy of Sciences of the United States of America*, *75*(3), 1309–1313. https://doi.org/10.1073/pnas.75.3.1309
Ule, J., & Blencowe, B. J. (2019). Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Molecular Cell*, *76*(2), 329–345. https://doi.org/10.1016/j.molcel.2019.09.017
Ustianenko, D., Weyn-Vanhentenryck, S. M., & Zhang, C. (2017). Microexons: discovery, regulation, and function. *Wiley Interdisciplinary Reviews. RNA*, *8*(4). https://doi.org/10.1002/wrna.1418
Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., González-Vallinas, J., Lahens, N. F., Hogenesch, J. B., Lynch, K. W., & Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. *ELife*, *5*, e11752. https://doi.org/10.7554/eLife.11752
Wall, R., Philipson, L., & Darnell, J. E. (1972). Processing of adenovirus specific nuclear RNA during virus replication. *Virology*, *50*(1), 27–34. https://doi.org/10.1016/0042-6822(72)90342-x
Wan, R., Bai, R., Yan, C., Lei, J., & Shi, Y. (2019). Structures of the Catalytically Activated Yeast Spliceosome Reveal the Mechanism of Branching. *Cell*, *177*(2), 339-351.e13. https://doi.org/10.1016/j.cell.2019.02.006
Wan, R., Yan, C., Bai, R., Wang, L., Huang, M., Wong, C. C. L., & Shi, Y. (2016). The 3.8 Å structure of the U4/U6.U5 tri-snRNP: Insights into spliceosome assembly and catalysis. *Science (New York, N.Y.)*, *351*(6272), 466–475. https://doi.org/10.1126/science.aad6466
Wang, E. T., Cody, N. A., Jog, S., Biancolella, M., Wang, T. T., Treacy, D. J., Luo, S., Schroth, G. P., Housman, D. E., Reddy, S., Lecuyer, E., & Burge, C. B. (2012). Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell*, *150*(4), 710–724. https://doi.org/10.1016/j.cell.2012.06.041 [doi]
Wang, P. L., Bao, Y., Yee, M. C., Barrett, S. P., Hogan, G. J., Olsen, M. N., Dinneny, J. R., Brown, P. O., & Salzman, J. (2014). Circular RNA is expressed across the eukaryotic tree of life. *PloS One*, *9*(6), e90859. https://doi.org/10.1371/journal.pone.0090859 [doi]
Wang, W., Han, B. W., Tipping, C., Ge, D. T., Zhang, Z., Weng, Z., & Zamore, P. D. (2015). Slicing and Binding by Ago3 or Aub Trigger Piwi-Bound piRNA Production by Distinct Mechanisms. *Molecular Cell*, *59*(5), 819–830. https://doi.org/10.1016/j.molcel.2015.08.007
Wang, Z, Rolish, M. E., Yeo, G., Tung, V., Mawson, M., & Burge, C. B. (2004). Systematic identification and analysis of exonic splicing silencers. *Cell*, *119*(6), 831–845. https://doi.org/S0092867404010566 [pii]
Wang, Zefeng, & Burge, C. B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA (New York, N.Y.)*, *14*(5), 802–813. https://doi.org/10.1261/rna.876308
Wang, Zhen, Murigneux, V., & Le Hir, H. (2014). Transcriptome-wide modulation of splicing by the exon junction complex. *Genome Biology*, *15*(12), 551. https://doi.org/10.1186/s13059-014-0551-7
Wei, P., Xue, W., Zhao, Y., Ning, G., & Wang, J. (2020). CRISPR-based modular assembly of a UAS-cDNA/ORF plasmid library for more than 5500 Drosophila genes conserved in humans. *Genome Research*, *30*(1), 95–106.

https://doi.org/10.1101/gr.250811.119
Weinzierl, R., Komfeld, K., Hogness, D., Connor, M. O., Binari, R., & Bender, W. (1987). Ultrabithorax mutations in constant and variable regions of the protein coding sequence. *Genes & Development*, *1*(4), 386–397. https://doi.org/10.1101/gad.1.4.386
Westholm, J O, Miura, P., Olson, S., Shenker, S., Joseph, B., Sanfilippo, P., Celniker, S. E., Graveley, B. R., & Lai, E. C. (2014). Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Reports*, *9*(5), 1966–1980. https://doi.org/10.1016/j.celrep.2014.10.062 [doi]
Westholm, Jakub O, Miura, P., Olson, S., Shenker, S., Joseph, B., Sanfilippo, P., Celniker, S. E., Graveley, B. R., & Lai, E. C. (2014). Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Reports*, *9*(5), 1966–1980. https://doi.org/10.1016/j.celrep.2014.10.062
White, R. L., & Hogness, D. S. (1977). R loop mapping of the 18S and 28S sequences in the long and short repeating units of Drosophila melanogaster rDNA. *Cell*, *10*(2), 177–192. https://doi.org/10.1016/0092-8674(77)90213-6
Wilkinson, M. E., Fica, S. M., Galej, W. P., Norman, C. M., Newman, A. J., & Nagai, K. (2017). Postcatalytic spliceosome structure reveals mechanism of 3'-splice site selection. *Science (New York, N.Y.)*, *358*(6368), 1283–1288. https://doi.org/10.1126/science.aar3729
Wu, J., & Manley, J. L. (1989). Mammalian pre-mRNA branch site selection by U2 snRNP involves base pairing. *Genes & Development*, *3*(10), 1553–1561. https://doi.org/10.1101/gad.3.10.1553
Wu, J. Y., & Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell*, *75*(6), 1061–1070. https://doi.org/10.1016/0092-8674(93)90316-i
Wu, N.-Y., Chung, C.-S., & Cheng, S.-C. (2017). Role of Cwc24 in the First Catalytic Step of Splicing and Fidelity of 5' Splice Site Selection. *Molecular and Cellular Biology*, *37*(6). https://doi.org/10.1128/MCB.00580-16
Yan, C., Wan, R., Bai, R., Huang, G., & Shi, Y. (2016). Structure of a yeast activated spliceosome at 3.5 Å resolution. *Science*, *353*(6302), 904–912. https://doi.org/10.1126/science.aag0291
Yan, C., Wan, R., Bai, R., Huang, G., & Shi, Y. (2017). Structure of a yeast step II catalytically activated spliceosome. *Science (New York, N.Y.)*, *355*(6321), 149–155. https://doi.org/10.1126/science.aak9979
Yu, C., Wan, K. H., Hammonds, A. S., Stapleton, M., Carlson, J. W., & Celniker, S. E. (2011). Development of expression-ready constructs for generation of proteomic libraries. *Methods in Molecular Biology (Clifton, N.J.)*, *723*, 257–272. https://doi.org/10.1007/978-1-61779-043-0_17
Yuryev, A., Patturajan, M., Litingtung, Y., Joshi, R. V, Gentile, C., Gebara, M., & Corden, J. L. (1996). The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(14), 6975–6980. https://doi.org/10.1073/pnas.93.14.6975
Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N. M., & Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell*, *152*(3), 453–466. https://doi.org/10.1016/j.cell.2012.12.023
Zhang, X.-O., Fu, Y., Mou, H., Xue, W., & Weng, Z. (2018). The temporal landscape of

recursive splicing during Pol II transcription elongation in human cells. *PLoS Genetics*, *14*(8), e1007579. https://doi.org/10.1371/journal.pgen.1007579
Zhang, X., Yan, C., Zhan, X., Li, L., Lei, J., & Shi, Y. (2018). Structure of the human activated spliceosome in three conformational states. *Cell Research*, *28*(3), 307–322. https://doi.org/10.1038/cr.2018.14
Zhuang, Y., & Weiner, A. M. (1986). A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell*, *46*(6), 827–835. https://doi.org/10.1016/0092-8674(86)90064-4