Regulation and Function of Non-Exonic Recursive Splicing

Michael Yen-Minn Chen

Thesis Advisor: Dr. A. Javier Lopez

submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

Department of Biological Sciences

Carnegie Mellon University

Pittsburgh 2011

Regulation and Function of Non-Exonic Recursive Splicing

Michael Yen-Minn Chen
Department of Biological Sciences
Carnegie Mellon University
Pittsburgh, PA

Abstract

Removal of intron sequences from pre-mRNA by splicing is an essential step in the expression of most human genes. Recursive splicing is a mechanistic variant shown to mediate the stepwise removal of very long introns and a subset of alternative splicing events in *Drosophila melanogaster*. Recursive splice sites (RSSs) consist of juxtaposed 3' and 5' splice site motifs that define a zero-nucleotide exon whose 3' and 5' splice sites are coincident. Only 10-20% of RSSs in Drosophila are involved in alternative splicing, so most RSSs are non-exonic. Consequently, their biological role is unclear despite their high conservation and preferential association with long introns. Although RSSs can be predicted in human genes as well, their function has not been demonstrated. In this work, I study three problems: (1) How cis-elements contribute to the proper use of a non-exonic RSS. (2) The biological role of a non-exonic RSS in its natural context. (3) Verification of recursive splicing in human genes and some of its functional consequences.

It is unclear how RSSs can function without interference between their 3' and 5'ss components. The consensus RSS motif suggests that these elements are biased to function initially as a 3'ss due to enhanced features of the 3'ss component. In addition, most non-exonic RSSs are (surprisingly) associated with silent downstream 5'ss motifs at a position where they would be expected to define an exon. This could also help activate the RSS as a 3'ss by interaction with the non-overlapped 5'ss ("pseudo-exon definition"), but some mechanism would have to redirect splicing subsequently to the regenerated 5'ss rather than the downstream site. Intriguingly, the silent downstream 5' splice sites are at a distance from the RSS where regular 5' splice sites exhibit a peak in the distribution of enhancers. Experimental dissection of an example associated with non-exonic RSS RP3 in the *Ultrabithorax* gene of Drosophila revealed that the downstream pseudo-5'ss is not required for activation of the RSS as a 3'ss. Instead, it functions as part of a conserved module that stimulates use of the 5'ss that is regenerated by the RSS. This prevents inappropriate use of competing alternative and cryptic sites. The result is consistent with the hypothesis that the enhanced 3'ss component and branch site are sufficient to ensure their activity, but that activation of the regenerated 5'ss requires assistance. The regulatory function of the pseudo-5'ss requires base-pairing with U1 snRNA, even though splicing does not occur at this position. Similar modules may assist the efficient and correct sequential activity of many RSSs. These arrangments may reflect a dynamic evolutionary history of interconversions between exonic and nonexonic RSSs.

To test the hypothesis that recursive splicing is important for the correct and/or efficient expression of genes with long introns, a two-step gene replacement strategy was used to delete the non-exonic RSS RP3 from within a 50 kb intron in the endogenous *Ultrabithorax* gene and to generate isogenic wild-type control chromosomes. A change in

the alternative isoform ratios was detected by semi-quantitative RT-PCR, and phenotypic analyses indicate that deletion of RP3 leads to a mild loss of function. Additionally, a *white* marker gene inserted near RP3 is profoundly silenced but can be reactivated by deletions extending upstream, suggesting a repressive chromatin structure in this region.

A sample of predicted RSSs from human was tested using the same approaches as used previosuly in Drosophila. These tests made use of a recursive splicing reporter system or minigenes transfected into human cell lines. For three out of eight human RSS candidates tested, it was possible to detect the predicted recursive intermediates and a shift to use of an alternative 5'ss after mutation of the RSS 5'ss motif. A RSS associated with a novel ORF-truncating cassette exon (E3b) in the human dopamine reuptake transporter gene *SLC6A3* was also validated. Alternative splicing of exon E3b was verified in endogenous transcripts in the *substantia nigra* of adult human brain and in reporter and minigene transcripts in a transfected neuronal cell line. E3b is flanked by single-nucleotide polymorphisms (SNPs) that appear to be associated with differential risk for schizophrenia. The risk-associated haplotype increases the inclusion of E3b in cell transfections assays and thus might be associated with reduced expression of dopamine reuptake transporter in vivo. An ORF-truncating exon E3b is present in all sequenced mammalian genomes except mouse, rat and rabbit, suggesting that recursive splicing of this cassette exon normally plays a role in regulating dopamine activity, and that genetically determined differences in regulation can underlie or exacerbate dopaminergic dysfunction.

Acknowledgements

I would first like to thank my advisor Dr. Javier Lopez for his infinite patience in dealing with my many faults, while training me to become a better scientist. I thank him for allowing me to research independently while still answering dumb questions from time to time and withstanding my terrible presentation skills.

Additionally, I would like to thank the members of my committee: Dr. John Woolford, Dr. Mark MacBeth, and Dr. Michael Palladino. They have always been there to support me and answer my questions during my time here, as well as making sure that I would be prepared for life after CMU.

I am grateful to the many former and current members of the Lopez lab for their help. Dr. James Burnette for his analysis of non-exonic RSSs; Dr. Panagiotis Papasaikas for helping me understand the computational facet of my projects; Steve Reilly and Sherry He for their help with the deletion of RP3; Kathleen McCann and Sandy Roh for her help with the mammalian recursive splicing project and for the heavy lifting of the DAT project; Rachel Ehrlich and Maria McDonald and Sherry He for continuing the RSS deletion project; Jason Talkish, Anmol Grover, and Stephanie Hughes for allowing me to use them during their rotations for my own projects. Thanks to Sladjana Stratimirovic for showing me the ropes when I first joined the lab and Andrea Zonneveld for keeping the lab stocked. Thanks to Yevgeniya Monisova and Debbie Makin for their conversations and ideas.

I would like to thank a number of labs from the department, including the Woolford, MacBeth, Minden, Jarvik, Berget, Linstedt, Lee, McCartney, and Armitage Labs for their help with various reagents, equipment, and advice.

I would like to thank my parents for helping me in my pursuit for a PhD, if it wasn't for them, I would not be at CMU.

Finally, I would like to thank my wife Dianne for everything.

Table of Contents

Index of Figures and Tables

Chapter **Page**

3. Analysis of Biological Function of Non-Exonic Recursive Splice Sites

4. Recursive Splicing in Mammalian Introns

Chapter 1: General Introduction

In eukaryotes, genes commonly contain two operationally defined types of sequences called exons and introns. Both exons and introns are transcribed into pre-mRNA, but the introns are removed and the exons are joined together in a process called splicing. Most introns are removed by the spliceosomal pathway, in which a complex cellular machinery composed of ribonucleoprotein assemblies recognizes the boundaries between introns and exons and removes the intron. A mechanism called alternative splicing can increase the diversity of mRNA and protein products made from a given gene by optionally including or excluding specific exons or portions of exons. Alternative splicing can function as a key developmental switch, as in the Drosophila sex determination pathway (review Salz 2011), or fine-tune gene function as in the Drosophila Hox gene *Ultrabithorax* (Subramaniam et al 1994, Reed et al 2010). Alternative splicing occurs in all cells of the human body, but is most prevalent in the brain (Stamm et al 2000, Xu et al 2002). Additionally over 94% of human genes undergo alternative splicing (Wang et al 2008). Alternative splicing utilizes *trans*-acting factors that bind to *cis*-acting elements to influence the activity of the spliceosome at specific splice sites. Mutations that affect constitutive or alternative splicing in *cis* or *trans* can lead to genetic disease (Faustino and Cooper 2003) and have been implicated in the development of tumors (Venables 2004).

A particular mechanism for alternative splicing involves the reutilization of a spliced junction to remove an already spliced exon as an intron (Hatton et al 1995). This mechanism, known as recursive splicing, was discovered in *Drosophila melanogaster*

and appears to be common at least in Diptera and other insects (Burnette et al 2005, Conklin et al 2005, Papasaikas et al 2010). Recursive splicing allows the removal of an intron as a series of subfragments. In Drosophila and other Diptera, recursive splicing is associated specifically with long introns (>5 kb). Most recursive splice sites (RSSs) appear to be non-exonic, i.e. they subdivide an intron without defining a detectable alternatively spliced exon. Nevertheless such non-exonic recursive splice sites are highly conserved, implying that they serve an important but currently unknown function distinct from alternative splicing. Furthermore, given current knowledge of mechanisms for splice site recognition and activation, it is unclear how a non-exonic recursive splice site can function. Although recursive splicing was first proposed in mammals (Mineo et al 1990), it has not been demonstrated conclusively in that group of organisms. This thesis addresses aspects of all three problems. Before presenting my work, I will review relevant aspects of the mechanism of intron removal by the major spliceosome pathway, as well as our current understanding of recursive splicing,

The Basic Splicing Mechanism

Removal of introns is required for the expression of most protein–coding genes and many non-coding RNAs in multicellular eukaryotes. There are two major groups of introns, self-splicing introns and spliceosomal introns. Self-splicing introns include group I and group II introns. Removal of Group I introns is initiated by a nucleoside or nucleotide that is non-covalently bound by the folded intron; these introns are mostly found in algae, lichen, fungi, and some bacteria (Haugen et al 2005). Group II introns use an adenosine that is part of the intron itself; these introns are found in bacteria and

in the organelles of various eukaryotic organisms (reviewed by Federova and Zingler 2007). Spliceosomal introns are removed by the same chemistry as group II introns, but this is catalyzed by the spliceosome, a large and dynamic complex assembled through interactions between the mRNA and five small nuclear ribonucleoprotein subcomplexes (snRNPs) (Jurica and Moore 2003). The spliceosome catalyzes two sequential transesterification reactions that excise the intron and join the flanking exons (reviewed in Wahl 2009) (Figure 1.1). First, the 2'-OH of a specific nucleotide within the intron (called the branchpoint nucleotide) attacks the phosphodiester bond at the 5' end of the intron (called the 5' splice site). This frees the 3'-OH of the upstream exon, which then attacks the phosphodiester bond at the 3' end of the intron (called the 3' splice site). This second step joins the exons via a 3'-5' phosphodiester bond and releases the intron as a lariat structure with a free 3'-OH and a 2'-5' phosphodiester bond at the branchpoint.

Cells can have two spliceosome variants, called major and minor because of their difference in abundance and the proportion of introns that they act on. The minor spliceosome only splices introns in 700 to 800 genes in the human genome. Many of the minor spliceosomal components are analogs of the major spliceosome, but recognize a different consensus for the splicing signals on mRNA (reviewed in Pess and Frilander 2011, Will and Luhrmann 2005). The core splicing signals for the major spliceosomal pathway consist of sequences surrounding the splice sites and branch point sequence (Figure 1.2, top panel) (reviewed in Wang and Burge 2008). The 5' splice site (5'ss) consensus is (AG)GURAGU and the 3' splice site (3'ss) consensus is NYAG(G), which is preceded by a poly-pyrimidine tract of 10-15 nucleotides (the

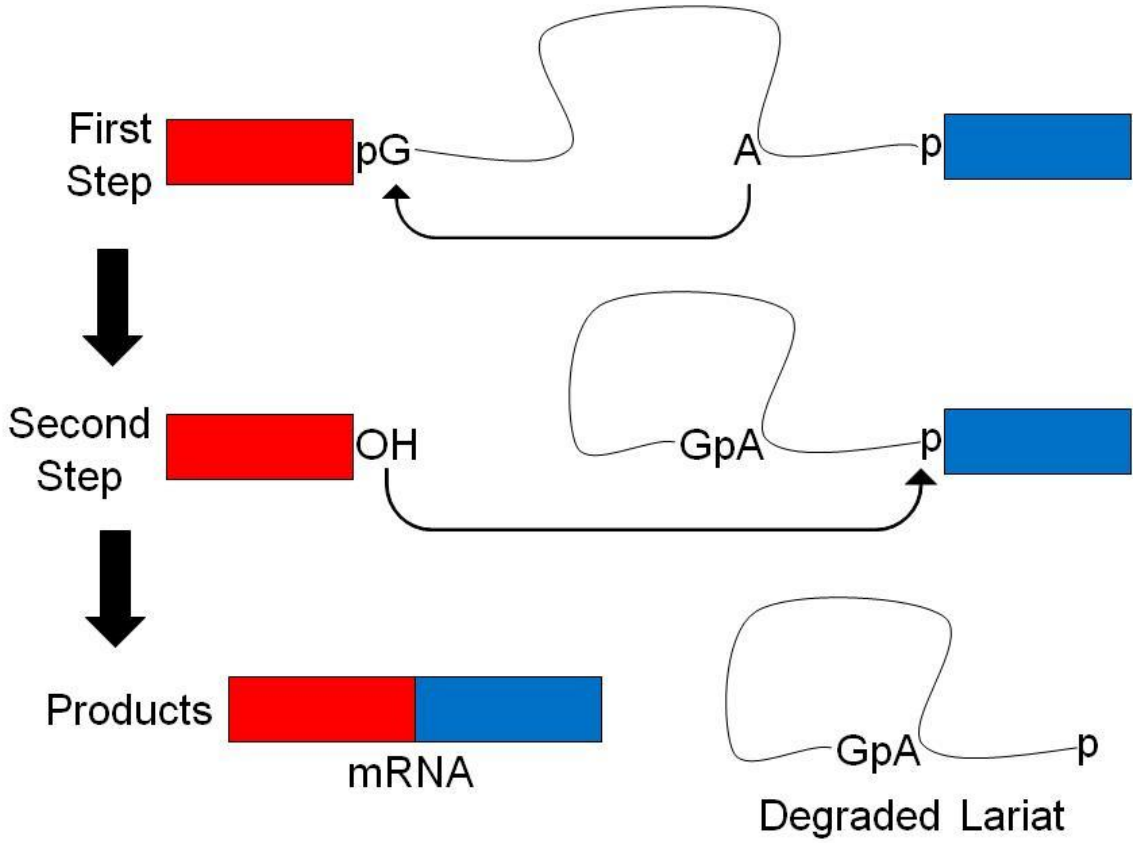


Figure 1.1. Diagram of the two-step biochemistry of pre-mRNA splicing. Phosphates are represented by 'p', nucleotides by capital letters, and transesterification reactions by arrows showing the direction of attack.

nucleotides in parentheses are located within the exons). The branchpoint nucleotide is located 15-40 nucleotides upstream of the 3'ss and is defined by the consensus sequence YNYUR**A**Y (the bolded adenosine is the branch point).

Assembly of the spliceosome (Figure 1.2, bottom panel) begins when the U1 snRNP binds to the 5'ss, with U2AF and SF1 following shortly to cooperatively bind the 3'ss and branch point respectively, to form the E complex. U2 snRNP then displaces SF1 and binds to the branch point sequence, forming the A complex. U4, U5, and U6 snRNPs are pre-assembled as a tri-snRNP before binding to the mRNA, U1, and U2 to form the B complex. The U1 and U4 snRNPs are destabilized or released, the B complex is activated to catalyze the first transesterification step and rearrangements occur to form the C complex. The C complex catalyzes the second transesterification step, which ligates the exons together. U2, U5, and U6 are release and recycled, and the free intron lariat product is linearized by lariat debranching enzyme (DBR) and subsequently degraded by nucleases (Ooi et al 2001).

Auxiliary Elements

The core splicing signals by themselves cannot define true exon-intron boundaries in organisms with high genomic complexity. In budding yeast the majority of introns are 300 nt or less, and introns shorter than 150 nt use intron definition to facilitate their removal (Berget 1995). In this mechanism the early-binding spliceosome assembly factors (U1 snRNP, U2AF, SF1) interact between the splice sites across the intron. In multicellular eukaryotes, expanded intron sizes require a different mechanism in order to recognize authentic splice site signals and to pair splice sites correctly across longer

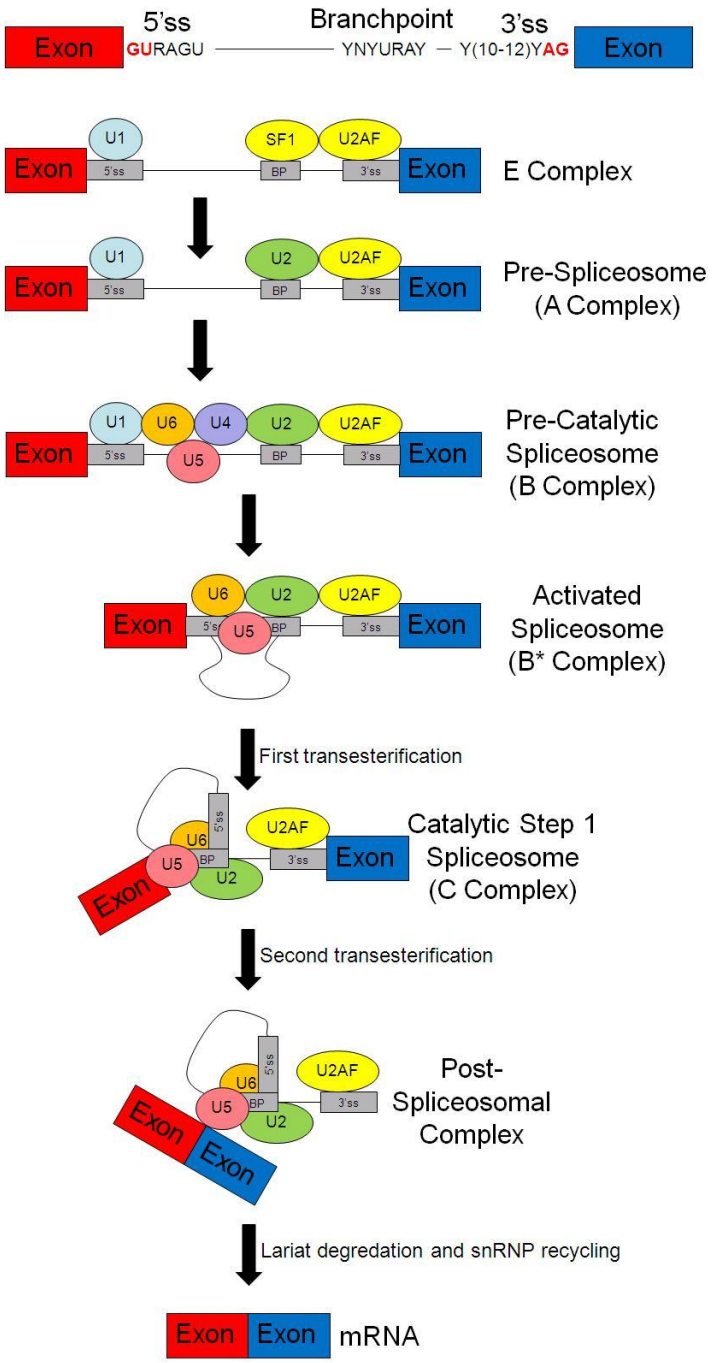


Figure 1.2. Sequential assembly of spliceosome during pre-mRNA splicing. The consensus sequences of metazoans are at the top. Below that is the step-wise recruitment of major snRNPs during removal of an intron between the red and blue exons.

distances. The exon definition model invokes stabilizing interactions between components of the splicing machinery across the exon instead of the intron (Figure 1.3) (reviewed by Berget 1995). Initially during transcription, the splicing machinery searches for closely spaced splice sites. As exons generally are bounded by a 3'ss and 5'ss in close proximity (whereas they are far apart for typical introns), the splicing machinery first defines an exon by the binding of U1 and U2 snRNPs at its ends. The stable binding of U1 and U2 snRNPs is facilitated by SR proteins (Serine- and aRginine-rich proteins) and hnRNPs (heterologous nuclear ribonucleoparticles (Jurica and Moore 2003). SR proteins contain one or more RRM domains for binding RNA, as well as serine-arginine repeats (called RS domains) for binding to other SR proteins and splicing factors such as U2AF (Long and Caceres 2009). The hnRNP protein family also associates with the splicing machinery to regulate the interactions of splicing machinery to correct splice sites. An example is the hnRNP protein PTBP (polypyrimidine-tract binding protein), which plays multiple roles in splicing such as inhibiting exon definition by binding to the polypyrimidine-tract of 3' splice sites, as well as preventing intronic cross talk of U1 and U2 by binding to polypyrimidine-tracts located in introns between U1 and U2 (reviewed by Schellenberg et al 2008).

The trans-acting factors such as SR and hnRNP proteins bind to *cis*-elements on either the intron or exon (reviewed by Wang and Burge 2008) (Figure 1.4). Exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs) are located in exons and promote inclusion or exclusion of the exon, respectively. Conversely, intronic splicing enhancers (ISEs) and intronic splicing silencers (ISSs) are located in introns and function in a similar manner as their exonic counterparts. ESEs and ESSs have

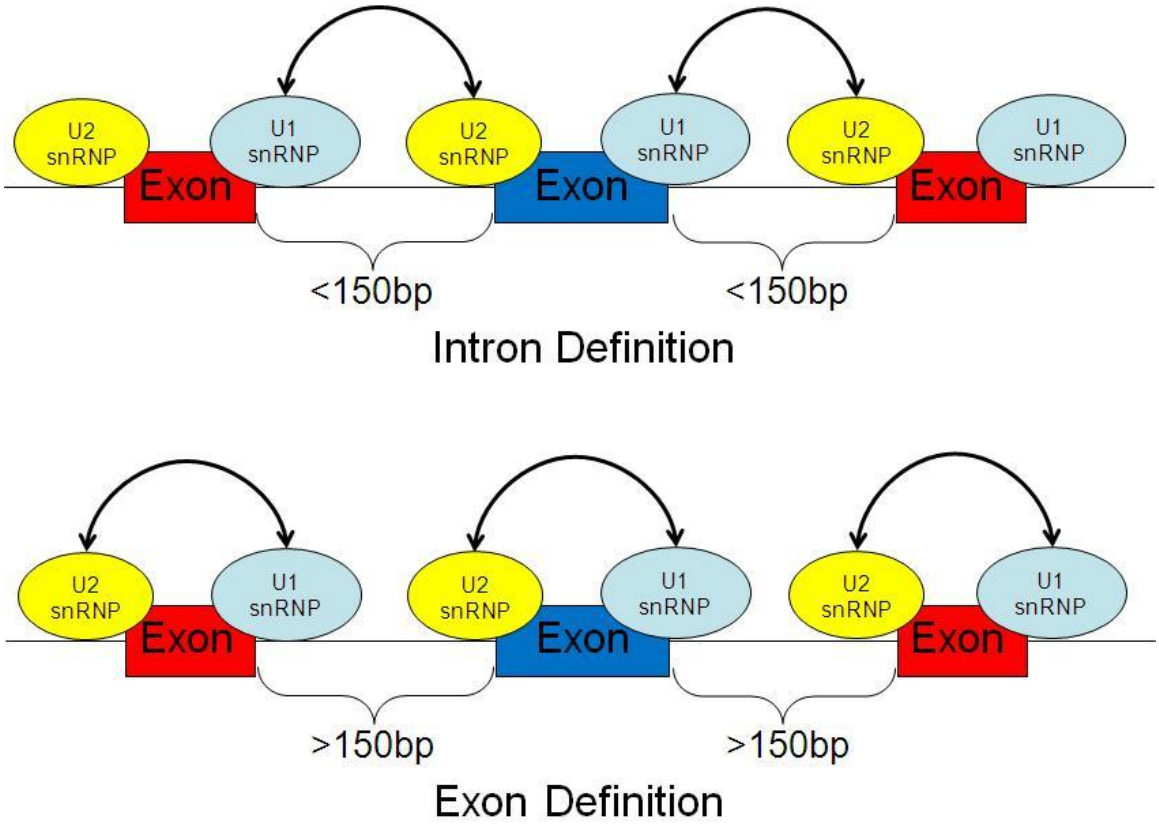


Figure 1.3. Intron and exon definition in pre-mRNAs. U1 and U2 snRNPs interact across the intron when the intron is less than 150 bp. Larger introns force the snRNPs to interact across the exons to license the preceding intron for splicing.

been better studied, with a combination of mutational approaches, high-throughput screens, and computational methods used to identify a large number of exonic elements (Chasin 2007). There have been fewer approaches to screen for intronic elements, with only specific elements identified for specific *trans*-acting factors. Splicing enhancers and silencers generally function in multiple copies and additively promote the use of a particular splice site (Matlin et al 2005), either by increasing the affinity for binding trans-acting factors (Dominguez and Allain 2006), or by increasing the number of factors in the local area. Additionally, splicing enhancers and silencers are highly contextual. These elements may be location dependent, either strengthening or weakening splicing based on distance to the splice site, or switching from enhancer to silencer based on location in either exons or introns, or changing activity based on orientation and location in different genes. The complexity of recognizing 3' and 5' splice sites and defining exons allows primary transcripts to undergo alternative splicing.

Alternative Splicing

Alternative splicing selectively uses competing 5' and/or 3' splice sites to form different mRNA structures from the same gene (reviewed by Black 2003). The alternative splicing of mRNAs can produce proteins with altered/different functions, or it can be used for quantitative regulation of protein expression. Alternative splicing events can generally be categorized into a few major forms (Figure 1.5): **alternative 5'ss use**, in which competing 5'ss are selected in order to change the 3' boundary of the upstream exon; **alternative 3'ss use**, in which competing 3'ss are selected in order to change the 5' boundary of the downstream exon; **exon skipping**, in which an exon may be left out

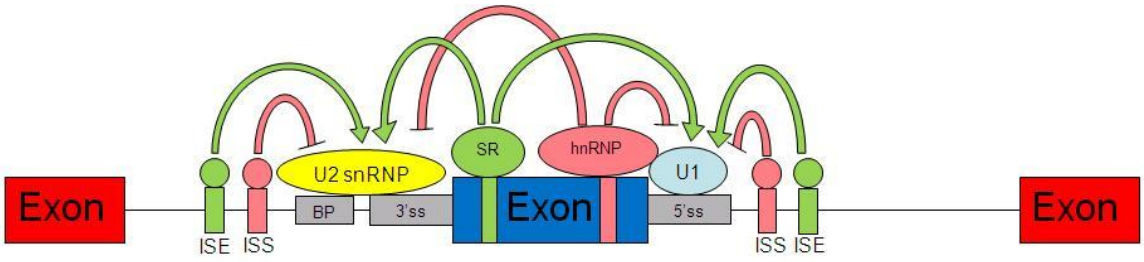


Figure 1.4. Schematic of exonic and intronic cis-regulatory elements. Splicing is regulated by combinatorial enhancement or suppression through cis-elements (ISE, ISS, ESE, ESS) and trans-acting factors (SR proteins, hnRNPs).

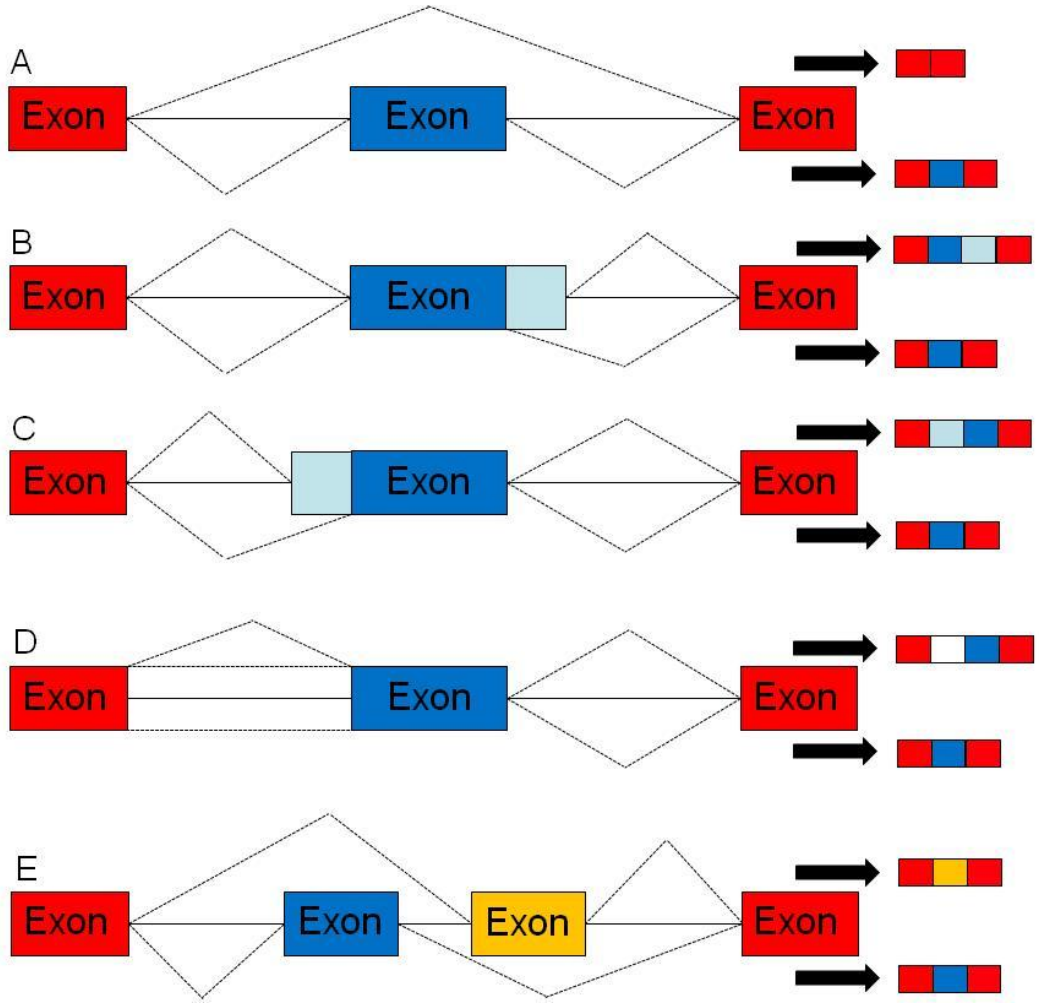


Figure 1.5. Major patterns of alternative splicing. The black arrows point to the sequence of ligated exons from alternative splicing using either the top pathway or the bottom pathway. (A) Exon skipping. (B) Alternative 5'ss usage. (C) Alternative 3'ss usage. (D) Intron inclusion. (E) Mutually exclusive exons.

of the mRNA by failure to recognize its splice sites, by active suppression of splicing, or through removal by recursive splicing; **intron retention**, in which the flanking 5'ss and 3'ss are not activated, leaving the intron in the mature mRNA; and **mutually exclusive exons**, in which the splicing of one exon prevents the splicing of another exon. The Drosophila Down Syndrome Cell Adhesion Molecule (DSCAM) is a notorious example of the diversity that can be produced by combinatorial alternative splicing (reviewed by Schumucker and Chen 2009), with potentially over 38000 mRNA isoforms. Drosophila DSCAM is composed of 115 exons, 95 of which are alternatively spliced as members of four exon clusters. Exon clusters 4, 6 and 9 contain 12, 48, and 33 mutually exclusive exons respectively and comprise the extracellular immunoglobulin domain region of the protein, whereas cluster 17 contains two mutually exclusive exons that affect the transmembrane domain.

Alternative splicing can be regulated by cues from cell type, developmental stage, gender, and external or internal signals (Faustino and Cooper 2003). Despite intensive study, the mechanisms that regulate alternative splicing are not well understood. In general, splice site selection is influenced through the action of cis-elements and *trans*-acting factors such as SR proteins, hnRNPs, and other proteins that suppress or enhance the use of alternate splice sites in particular cell types or under different physiological conditions (reviewed in: Pozzoli and Sironi 2005, Smith and Valcarcel 2000, Jurica and Moore 2003, Black 2003). These factors act by influencing the binding of U1 snRNP, U2 auxiliary factor (U2AF), and U2 snRNP early in

spliceosome assembly (reviewed in: Pozzoli and Sironi 2005, Smith and Valcarcel 2000, Jurica and Moore 2003, Black 2003).

It has been estimated that 94% of human genes are alternatively spliced (Wang and Burge 2008). The accurate splicing of these genes is essential to not just the health of the cell, but also the organism (reviewed by Cooper et al 2010). Mutations can directly affect the splicing of a human gene, causing a change in its alternative splicing or creating a novel splice junction. For example, mutations in the eighth intron of *CFTR*, the gene mutated in Cystic Fibrosis, lead to skipping of exon nine, increasing the severity of Cystic Fibrosis (Buratti et al 2007). Another example involves deletions encompassing the 3'ss of intron 10 in the receptor tyrosine kinase KIT, which creates an aberrant 3'ss within exon 11. This leads to an in-frame loss of 27 nucleotides and abolishes auto-inhibition, leading to constitutive kinase activity that may play a key role in the formation of gastrointestinal tumors (Chen et al 2005).

Cotranscriptional Splicing

For sufficiently long genes, alternative splicing and constitutive splicing are generally thought to occur during transcription (reviewed in Kornbhitt et al 2004, Pandya-Jones and Black 2009). Once transcription synthesizes the 5' and 3' splice sites, splicing can take place within 5-10 minutes (Singh and Padgett 2009). Splicing factors can associate with the C-terminal domain (CTD) of RNA polymerase II (reviewed by Bentley 2007). Examples include spliceosomal components (Das et al 2006) and SR proteins (Das et al 2006). Splicing-associated factors can stimulate transcriptional elongation (Fong and Zhou 2001, Lin et al 2009, Lenasi and Barboric 2010, Alexander et al 2010).

Conversely, transcription can affect splicing through the rate of elongation, with slower polymerases favoring alternative exon inclusion (de la Mata et al 2003, de la Mata 2010, reviewed in Kornbhlitt et al 2004). As genes increase in length with organism complexity, transcription may need to be coupled to splicing and other mRNA processing pathways for efficient gene expression.

Recursive Splicing

Recursive splicing involves coincident 3' and 5' splice sites that essentially define zero-nucleotide exons (Figure 1.6A; Hatton et al 1998, Burnette et al 2005). The recursive splice site (RSS) functions first as a 3' splice site and regenerates a 5' splice site that can be used for subsequent splicing to a downstream 3' splice site (Figure 1.6A). Recursive splicing was first demonstrated as the alternative splicing mechanism for exclusion of cassette exons within the *Ultrabithorax* (*Ubx*) gene in *Drosophila melanogaster* (Hatton et al 1998). Several examples of RSSs associated with cassette exons in this way have been identified and verified (dashed rectangles in Fig 2; Hatton et al, Burnette et al 2005, Conklin et al 2005, Papasaikas et al 2010). However, it currently appears that most RSSs (~90%) are non-exonic; that is, they are not associated with any annotated or detectable alternatively spliced exons. Computational approaches have predicted hundreds of RSSs in Drosophila (Burnette et al 2005; Papasaikas et al 2010), and initial attempts in mammals yield similar results (Papasaikas and Lopez, personal communication). In Drosophila, 16 different RSSs in 10 genes have been experimentally verified to undergo recursive splicing (Burnette et al 2005; Conklin et al 2005; Papasaikas et al 2010).

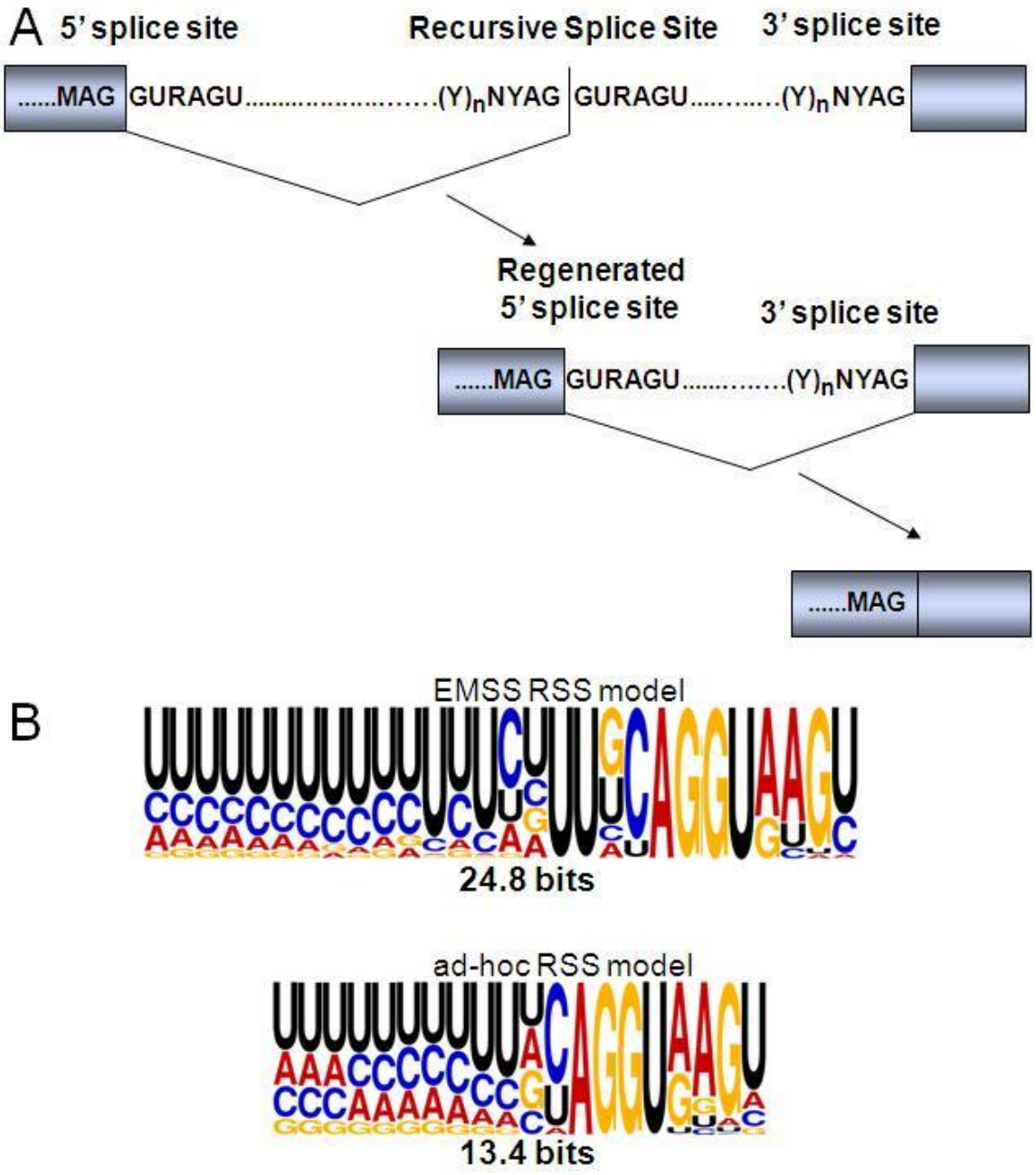


Figure 1.6. The recursive splicing mechanism. (A) The first step pairs the upstream 5'ss and the 3'ss component of the RSS for removal of the intron. The 5'ss regenerates, and is able to pair with a downstream 3'ss for the second step of intron removal. (B) Comparison between the EMSS and ad-hoc RSS models. (from Papasaikas et al, 2010)

Initially, an ad-hoc model was used to predict recursive splice sites in Drosophila introns. This model was created by juxtaposing the known position-specific nucleotide preference matrix (PSSM) for intronic positions of regular 3' splice sites with the known PSSM for intronic positions of regular 5' splice sites (Burnette et al 2005). A better model (known as the EMSS RSS model) was developed subsequently by semi-supervised machine learning. For this purpose, a small number of experimentally verified sites was used to seed a reiterative model-building algorithm that operated on the entire Drosophila genome sequence (Figure 1.6B). This refined model exhibited a longer RSS motif than the ad-hoc model and had nearly twice the information content (24.8 bits; Papasaikas et al 2010). In particular, the EMSS model extends the length of and tightens the nucleotide preferences of the 3'ss motif, and it shows a preference for a C nucleotide at position -8 with regards to the 3'ss/5'ss position. At the same time, the nucleotide preferences at the 5'ss motif of the EMSS model conform more tightly to the standard consensus sequence.

Analysis of the distribution of RSSs originally predicted by the ad-hoc model showed that they are found preferentially within introns longer than 10 kb (Figure 1.7). The observed frequency of RSSs is in large excess over the random expectation derived by Monte-Carlo simulations and by direct calculation (Burnette et al 2005). The refined model predicted an even greater number of RSSs, and these were also distributed preferentially in large introns (Figure 1.7; Papasaikas et al 2010). Every Drosophila intron larger than 50 kb has at least one predicted RSS, with an average of 1 RSS per 25 kb.

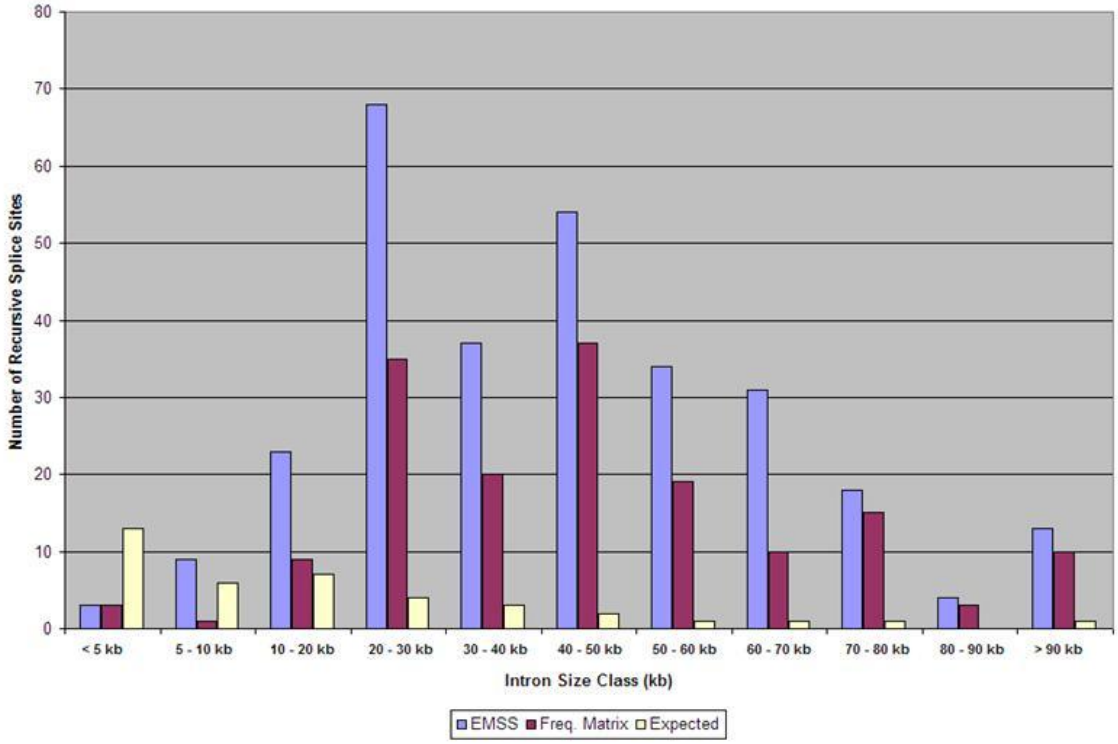


Figure 1.7. Distribution of non-exonic recursive splice sites by intron size class. The observed number of predictions using the EMSS model (blue bar/left bar) versus the ad-hoc model (red bar/middle bar) and the expected number of predictions (yellow bar/right bar) (from Papasaikas et al 2010).

Phylogenetic analysis involving 12 species of the genus Drosophila with completely sequenced genomes revealed that predicted RSSs are highly conserved (Figure 1.8). Over 90% of sites are conserved between species that diverged 40 Myr ago, and the relative positions of individual RSSs within introns are also highly conserved across these Drosophila species (Fig. 1.8). Additionally, 82% of RSSs from *D. melanogaster* can also be found in the mosquito Anopheles. Furthermore, the RSSs represent very strong peaks of local sequence conservation within their host introns (Figure 1.9); Papasaikas et al 2010). The strongest conservation centers in the core of the RSS (the AG|GT), with conservation decreasing with distance away from the core; this is consistent with the known strength of functional constraint on nucleotide preferences at different positions within 3' and 5' splice sites. A second, smaller peak of conservation is centered at 33 nt upstream of the 3'/5'ss and corresponds to the branch site sequence; this is remarkable, given that the branch site consensus is generally very loose for standard splice sites.

The enrichment of RSSs in introns larger than ~10kb, their excess over random expectation, and their conservation during evolution indicate that RSSs have an adaptive role in the context of large introns. That this role is specific to long introns is also supported by the observation that deleting a non-exonic recursive splice site from a minigene construct has no effect on splicing or expression (Burnette et al 2005). However, this role is currently unknown. Possible hypotheses are discussed in Chapter 3.

Interestingly, a large number of predicted RSSs (33%) in the Drosophila genome are located within non-LTR retroelements. The majority are accounted for by telomere-

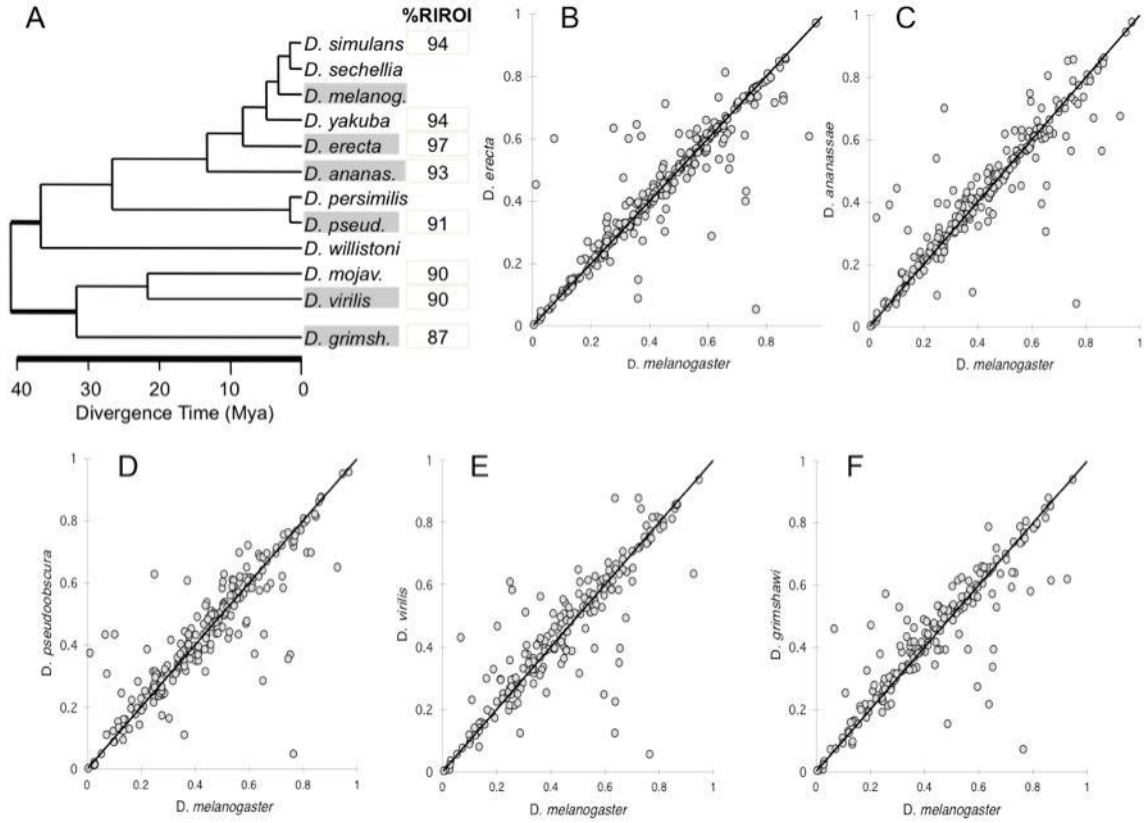


Figure 1.8. Phylogenetic analysis of RSSs among Drosophilids. (A) Cladogram of 12 Drosophilids used in the phylogenetic analysis. %RIROI is the percent RSS identification rates in orthologous introns. (B-F) The relative positions of RSSs for 5 target species (shaded in panel B) covering the complete divergence period are plotted against the corresponding positions in *D. melanogaster*. Points lying along the diagonal correspond to RSSs found at the same relative position within the same intron of *D. melanogaster* and the target species. (from Papasaikas et al 2010)

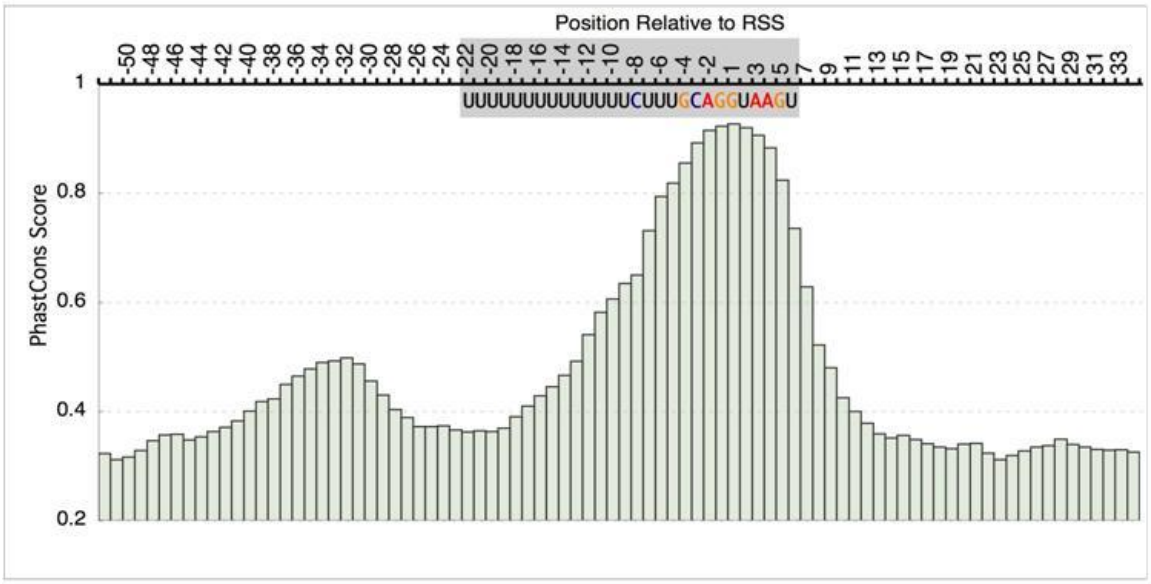


Figure 1.9. RSSs represent local peaks of conservation. Average PhasCon scores among 15 insect species plotted along each nucleotide position centered around the RSS. Position -1 is the last nt of the 3'ss component and +1 is the first nt of the 5'ss component (from Papasaikas et al 2010).

associated Het-A and TAHRE telomeric retroelements (Papasaikas et al 2010). These RSSs are located on the antisense strand of the retroelement, which is transcribed but has unknown function except in repeat-associated siRNA-mediated regulation of retroelement transposition and telomere elongation (Shpiz et al 2009, Savitsky et al 2006). The rasiRNA pathway is distinct from the canonical siRNA pathway by not requiring Dicer, a key enzyme in siRNA and miRNA processing (Vagin et al 2006, Savitsky et al 2006). The Het-A and TAHRE elements are found in long head-to-tail arrays at telomeres. Transcription can extend across multiple elements in an array. Thus, it is possible that the RSSs mediate recursive splicing of the multi-element transcripts and could modulate accumulation of the antisense strand and thus production of rasiRNA (Papasaikas et al. submitted). Alternatively, these RSSs may promote interactions between splicing and transcription machinery that facilitate transcription within the telomeric chromatin. Intriguingly, the independently-evolved telomere-associated non-LTR retroelements of Lepidoptera are one of the few other elements that also contain RSSs.

The presence of RSSs in retroelements that took over the telomeres early in dipteran evolution also suggests an explanation for the origin of RSSs in large introns of these insects (Papasaikas et al 2010). During initial expansion of the retroelement population, insertion of such an element into an intron of an active gene would simultaneously increase that intron by the size of the retroelement (generally 6-14 kb) and install a RSS. RSSs installed in this manner that conferred a selective advantage would remain conserved even as the rest of the retroelement degraded with time.

Chapter 2: The landscape of auxiliary elements around non-exonic Recursive Splice Sites: A U1snRNA-dependent module promotes efficient use of a regenerated 5' splice site.

ABSTRACT

Recursive splicing at non-exonic elements mediates the stepwise removal of very long introns in Drosophila and probably other metazoans. Recursive splice sites (RSSs) consist of juxtaposed 3' and 5' splice site motifs such that the actual splice sites are coincident and define a zero-nucleotide exon. The potential conflict between these 3' and 5'ss components, which must function sequentially, and the long intronic distances spanned by recursive splicing pose special challenges. Recent work revealed that Drosophila RSS motifs have higher information content than regular splice sites due to enhanced 3'ss components with distinctive features that are likely adaptations to these constraints. We have now analyzed the landscape of candidate *cis*-acting elements around non-exonic RSSs. As a consequence of their inherently stronger 3'ss components, non-exonic RSSs may depend less on the types of intronic splicing enhancers that are associated with regular splice sites. Surprisingly, most non-exonic RSSs are associated with downstream 5'ss motifs at a position where they would be expected to define an exon, but where regular 5' splice sites exhibit a peak of enhancers. Experimental dissection of an example associated with RSS RP3 in *Ultrabithorax* revealed that the downstream pseudo-5'ss does not normally define an exon and instead functions as part of a conserved enhancer module that

stimulates use of the 5’ss regenerated by the RSS. This prevents inappropriate use of competing alternative and cryptic sites. Similar modules may assist the efficient and correct sequential activity of many RSSs. This may reflect a dynamic evolutionary history for exonic and nonexonic RSSs.

Note: this chapter corresponds to a manuscript submitted for publication (authors: Michael Chen, Panagiotis Papasaikas and A. Javier Lopez) and includes computational analyses performed by fellow graduate student Panagiotis Papasaikas. These are included in full detail because they are important for understanding and interpreting the experimental studies, which I performed. The contributions by P. Papasaikas are identified in the text.

INTRODUCTION

Most protein-coding genes in multicellular organisms are interrupted by introns that must be removed from primary transcripts by pre-mRNA splicing. This provides the opportunity to regulate gene expression and expand the proteome by alternative splicing, which can define different starting and ending points for individual exons or exclude them altogether from mRNAs (reviewed by Stamm et al. 2005, Tazi et al. 2009). Several classes of sequence motifs in the pre-mRNA help to orchestrate constitutive and alternative splicing. The earliest recognized signals were the 5' splice site ("donor") and 3' splice site ("acceptor") motifs that mark the exon/intron and intron/exon boundaries, respectively, as well as the branch site motif, which is located near the 3' end of the intron and contains the adenosine involved in the first trans-esterification reaction of splicing (reviewed by Burge et al 1999). The 3' and 5' splice site motifs are normally separated by exons (between a 3' and a 5' splice site) or introns (between a 5' and a 3' splice site). In some functional splicing elements, however, a 3' and a 5' splice site are coincident, defining a zero-nucleotide exon. Depending on the context and the strength of the corresponding motifs, such elements can function alternatively as competing 3' or 5' splice sites ("dual specificity splice sites") (Zhang et al 2007) or sequentially as 3' and 5' splice sites ("recursive splicing"; Figure 2.1) (Hatton et al 1998, Burnette et al 2005).

Recursive splicing was described originally as a mechanism that excludes cassette exons from mRNA not by skipping them but by reutilizing an exon-exon

junction as a 5’ splice site (Mineo et al 1990, Hatton et al 1998). Bioinformatic and phylogenetic analyses using a hypothetical model based on standard 3’ and 5’ splice site motifs showed that recursively spliced cassette exons are flanked by large introns, and they suggested that a much larger population of apparently non-exonic recursive splice sites (Figure 2.1) frequently mediate the stepwise removal of very large introns in Drosophila and other insect species (Burnette et al 2005, Shepard et al 2009). A preferential and conserved association of RSSs with very long introns has been confirmed with a more accurate model (“EMSS-RSS”; Figure 2.1) that was generated by semi-supervised learning and validated experimentally (Papasaikas et al 2010). The EMSS-RSS model has higher information content and distinctive features that are consistent with a specialized role of recursive splice sites in long introns. In addition, its features help explain how these elements can function effectively despite the coincidence of the 3’ and 5’ splice sites, which is expected to create a conflict between binding of recognition factors for each of the associated motifs (Papasaikas et al 2010)

Given the complexity of recursive splicing, the efficient recognition and proper sequential function of RSSs may also be expected to depend on additional *cis*-acting auxiliary elements. In complex genomes the efficiency and fidelity of regular splicing depend not only on the strength and arrangement of the splice site motifs and branch site but also on flanking elements such as splicing enhancers and silencers that can be located within exons and introns (reviewed by Wang and Burge 2008). Enhancers stimulate use of correct splice sites, and

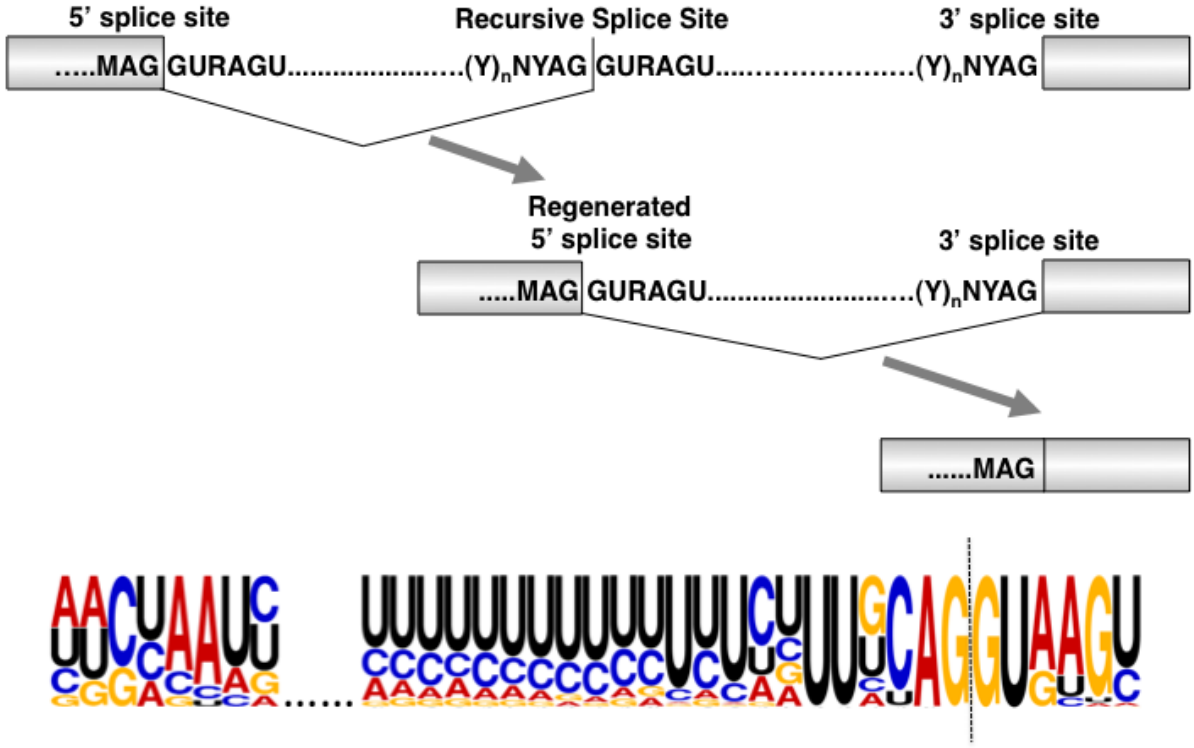


Figure 2.1. Recursive splicing. Top: The diagram shows the removal of an intron in two steps by use of a non-exonic recursive splice site. The vertical line marks the coincident 3' and 5' splice sites within the RSS motif. In some cases a downstream 5'ss defines a cassette exon whose exclusion is mediated by recursive splicing. Bottom: The sequence logos represent the recursive splice site motif generated by semi-supervised learning ["EMSS-RSS" model] and the associated branch site model, located approximately 32 nt upstream of the recursive splice site (vertical line) (Papasaikas et al 2010).

silencers prevent use of cryptic splice sites or pseudo-exons, which can be found frequently in random intron squences (Fairbrother and Chasin 2000, Sironi et al 2004, Zhang and Chasin 2004). Enhancers and silencers also control the activity of alternative splice sites (reviewed by Black 2003, Wang and Burge 2008).

Here we use a combination of bioinformatic, phylogenetic and experimental approaches to investigate the involvement of auxiliary sequences in the action of non-exonic recursive splice sites. First, we use the much larger set of validated regular splice sites to identify candidate intronic enhancer and silencer motifs in Drosophila, and we compare their distribution around RSSs with those around regular constitutive and alternative splice sites. In all three cases, enhancers are overrepresented and silencers are underrepresented, although the biases are less pronounced around RSSs. This may reflect a reduced dependence on standard enhancers due to the higher information content in the 3'ss component of RSSs and the need to avoid premature activation of the 5'ss component.

Despite the absence of documented exons, a prominent peak of predicted 5'ss motifs is found at ~50 nt downstream of the non-exonic RSSs, nestled between enhancer peaks. Detailed mutational analysis of these features downstream of RSS RP3 from the *Ultrabithorax* gene reveals that the pseudo-5'ss motif is not required to activate the RSS as a 3'ss, as might have been the case if it defined a pseudo-exon or a rare cassette exon. Instead it acts as part of an enhancer module to promote use of the regenerated 5' splice site and prevent use of competing cryptic and alternative 5' splice sites. This function

requires base-pairing interaction with U1 snRNA. We propose that similar modules function downstream of many non-exonic RSSs, and we discuss evolutionary implications of this arrangement.

METHODS AND MATERIALS

Prediction of Intronic Auxiliary Elements

(This computational work was performed by graduate student Panagiotis Papasaikas).

Several computational methods have been developed for the prediction of *cis*-acting elements that can control constitutive or alternative splicing. Typically these methods involve hypothesis-testing on nucleotide words (*k*-mers) in mRNA regions that are expected to be involved in splicing regulation. They assume that the distribution of a test statistic for *k*-mers involved in splicing regulation will be significantly different within these regions when compared to the background distribution that is used to formulate the null hypothesis. The test statistic is usually based on the frequency and/or cross-species conservation rates of the *k*-mers. Different mRNA regions can be used to formulate and evaluate the null hypothesis, depending on the type of element that is to be identified (e.g. exonic or intronic, silencers or enhancers) and the application (Fairbrother et al 2002, Sugnet et al 2006, Zhang and Chasin 2004, Zhang et al 2005, Yeo et al 2007, Voelker and Berglund 2007). We used a strategy that combines multiple attributes in order to identify sequences that function as part of intronic auxiliary elements and to predict their role as either silencers or enhancers. First we formulated a set of plausible attributes for intronic splicing silencers and enhancers in the genome. Each attribute can be translated into separate statistical variables for nucleotide words of size *k*. For this analysis, we

determined that *k*=5 was the maximum that could be used while being able to obtain accurate estimates for each of the statistical tests described below. Next, these variables were estimated for all possible 4^5=1024 pentanucleotides. Finally, these variables were combined in a statistically principled way to identify putative *cis*-regulatory intronic elements among the nucleotide pentamers. Only sequences around constitutive internal exons (according to Flybase 4.15 annotation [FlyBase Consortium, 2003]) and flanked by introns longer than 300bp were used in this analysis unless stated otherwise. The set of attributes for intronic enhancers and silencers is listed below:

1. Frequency in splice site proximal vs. splice site distal intronic regions: Intronic splicing enhancers are expected to be overrepresented in the regions immediately downstream of 5' splice sites or upstream of 3' splice sites as compared to intronic regions far from intron-exon boundaries. Conversely, intronic splicing silencers are expected to be underrepresented in this comparison. The splice-site proximal intronic frequencies of pentamers for this as well as for all subsequent tests were estimated on the 150 nt immediately upstream of the 3’ss and 150nt immediately downstream of the 5’ss after masking of the splice site signals. To quantify significance of over or under-representation, a *p*-value is calculated for every pentamer using the normal approximation for the binomial distribution. The test procedure is described in detail in (Zhang et al. 2005).

2. Frequency near "weak" vs. "strong" splice sites: Splicing enhancers have been shown to have a compensating role in the recognition of "weak" (non-consensus) splice signals. Therefore, we expect donor and acceptor intronic enhancers to be overrepresented in the intronic regions near weak versus strong 5' and 3' splice sites, respectively. We define as "strong" those splice sites found in the third quartile (top 25%) of scores using PSSMs constructed for canonical splice site motifs from the high-confidence set of constitutive exons defined above. We define as "weak" those splice sites found in the first quartile (bottom 25%) of the same scores. We calculated *p*-values as above to quantify the differential enrichment in intronic regions flanking strong vs. weak splice sites.

3. Overrepresentation in Conserved regions: We hypothesize that functional splicing sequence elements are subject to purifying selection. As a result we expect splicing auxiliary elements to be overrepresented in conserved regions near splice sites in multiple alignments of the sequenced genomes from *Drosophila* species. The frequency of each pentamer in conserved intronic regions flanking the 3' and 5' splice sites of both constitutive and alternative internal exons was estimated using the *PhastCons* (Siepel et al 2005) conserved elements that mapped within these regions in the 15-way insect multiple alignments available from the UCSC Genome Browser database (http://genome.ucsc.edu/) (Rhead et al 2010). We calculated *p*-values for each pentamer as above to quantify the differential enrichment in conserved vs. all

intronic regions flanking splice sites from both constitutive and alternative internal exons.

4. Positional Bias: The arrangement of auxiliary elements near splice sites is dictated by spatial constraints imposed by their trans-acting counterparts. As a result, we expect auxiliary elements to be distributed non-uniformly near splice sites (see also Yeo et al 2007). We estimate departure from uniformity using the two-sided one-sample Smirnov-Kolmogorov (SK) goodness of fit test. First we calculated the frequency of each pentamer in each bin of length=10nt, for the 150nt of intronic sequence flanking the 3'ss and 5'ss of both constitutive and alternative exons. Next, we applied Laplacian smoothing to each pentamer frequency for every bin. Finally, we calculated SK test *p*-values for the observed positional distribution versus the uniform distribution using the ks.test R function. We combined the resulting *p*-values from the four analyses using Fisher's method (Fisher 1948) in order to obtain a single *p*-value for every pentamer and infer splicing regulatory sequence elements. Overall, for classifying a pentamer as an intronic silencer, we required a combined p-value of <0.01 using criterion 1 (underrepresentation within intronic regions <150 nt from splice sites), 3 (overrepresentation in conserved regions), and 4 (positional bias), and an individual *p*-value of <0.1 for criterion 1. Conversely, for classifying a pentamer as an intronic enhancer we required a combined *p*-value of <0.01 using criteria 1 (overrepresentation in intronic regions <150 nt from splice sites), 3

(overrepresentation in conserved regions), and 4 (positional bias), and an individual *p*value of <0.1 for criterion 1.

Mutational analysis in *Ubx* minigenes

The *Ubx* minigene plasmids pUB.*Ubx*.4F12.RP and pUB.*Ubx*.RP* used in these experiments were described previously (Burnette et al 2005). The plasmid polylinker region was simplified to remove interfering restriction sites by digesting with *BamH* I and religating. This plasmid was designated pUB.*Ubx*.4F12.ΔBam. Targeted deletions and nucleotide substitutions in the candidate regulatory elements downstream of RSS RP3 in pUB.*Ubx*.4F12.ΔBam were made by PCR (see plasmid maps in appendix). Deletion ΔDE1 was constructed by deleting the region between +29 and +85 nucleotides downstream of RP3. The nucleotide substitutions in mut1, mut2, and mut3 were constructed by replacing the sequence corresponding to element A, Ψ5a, and element B with random sequence containing a *BamH* I site (Table 1).

Double mutants containing mΨ5b were constructed by repeating the mΨ5b mutagenesis in a mut1, 2, or 3 template. The double mutants mut1mut2, mut1mut3, and mut2mut3 were made by mutating element A or Ψ5a in the mut2 or mut3 templates. *EcoRI* was introduced instead of *BamH* I in the leading mutation to differentiate the two mutations. These double mutants were then combined with mΨ5b. Double mutants containing RP3* were made by mutating element A, Ψ5a, or element B in the previously described RP3* mutant template.

Table 1. Mutations introduced at features downstream of RP3

Construct	Mutated Element	Wild Type Sequence	Mutant Sequence
mut1	A	AACCAAAACAAAAACATTGACAAA	CGAAAATCGGATCCCGGTTTGGCC
mut2	Ψ5a	GTGAGT	GGATCC
mut3	B	AAATAAGTATAATAATAAA	ATCGGATCCCGGTTTGGCC
mΨ5b	Ψ5b	GTGAGT	CCATGG
1C,6C	Ψ5a	GTGAGT	CTGAGC

All constructs were verified by sequencing to confirm the presence of the desired mutation and the absence of spurious changes.

Compensatory Mutations in Ψ5a and U1 snRNA

We mutated Ψ5a from GTGAGT to CTGAGC (mutant 1C,6C in Table 1). We amplified a fragment of Drosophila genomic DNA containing the gene for the major U1 snRNA (U1a, encoded by *snRNA:U1:21D*), including its promoter and spanning from the upstream *Eco*RI site through the downstream *Bam*HI site. We inserted this fragment between the *Eco*RI and *Bam*HI sites of pBluescript-KS to yield pKS.21D.U1. We then introduced mutations that restored base-pairing complementarity with Ψ5a 1C,6C, yielding (pKS.21D.U1m). The compensatory mutations were U1-3G and U1-8G, changing ACTTAC to GCTTAG.

Analysis of Mutant Minigene Splicing in SL2 Cells

We maintained Drosophila SL2 cells in Schneider's Drosophila medium (Whittaker) supplemented with 12.5% Fetal Bovine Serum and 1% Penicillin/Streptomycin. We transfected minigene constructs into SL2 cells using Effectene reagent (Qiagen). We used 0.4 ug of *Ubx* minigene and 0.4 ug of a co-transfected standard (pPac.LacZ [Burnette et al 2005]) per ~10^6 cells in each well of a 6 well plate. We harvested cells after 48 hours and extracted total RNA using Trizol (Invitrogen). We primed reverse transcription with Superscript II (Invitrogen) on 1 ug of RNA with random hexamers, followed by treatment with RNase H. We subjected one tenth of the product to PCR amplification as

described previously (Burnette et al 2005) to detect the minigene mRNA, the minigene recursive intermediate, or the *LacZ* mRNA from the transfection standard. For suppression analysis with U1 snRNA, we co-transfected different combinations of wild type and mutant versions of the *Ubx* minigene and the U1 snRNA plasmid at a 1:2 ratio, as described for a similar analysis of *white* transcript splicing (Lo et al 1994). Products were analyzed by agarose gel electrophoresis in the presence of GelStar dye (Lonza). Gel images were captured digitally and bands were quantitated using ImageJ 1.41o software.

RESULTS

The landscape of flanking sequences around RSSs.

(This computational work was performed by fellow graduate student Panagiotis Papasaikas)

We began to explore the involvement of auxiliary sequences in RSS function by developing an in-house algorithm for prediction of intronic splicing silencers and enhancers in *D. melanogaster*. Our approach followed the rationale of previous methods for prediction of splicing regulatory signals (Fairbrother et al. 2002; Zhang and Chasin 2004; Zhang et al. 2005; Sugnet et al. 2006; Voelker and Berglund 2007; Yeo et al. 2007) and is described in detail in the Methods section. Briefly, we devised a simple statistical test that assesses the conservation, positional bias and over- or under-representation of *k*-mers in the proximity of constitutive 3' or 5' splice sites, and we used this test to infer pentamer motifs that are either stimulatory or inhibitory to splicing. For this analysis, we determined that *k*=5 was the maximum *k* that could be used while being able to obtain accurate estimates for the statistical tests. We derived two sets of pentamers corresponding to intronic enhancer motifs found upstream of regular 3' splice sites or downstream of regular 5' splice sites, and two analogous sets of silencer motifs. Downstream of 5' splice sites we predicted 156 *k*-mers as intronic splicing silencers and 85 *k*-mers as intronic splicing enhancers. Upstream of 3'splice sites we predicted 166 *k*mers as intronic splicing silencers and 117

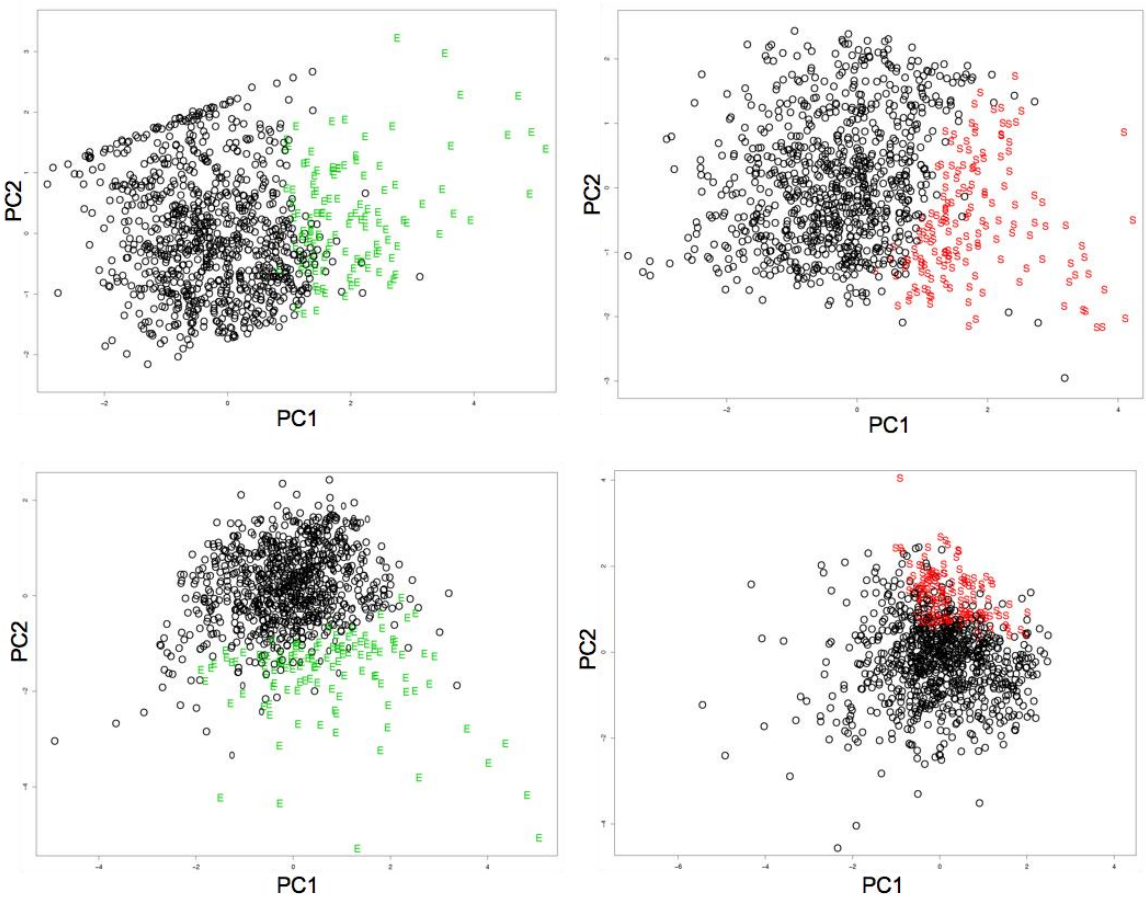


Figure 2.2. Analysis of all 1024 pentamers in the first two principal components of the attributes used for classification. Left: splicing enhancers (green *k*-mers) downstream of 5'splice sites (top) or upstream of 3' splice sites (bottom). Right: splicing silencers (red *k*-mers) downstream of 5' splice sites (top) or upstream of 3'splice sites (bottom).

*k*mers as intronic splicing enhancers. The analysis of all 1024 pentamers in the first two principal components of the attributes used for our predictions is shown in Figure 2.2.

The distribution of these motifs was then analyzed in a window of 500nt centered on 293 known or predicted RSSs that are not associated with known exons but were previously shown to exhibit high phylogenetic conservation and association with strong branch site predictions (Papasaikas et al 2010). For comparison we also performed the same analysis on two subsets of constitutive and cassette exons that are flanked by at least 500nt of intronic sequence and are supported by EST and cDNA data. These two exon subsets were held back during the prediction of intronic regulatory motifs. The results of this analysis are summarized in Figure 2.3 (Top and Bottom panels). As expected for functional splice sites, there is an overrepresentation of enhancers and underrepresentation of silencers adjacent to non-exonic RSSs. However, these biases are less pronounced than for constitutive exons and resemble more the situation for cassette exons. This could mean that most or all RSSs correspond to the ends of cassette exons that are rarely retained in mRNA. Alternatively, the higher information content of RSSs, particularly the distinctive 3'ss and association with strong branch site motiofs (Papasaikas et al 2010), may render them less dependent on enhancers and less sensitive to silencers. Additionally, reduced dependence on downstream enhancers may contribute to the correct sequential function of RSSs as 3' and 5' splice sites, which might otherwise clash with one another. Conversely, some downstream silencers may be required to prevent

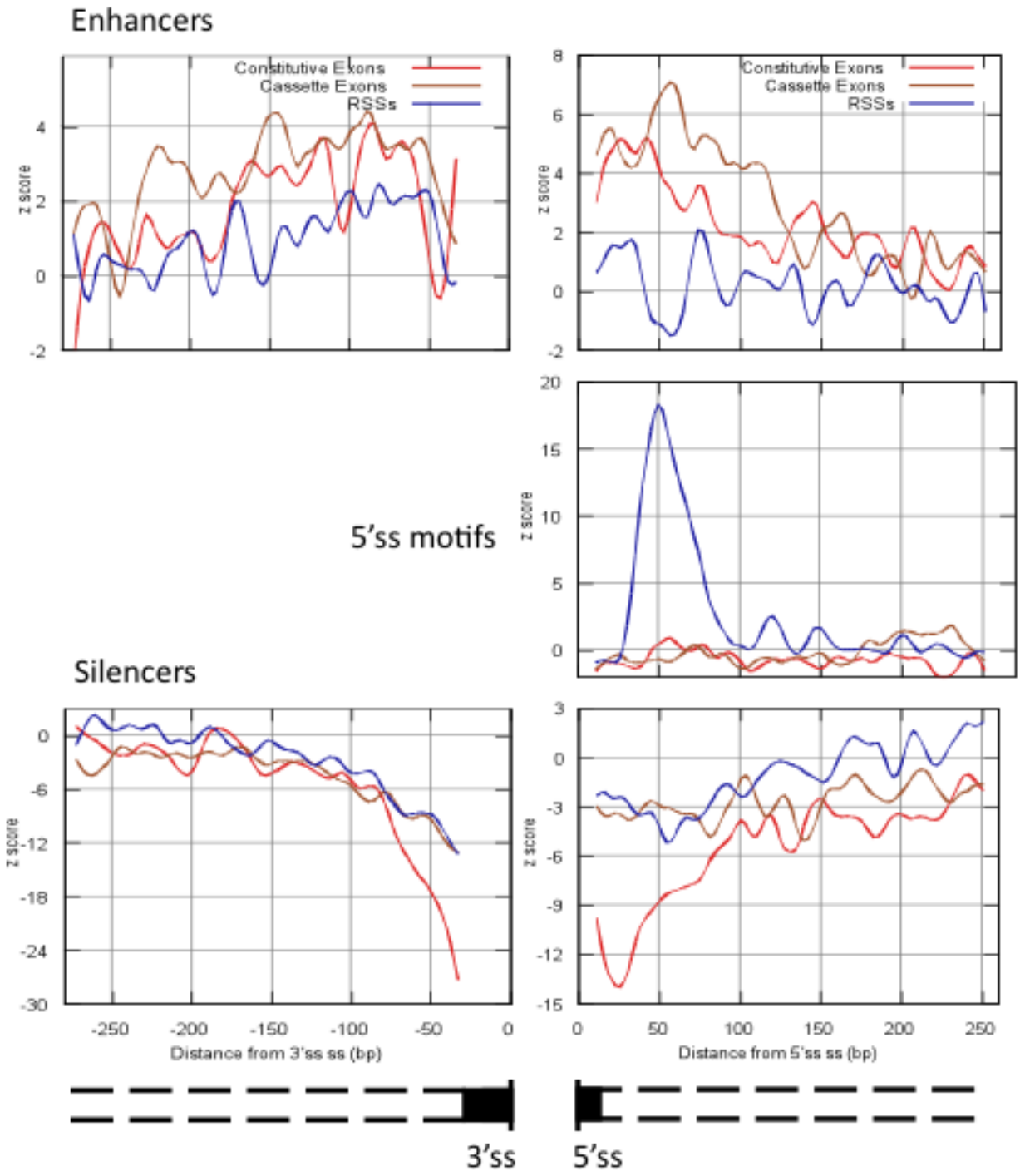


Figure 2.3. Distribution of predicted intronic auxiliary elements around non-exonic RSSs and constitutive or regular cassette exons of Drosophila. The z-scores in the y-axes correspond to standard deviations above or below background levels in random intronic sequences. The plots were soft smoothed using GNUplot's cubic spline function.

activation of cryptic 5’ splice sites in combination with the 3’ss component of the RSS.

To investigate these possibilities we constructed a position-specific scoring matrix (PSSM) for the canonical 5’ss motif that was derived from the high-confidence set of constitutive exons described in the Methods section. We assessed the distribution of 5’ss motifs around RSSs and exons as above, using a score cut-off that yields a False Positive Rate of 1 per kb against random intronic sequences. These random sequences were generated with a second-order Hidden Markov Model that captures the mono-, di- and tri-nucleotide composition of real Drosophila introns. The results are summarized in Figure 2.3 (Center panel). A striking peak of 5’ss motifs is found at ~50nt downstream of the RSSs, even though no documented exon is present in any of these cases. As evidenced by the lack of EST and cDNA records showing inclusion of these exons, and as confirmed by our experimental analysis for a selection of RSSs (Burnette et al 2005, Papasaikas et al 2010), these 5’ss motifs are rarely or never used for splicing in their normal physical context. We cannot exclude the possibility that these 5’ss motifs define recursive cassette exons that are only included in a few cells and/or trigger nonsense-mediated decay, but it does not appear that most of these regions are under selection for such a function (see Discussion).

It is interesting that the downstream 5’ss motifs are positioned between enhancer peaks in a region where standard cassette and constitutive exons are enriched for intronic enhancers (Figure 2.3). These patterns suggest that the 5’ss

motifs downstream of non-exonic RSSs play regulatory roles that coordinate the function of the 3'ss component and the regenerated 5'ss. For example, a downstream 5'ss motif might not serve as an actual point of splicing but still might interact with the 3'ss component of the RSS to aid its recognition via a mechanism analogous to regular exon definition (Berget 1995). Such an interaction would not be possible between the 3' and 5'ss components of the RSS, which are coincident. Subsequent to this intial step, the regenerated 5'ss in the rearranged substrate could be favored for splicing to the downstream exon. Alternatively, the downstream 5'ss motifs could function as components of splicing enhancers for the first or second steps of recursive splicing or as silencers to prevent inappropriate activation of pseudo-exons with other cryptic splice sites. In the sections that follow we describe tests of these hypotheses using a model system.

A conserved enhancer/pseudo5'ss module downstream of a non-exonic RSS in *Ubx*

The best-characterized non-exonic RSS is RP3 within the *Ultrabithorax* (*Ubx*) gene of *D. melanogaster* (Burnette et al 2005). This non-exonic RSS has been validated by analysis of recursive intermediates, splicing lariats and mutations. As illustrated in Figure 2.4, RP3 is located in the middle of a 50 kb intron, and it is followed 54 nt downstream by a strong pseudo-5'ss motif (designated Ψ5a) that is flanked by two regions (A and B) predicted to be splicing enhancers according to the analysis described above. This RSS--A-Ψ5a-B module, which exhibits the

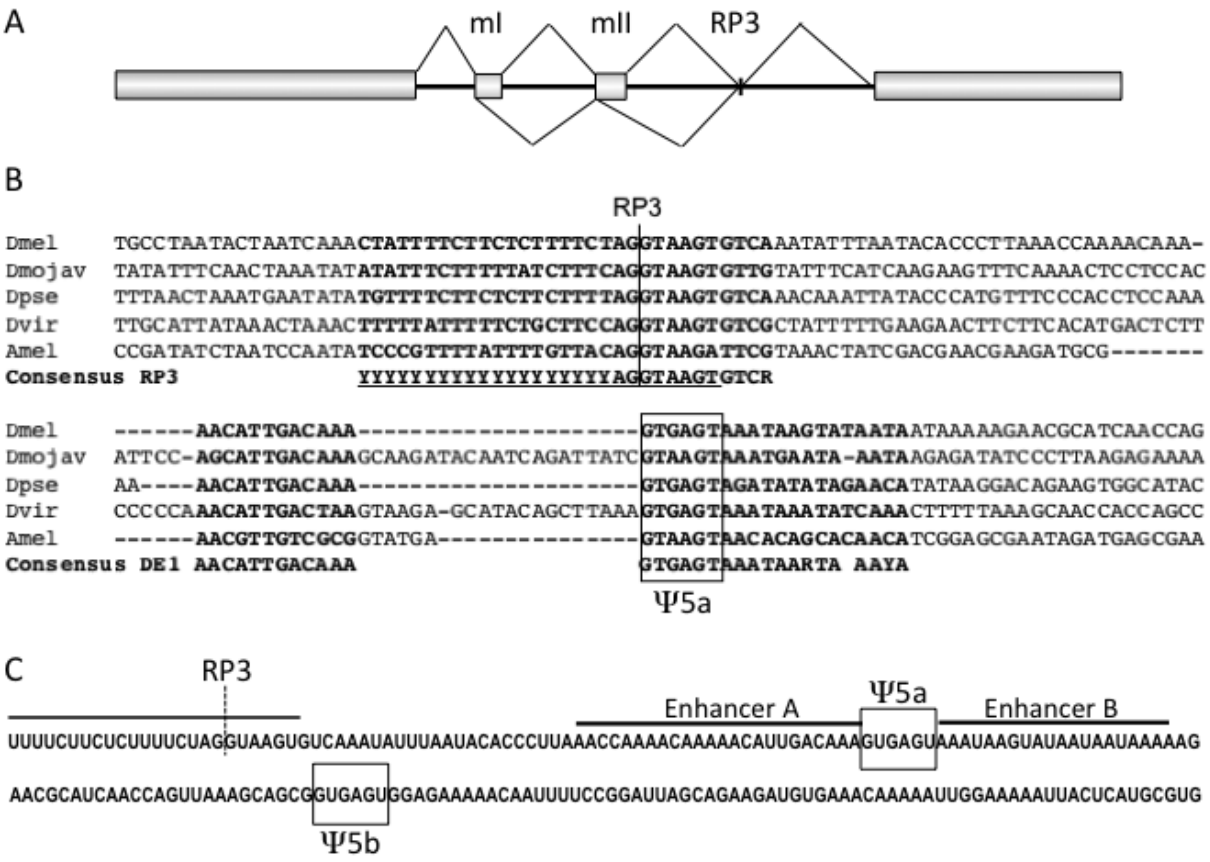


Figure 2.4. Non-exonic recursive splice site RP3 in *Ubx*. (A) Diagram of the *Ubx* transcription unit in Drosophila (not drawn to scale). mI and mII are 51 bp recursively spliced cassette exons. RP3 is a non-exonic recursive splice site in the middle of intron 3. The introns measure 7.4, 14.6, and 50 kb, respectively. mI, mII and RP3 are spliced co-transcriptionally (Lopez et al 1996, Burnette et al 2005). (B) Phylogenetic conservation of RP3 and a downstream module (DE1) containing a 5'ss motif (Ψ5a). (C) The sequences flanking Ψ5a within the conserved module correspond to predicted splicing enhancer regions A and B. Farther downstream is a non-conserved 5'ss motif (Ψ5b), located 107 nt from RP3. The sequence shown in this panel is from *D. melanogaster*.

same organization described above for non-exonic RSSs in the aggregate, is conserved among all sequenced *Drosophila* species and even more distantly related insects such as *Apis melifera* (honeybee), spanning ~300 MY of evolution (Russo et al 1995, Tamura et al 2004, Grimaldi and Engel 2005) (Figure 2.4). In some species, two pseudo-5' splice sites are located within the module (Figure 2.4). These observations suggest that the module plays an important functional role in *Ubx*.

RP3 presents an excellent model system to explore the role of 5'ss motifs downstream of non-exonic recursive splice sites. First, a *Ubx* minigene (*Ubx*.4F12.RP) has been shown to recapitulate highly efficient recursive splicing of RP3 in SL2 cells (Burnette et al 2005). Second, efficient recursive splicing at RP3 is essential for retention of two recursively spliced cassette exons (mI and mII) in *Ubx* mRNAs (Hatton et al 1998, Burnette et al 2005). The decision whether to resplice these cassette exons is made before and during their splicing to RP3 (mI to mII splicing and mI/mII- or mII- to RP3 splicing, respectively) (Burnette et al 2005) (Figure 2.4). Under normal circumstances, use of the 5'ss regenerated by RP3 is highly efficient and does not alter the choice whether to retain or remove mI and mII (Burnette et al. 2005) (Figure 2.4). A strong regenerated 5'ss motif (CAG/GUAAGU) is present at the junction between mI and its upstream exon, however, and if the 5' ss regenerated by RP3 is weakened by mutation, that junction is used instead to complete intron removal, excluding both mI and mII from the mRNA (Burnette et al. 2005). Finally, a

non-conserved cryptic 3'ss (designated Ψ5b: GCG/GUGAGU) is located in the intron farther downstream from Ψ5a, at +107 relative to RP3 (Figure 2.4). We exploit these properties in the experiments described below.

The enhancer/pseudo5' ss module is not required for activity of RP3 as a 3'ss but is required for efficient use of the regenerated 5'ss

We used a series of deletions and nucleotide substitutions in minigene *Ubx*.4F12.RP to probe the function of the conserved module downstream of RP3. Deletion of the entire module (Figure 2.5 lane 2) did not prevent production of the recursive intermediate, which instead accumulated consistently to higher levels than with the wild-type minigene. In addition, inappropriate use of the 5'ss at the mI exon junction increased slightly. Together, these results suggested that neither Ψ5a nor Ψ5b are required to activate RP3 as a 3'ss by exon definition, but that the A-Ψ5a-B module might instead be required for proper activity of the regenerated 5'ss. This would be consistent with previous results showing that a large deletion beginning 5 nt upstream of region A and extending downstream impaired the ability to use the mII/RP3 junction as a 5'ss in minigenes where RP3 was pre-spliced to the upstream exons (Burnette et al 2005). We explored this further by introducing nucleotide substitution mutations at element A, Ψ5a, element B, and Ψ5b, individually and in combination (Methods and Table 1).

Mutation of element A always resulted in increased accumulation of the recursive intermediate and activation of competing upstream and downstream 5' splice sites (Figure 2.5 lanes 3, 7, 10, 11, 13, 14), suggesting that this element

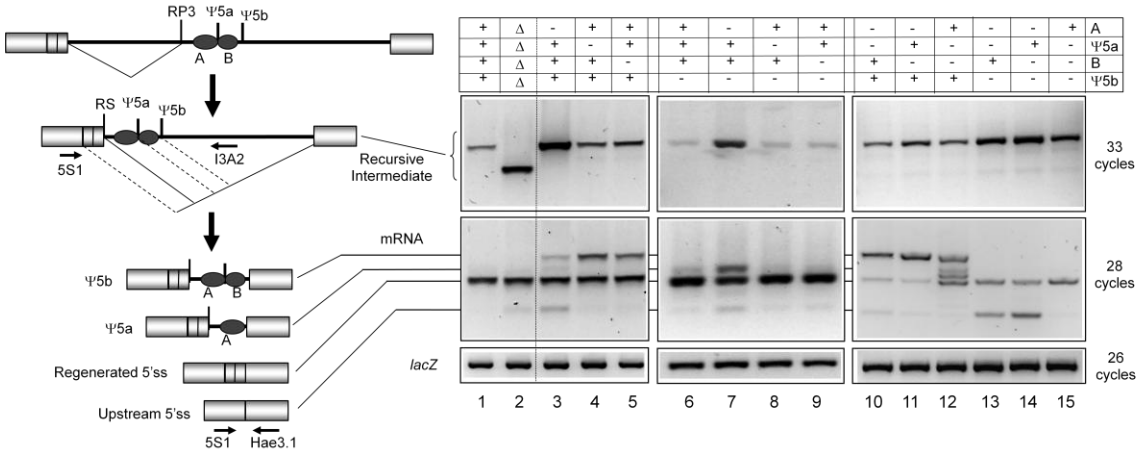


Figure 2.5. Mutational analysis of downstream elements. Left: Diagram of splicing pathway for minigene Ubx.4F12.RP. The first exon and cassette exons mI and mII are already joined in the minigene. Filled ovals represent predicted enhancer regions A and B. Small arrows represent primers for RT-PCR analysis. The first splicing event for minigene transcripts removes the first part of the intron using the 5'ss at the end of mII and RP3 as a 3'ss (Burnette et al 2005). This yields the Recursive Intermediate, which contains 4 potential 5' splice sites for removal of the second half of the intron (diagonal splicing lines). Normally the regenerated 5'ss (labeled "RS") outcompetes the others, as indicated by the solid splicing line, yielding a single mRNA species that contains mI and mII. Activation of the other potential 5' splice sites (dashed splicing lines) by mutations results in the alternative mRNA structures shown at the bottom, labeled according to the 5'ss used. **Right:** Mutant effects on minigene splicing. RT-PCR assays on total RNA from transfected SL2 cells were used to detect the recursive intermediates (top panels; primers 5S1+I3A2), the mRNAs (center panels; primers 5S1+Hae3.1) and the *lacZ* cotransfection control (bottom panels). The number of PCR cycles is indicated at the right. The minigene genotype for each lane is identified at the top: + indicates the element is wild type; - indicates the element is mutated as in Table 1; Δ indicates the element is missing because of a deletion spanning nucleotides +29 to +179 downstream of RP3.

acts as an enhancer for the regenerated 5'ss. Mutation of Ψ5a and/or element B increased the accumulation of recursive intermediate to a lesser extent but shifted splicing strongly to Ψ5b when this was available (Figure 2.5 lanes 4, 5, 12), or to the upstream 5'ss (although more weakly) when Ψ5b was inactivated by mutation (Figure 2.5 lanes 8, 9, 15). Ψ5a functioned effectively as a 5' splice site when element A was mutated, particularly if competition from Ψ5b was eliminated by mutating this 5'ss (Figure 2.5 lanes 3, 7). However, this aberrant use of Ψ5a as a 5'ss required wild-type element B (compare lanes 7 and 12 in Figure 2.5). Thus, element B functions as an enhancer for Ψ5a, but splicing at Ψ5a is normally inhibited by element A. These results are consistent with the function of A-Ψ5a-B as a functionally integrated module.

The effect of mutating this module or its subcomponents differed from that of mutating the 5'ss component of RP3 (RP3* in Figure 2.6, lane 1), which weakens the regenerated 5' splice site (Burnette et al 2005). All of these changes increased accumulation of the recursive intermediate, which is consistent with a less efficient turnover of this intermediate to give spliced mRNA (Figure 2.5 and Figure 2.6). However, mutations of the enhancer module favored activation of downstream 5' splice sites, whereas mutation of the 5'ss component of RP3 led almost exclusively to activation of the upstream 5'ss at the mI junction (Figure 2.6) (see also Burnette et al 2005). These differences are consistent with the hypothesis that A-Ψ5a-B acts as an enhancer module to stimulate use of upstream 5' splice sites, and that normally this effect is captured by the 5'ss regenerated by RP3, which is closer than the mI junction. We tested this by

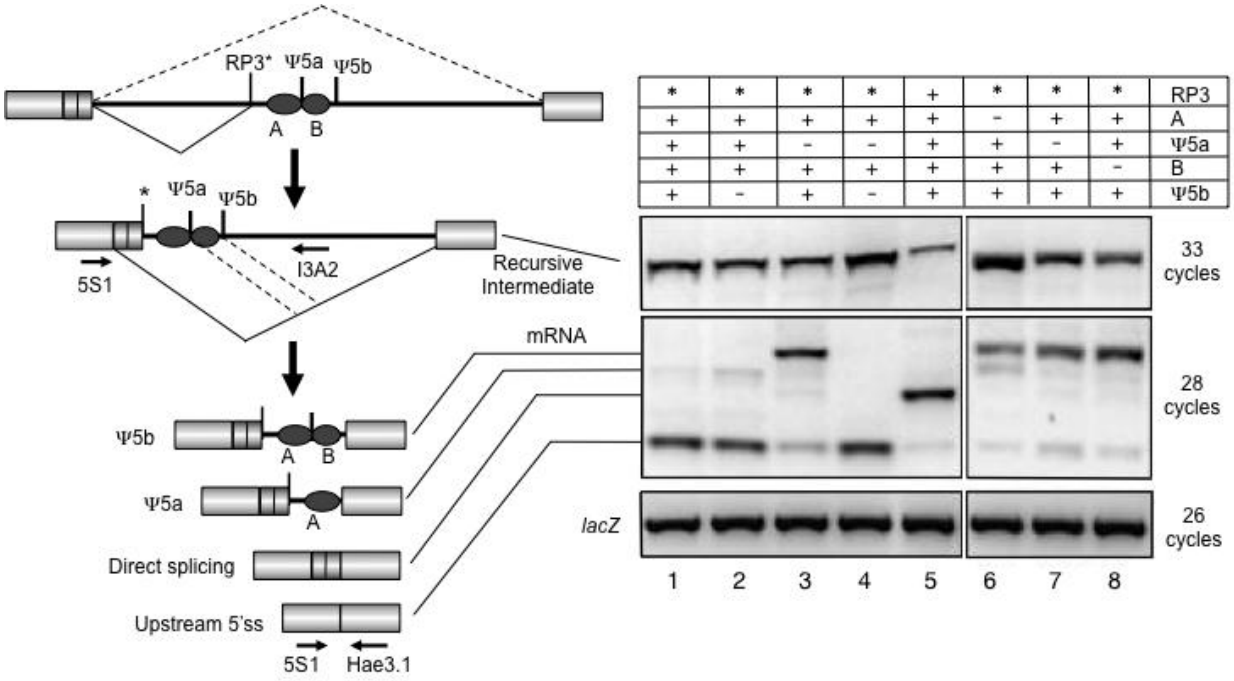


Figure 2.6. The downstream pseudo-5' splice sites are not required to activate RP3 for use as a 3'ss. Left: Diagram of splicing pathway for minigene Ubx.4F12.RP*, represented as in Figure 2.5. In this case RP3* is a mutant version of RP3 in which the regenerated 5'ss (*) has been weakened by mutation (Burnette et al. 2005). The first splicing event for minigene transcripts still removes the first part of the intron using the 5'ss at the end of mII and RP3* as a 3'ss [Burnette et al. 2005]. This yields a mutant Recursive Intermediate, which contains 3 potential 5' splice sites for removal of the second half of the intron (diagonal splicing lines). Normally the upstream 5'ss at the junction between the first exon and mI outcompetes the others, as indicated by the solid splicing lines, yielding a major mRNA species that lacks mI and mII. Mutations that impair activation of RP3 as a 3'ss should result in direct splicing (dashed splicing lines at top), restoring production of the normal mRNA structure containing mI and mII. **Right:** Mutant effects on minigene splicing. RT-PCR assays on total RNA from transfected SL2 cells were used as in Figure 2.5 to detect the recursive intermediates (top panel; primers 5S1+I3A2), the mRNAs (center panel; primers 5S1+Hae3.1) and the *lacZ* cotransfection control (bottom panel).

mutating module components in the context of the RP3* mutation. Now, 5'ss activity shifted almost completely to use of Ψ5a and/or Ψ5b (Figure 2.6 lanes 6-8), confirming that in the presence of the enhancer module upstream 5' splice sites are favored over downstream sites. In this case, too, increased use of Ψ5a was observed when element A was mutated, but not when element B was mutated (Figure 2.6 lane 1 vs 6, 8). Given that mutations in regions A, B or Ψ5a compromise turnover of the recursive intermediate, a possible additional role for these elements in enhancing the use of RP3 as a 3'ss might have been obscured in the initial mutant analyses of Figure 2.5. The results in Figure 2.6 (lanes 6-8) argue against a major role because increased accumulation of the recursive intermediate was still observed and the mutations did not restore significant production of the normal isoform containing mI and mII. Restoration of mI and mII inclusion would have been expected as a consequence of skipping RP3* (see diagram in Figure 2.6) because this is the mRNA structure that is produced when RP3 is deleted from the minigene and the intron is removed by one-step direct splicing (Burnette et al. 2005). However, since splicing shifted to Ψ5b in the double mutants that combine RP3* with disruptions of A, Ψ5A or B, it is possible that exon definition with this non-conserved 5'ss motif could substitute for the normal interactions. We tested this by mutating Ψ5b singly and in combination with Ψ5a in the RP3* background. Accumulation of the recursive intermediate continued to be observed in all cases, and there was no discernible shift to production of the normal mRNA containing mI and mII with the Ψ5a Ψ5b double mutant (compare lanes 2-4 with lanes 1 and 6 in Figure 2.6), supporting the

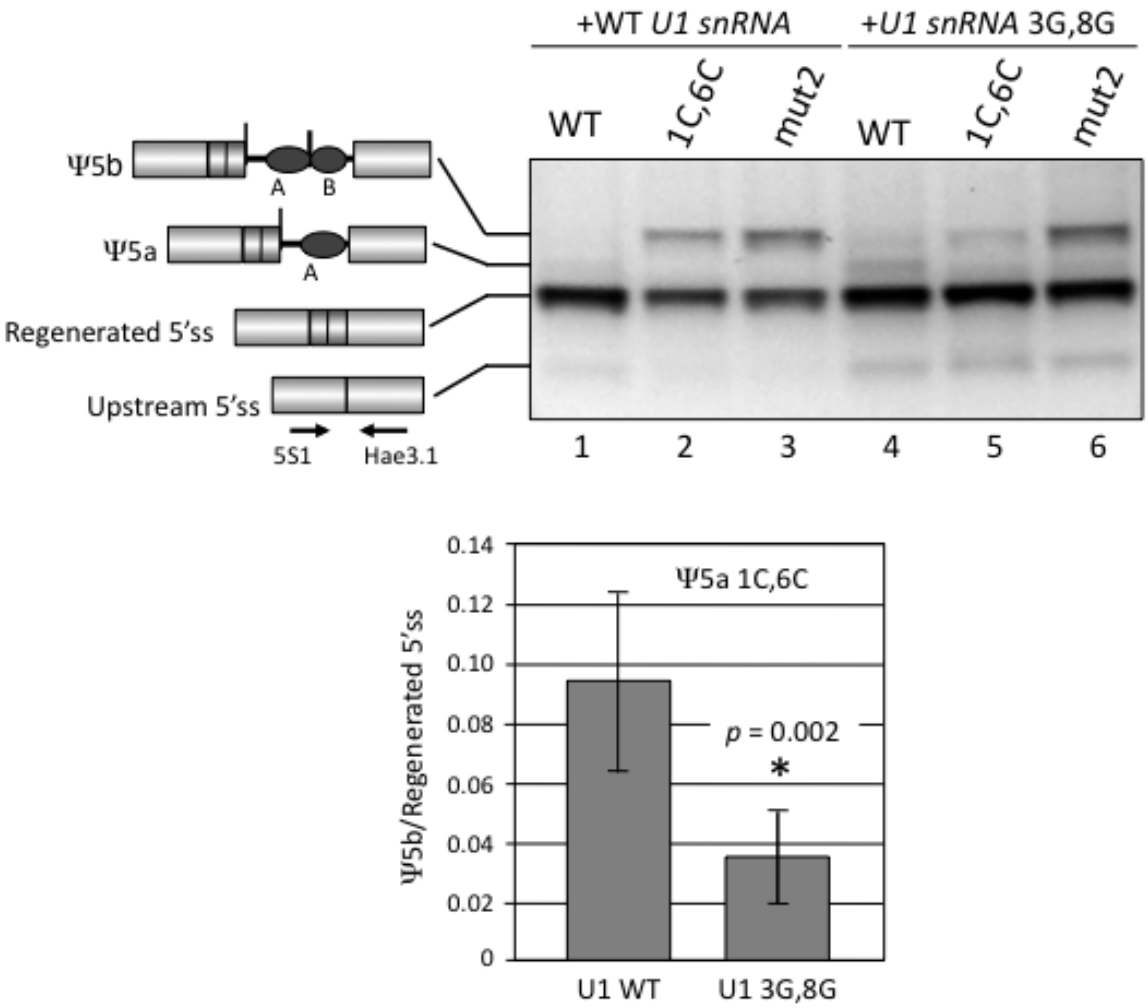


Figure 2.7. Specific suppression of Y5a resplicing defect by compensatory base changes in U1 snRNA. Minigenes bearing wild type Ψ5a or different Ψ5a mutations (mut2 or 1C,6C) were cotransfected into SL2 cells with either wild-type U1a snRNA or with mutant U1a snRNA (3G,8G) carrying base changes that restore complementarity to Ψ5a1C,6C but not to Ψ5amut2. RT-PCR assays on total RNA were used as in Figure 2.5 to detect the mRNAs (primers 5S1+Hae3.1). Refer to Figure 2.5 for the minigene recursive splicing pathway.

conclusion that neither pseudo-splice site assists recognition of the RP3 3'ss component by exon definition.

Function of the enhancer/Ψ5a module requires ability to base pair with U1snRNA

Results described above showed that Ψ5a can function as a 5'ss in some mutant backgrounds. This suggested that the conserved enhancer module might normally function by recruiting U1snRNA to Ψ5a and co-opting it for regulatory purposes instead of splicing. Alternatively Ψ5a may simply resemble a 5'ss motif but interact with a protein factor instead. We tested this by asking whether enhancer function disrupted by mutations in Ψ5a could be rescued by expression of U1 snRNA bearing compensatory mutations that restore base pairing. For this purpose we designed a second set of base substitutions in Ψ5a (1C,6C in Table 1) and we introduced a set of compensatory changes into the cloned U1 snRNA gene (3G,8G) (see Methods). Then we asked whether co-transfection with the mutant snRNA specifically rescued the enhancer defect of the complementary Ψ5a mutant. Figure 2.7 shows that this was the case. Cotransfection of wild type U1 snRNA did not rescue the enhancer defect of either Ψ5a mutant (mut2 or 1C,6C). In contrast the 3G,8G mutant U1 snRNA partially rescued the defect of the complementary Ψ5a mutant (1C,6C), and it did not rescue the defect of the non-complementary Ψ5a mutant (mut2).

The effect of splicing history on function of the *cis*-element module.

Transcripts from the minigene constructs analyzed above undergo one splicing event before the regenerated 5'ss is activated. During splicing, an exon junction complex (EJC) is deposited 20-24 nt upstream of the spliced junction, marking the position of a former intron and splicing event (reviewed by Bono and Gehring 2011). Recent data indicate that components of the EJC can affect splice site choice (Ashton-Beaucage et al 2010), although whether they do so before, during, or after deposition of the EJC is not known. Use of the regenerated 5'ss during the second step of recursive splicing would occur with an EJC already present. It is conceivable that this could have one of two effects. It could inhibit use of the regenerated 5' splice site, in which case downstream stimulatory elements might be required to remove the EJC. Alternatively, the presence of EJC components already in place might help stimulate use of the regenerated 5' splice site through interactions with additional factors, possibly with those bound at downstream enhancers. Either effect could account for the directionality of the downstream module at RP3. To test these possibilities, I introduced the same single-element mutations in element A, Ψ5'ss a and element B that were tested above into a minigene (pUB-UbxRI) where the upstream exons are pre-spliced to one another and to RP3, creating the same RNA architecture as in the recursive intermediate but without having gone through any previous splicing events.

The wild-type version of pUb-UbxRI showed no activation of Ψ5'ss a or b (Figure 2.8), suggesting that directionality of the downstream module does not depend on an upstream EJC. In addition, analysis of the mutant constructs

(Figure 2.8) showed that element A is still required for efficient use of the regenerated 5'ss despite the absence of an EJC. In fact, the effect of mutating element A was even stronger than in constructs that would acquire an EJC during splicing of upstream exons to RP3 (Fig. 2.8). This could be consistent with a partially redundant role for the EJC in helping to activate the regenerated 5'ss when it is created by a splicing event

Instead of activating Ψ5'ss a or Ψ5'ss b, the primary splice site used when element A is mutated in pUB-UbxRI is the upstream 5'ss at the E5'/ml junction, with only slight use of the normal regenerated 5'ss (Figure 2.8). The downstream pseudo-5' splice sites might be ignored in this contruct because the 3'ss component of RP3 is not available to define an exon with them, and/or the upstream sites are favored more strongly through interactions with the 5' cap of the mRNA as 5' splice sites for what is now the first exon.

Surprisingly, Ψ5'ss a and element B are not required for correct 5'ss choice in the pUB-UbxRI constructs. This suggests that the mechanism of activation for the regenerated 5'ss differs depending on whether it is created by a previous splicing event, despite the involvement of element A in both cases. This difference could involve alternative interactions to stabilize binding of relevant factors at enhancer element A, In the normal situation, Ψ5'ss a and element B may be required to overcome some hindrance to the recruitment of relevant factors to enhancer element A, for example because interaction with the RP3 5'ss component is blocked initially by 3'ss recognition factors and subsequently by the EJC.

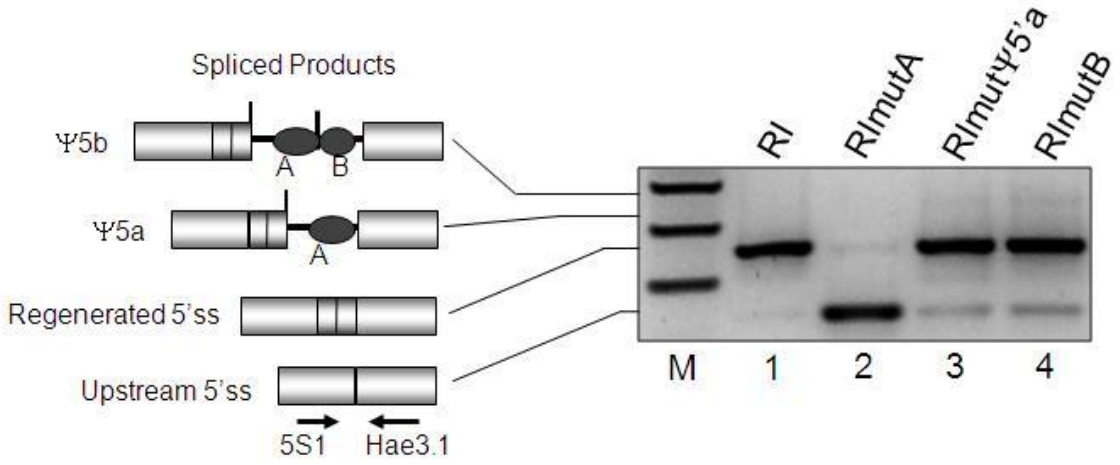


Figure 2.8. Effect of downstream element mutations when the regenerated 5'ss pre-exists in the RNA without previous splicing history. The recursive intermediate construct (pUB-UbxRI) mimics the state of the mRNA after the first step of recursive splicing, with the upstream exons pre-spliced to the recursive splice site.

DISCUSSION

In addition to sequence features of the RSS motif itself, it is likely that the activation and coordination of recursive splicing relies on additional auxiliary sequences. Analysis of upstream and downstream regions flanking the non-exonic EMSS-RSS predictions revealed an overrepresentation of predicted intronic enhancers and an underrepresentation of predicted intronic silencers, as would be expected for functional splicing elements. However, these biases were less pronounced than for constitutive exons and more similar to those observed for cassette exons. In contrast to cassette or even constitutive exons, however, the EMSS-RSS predictions were associated with a stronger overrepresentation of branch site predictions (Papasaikas et al 2010). These trends may be related to functional constraints arising from the location of non-exonic RSSs within very long introns and the coincidence of the 3' and 5'ss components. Stronger branch sites and higher information content of the 3'ss component, including longer Py tracts, may aid its recognition in the context of the long intron and avoid interference by the closely juxtaposed 5'ss. At the same time, these strengthened signals may render the 3'ss less dependent on enhancer mechanisms, which may also be deemphasized for the 5'ss in order to minimize competition during the first step of recursive splicing.

A puzzling observation is the presence of strong 5'ss predictions at ~50nt downstream of most non-exonic RSSs, including experimentally verified cases (Burnette et al 2005, Papasaikas et al 2010). One possibility is that they define

recursive cassette exons that have not yet been detected experimentally. Failure to detect these cassette exons during standard characterization of transcripts may be due to infrequent retention or to rapid degradation of mRNAs that retain them. For the majority of RSSs with a downstream 5'ss, inclusion of the potential cassette exon would truncate the ORF either by introducing in-frame stop codons and/or by shifting the frame and generating new downstream stop codons. The premature termination codons could potentially lead to destruction of the mRNA by the nonsense-mediated decay pathway (NMD), and this could have an important role in regulation of gene expression (reviewed by McGlincy and Smith 2008). This effect has been demonstrated recently for a likely recursive cassette exon in rat α-tropomyosin (Grellscheid and Smith 2006). Thus, some apparently non-exonic RSSs may define cassette exons with a negative regulatory role. True non-exonic RSSs do appear to exist, however, even when they are followed by downstream 5' splice sites. An example is RP3 in *Ubx* (Burnette et al 2005). That the majority of our non-exonic RSS predictions are neither coding cassette exons nor involved in NMD is also suggested by the fact that there is no significant over- or underrepresentation of in-frame versus out-of-frame stop codons (p>0.1 for both cases) in the region downstream of the RSS and up to the first 5'ss match. In addition, the low phylogenetic conservation scores downstream of the RSSs in *PhastCons* alignments (Papasaikas et al 2010) further suggest that, in most cases, the potential exon is not a part of the mature transcript.

An alternative hypothesis is that recognition of some RSSs is facilitated by exon definition-like interactions with a downstream 5'ss motif, even though that motif is not used for splicing. We tested this hypothesis by deletion and mutagenesis of conserved and non-conserved pseudo-5'splice sites downstream of non-exonic RSS RP3 in a *Ubx* minigene, and the results did not support a significant role in activation of RP3 as a 3'ss. On the contrary, the results suggest that the conserved pseudo-5'ss (Ψ5a) functions as part of an enhancer module that stimulates use of the regenerated 5' splice site, allowing it to compete effectively with surrounding alternative and cryptic 5' splice sites whose use would alter the structure of the mRNA and protein products. Suppression analysis with compensatory base changes showed that the ability to base pair with U1 snRNA is necessary for this function, suggesting that the enhancer module works by recruiting U1snRNA and coopting it for a positive regulatory role while preventing splicing at Ψ5a itself. The role of U1snRNA at Ψ5a may be to help stabilize binding of a ribonucleoprotein assembly, including factors bound at the flanking enhancer elements, that in turn helps to recruit U1 snRNP in a splicing-competent mode to the regenerated 5'ss. Simultaneously, this complex may prevent splicing at Ψ5a itself.

A working model that accounts for our experimental results is presented in Figure 2.9. Factors binding at region B help to recruit U1 snRNA to Ψ5a. This U1 snRNA (presumably in the context of U1 snRNP) also interacts with factors bound at region A, stabilizing a complex that in turn stimulates splicing at the nearest upstream 5'ss. The regenerated 5' splice site is normally the most

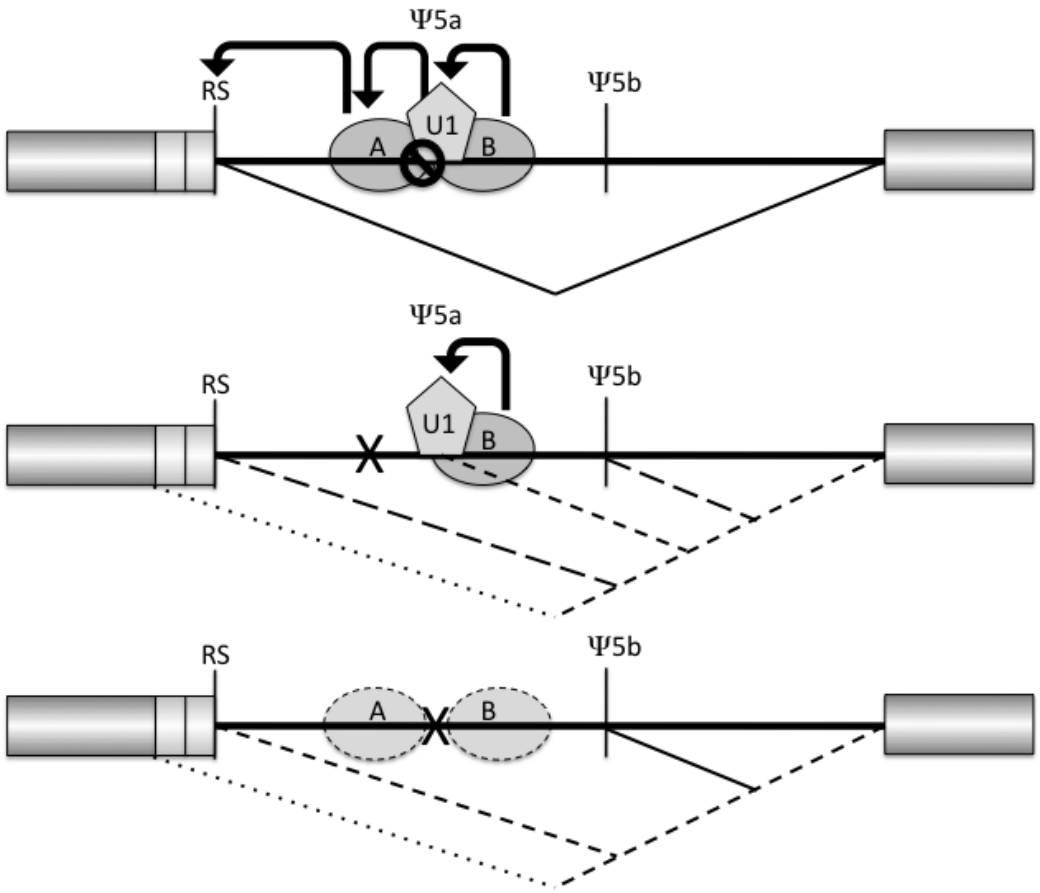


Figure 2.9. Model for the function of the A-Y5a-B module after use of RP3 as a 3'ss. Ovals represent protein complexes bound at enhancers A and B. The pentagon represents U1 snRNA bound at Ψ5a (presumably as part of U1 snRNP). Curved arrows represent interactions that stabilize binding and/or stimulate splicing. The NO sign indicates suppression of splicing at Ψ5a by factors recruited via enhancer A. RS denotes the regenerated 5'ss. X denotes inactivation of an element by mutation. **Top:** Wild type. **Middle:** Region A inactivated by mutation. **Bottom:** Ψ5a inactivated by mutation. See Discussion for details.

proximal and captures this effect, so that the upstream 5'ss at the mI junction is not activated. Downstream 5'splice sites (e.g. Ψ5b) are not influenced by the directional enhancer complex and are thus outcompeted. The interaction with factors at region A also prevents splicing at Ψ5a, so when A is mutated splicing occurs at this position, as long as region B is intact and helps recruit U1 snRNA. Mutation of region A also results in loss of enhancement at the regenerated 5'ss, so turnover of the recursive intermediate is slowed or inhibited and the upstream 5'ss, Ψ5a and Ψ5B can compete. Mutation of Ψ5a or region B also compromises turnover of the recursive intermediate and allows use of competing splice sites, because binding or activation of factors at region A is destabilized. This also explains why the genotypes in lanes 2 and 15 of Figure 2.5 produce similar splicing phenotypes even though region A is intact in lane 15 but deleted in lane 2; in both cases Ψ5a and region B are deleted or mutated so factors binding at region A are destabilized even when A is intact. Furthermore, the downstream 5'ss motifs and enhancer B are absent in both cases, leading to the same net result: increased accumulation of the recursive intermediate and slightly enhanced use of the upstream 5' splice site as a result of free competition with the regenerated 5'ss.

Recruitment of U1 snRNA in regulatory capacities has been observed previously in diverse situations. The effect observed depends on the position and sequence context of the pseudo-5'ss. In many cases the function is to inhibit processing, as in the regulation of *P*-element intron retention in Drosophila by a pseudo-5'ss (Siebel et al 1992), the suppression of splicing by retroviral 5'ss-like

elements (Hibbert et al 1999, Giles and Beemon 2005), and the suppression of a cryptic exon in the human *ATM1* gene by an internal pseudo-5'ss (Dhir et al 2010). The latter is a specific example of a general role for 5'ss-like sequences as exonic silencers that can prevent splicing of pseudoexons in human cells (Wang et al 2004). Positive effects include stimulation of polyadenylation by a U1snRNA-dependent enhancer (Lou et al 1996) and 5'ss stimulation by binding of U1 snRNA to a G-triplet intronic splicing enhancer (McCullough and Berget 2000). The positive regulatory function of Ψ5a downstream of RP3 may be a paradigm for other non-exonic RSSs, which are frequently associated with apparently cryptic 5'ss motifs at a similar position, as shown in Figure 2.3. Intuitively, it makes more sense for the highly conserved A-Ψ5a-B module to have evolved a primary function in promoting efficient use of the regenerated 5' splice site, a secondary consequence of which would be the suppression of non-conserved downstream pseudo-5' splice sites. For *Ubx* RP3, the enhancing role that we observe in the minigene context, with a 1.1 kb downstream intron segment, may be even more important for native transcripts, where the downstream intron segment measures 25 kb. In *Ubx* transcripts this enhancing function also has the effect of preventing inappropriate re-splicing at upstream exon-exon junctions in the RP3 recursive intermediate, thus ensuring the maintenance of tissue-specific alternative splicing choices that have already been made for recursive exons mI and mII.

The functional arrangement exemplified by RP3 and the A-Ψ5a-B module may also have important consequences in terms of evolutionary dynamics. The

system is poised for exonization as a consequence of mutations that disrupt the enhancers. Indeed, a spontaneous single-nucleotide change isolated in enhancer element B in the minigene was sufficient to activate alternative splicing of a cryptic exon defined between RP3 and Ψ5b (not shown). A possible example of exonization of a non-exonic RSS (exon 8 of *bruno3* in *D. pseudoobscura* and *D. persimilis*) has been described recently (Kandul and Noor 2009). Conversely, some current non-exonic RSSs could have evolved from recursive cassette exons by trapping the downstream 5'ss between enhancers and coopting it for enforcement of the regenerated 5'ss by the mechanism uncovered at *Ubx* RP3.

Chapter 3: Analysis of Biological Function of Non-Exonic Recursive Splice Sites

ABSTRACT

Genes that contain unusually long introns tend to have complex expression and play important roles in development and disease. Recursive splicing has been proposed to play a role in facilitating accurate splicing of such introns and or their transcription by elongating RNA Polymerase II. To test the hypothesis that recursive splicing is critical for the correct and/or efficient expression of genes with long introns, a two-step gene replacement strategy was used to delete the non-exonic RSS RP3 from within a 50 kb intron in the endogenous *Ultrabithorax* (*Ubx*) gene of *Drosophila melanogaster* and to generate isogenic wild-type control chromosomes. The effects of RP3 deletion on the expression and biological function of *Ubx* were assessed by RT-PCR across the developmental time course and by analysis of homeotic transformations in homozygotes and in different heteroallelic combinations. The results indicate that deletion of RP3 leads to a mild loss of function reflected in a morphological haltere phenotype and reduced viability. A change in alternatively spliced isoform ratios is detected during the larval stages. The RP3 deletion alleles also exhibit a synergistic interaction specifically with alleles of the RNA Pol-II 215 kd subunit that impair processivity. Additionally, a *white* marker gene inserted near RP3 is profoundly silenced but reactivated by deletions extending upstream, suggesting

a repressive chromatin structure in the RP3 region. Similar deletions of non-exonic recursive splice sites are in progress for the unrelated genes *frizzled* and *polychaetoid*.

INTRODUCTION

Recursive splice sites (RSSs) can mediate alternative splicing of cassette exons, 5'-terminal exons and 3'-terminal exons (Hatton et al 1998; Burnette et al 2005, Conklin et al, 2005, Grellscheid and Smith 2006, Kandul and Noor 2009, Papasaikas et al 2010), but most RSSs in Drosophila (~90%) are not associated with annotated or detectable exons. Nevertheless these "non-exonic" RSSs occur at much higher than expected frequency in the Drosophila genome and they are highly conserved across long evolutionary distances, suggesting that they have an important but unknown function in addition that of alternative splicing (Burnette et al 2005; Papasaikas et al 2010). RSSs in Drosophila are found only within introns that measure at least 5kb, and they are strongly enriched above expectation within introns of 10 kb or longer (Burnette et al 2005, Papasaikas et al 2010). This suggests a special role of RSSs in the expression of genes with very long introns.

The Advantages and Disadvantages of Long Introns

Introns in multicellular eukaryotes can be very large, extending through hundreds of thousands and even millions of nucleotides. Among metazoans, about 10% of human and 5% of Drosophila genes have introns longer than 10kb. These large introns can have important functions but they also present potential complications for gene expression and genome maintenance.

On the positive side, long introns frequently contain transcriptional regulatory elements, nested protein-coding genes, and non-coding RNA genes (reviewed by Fedorova and Fedorov 2003), although none of these have to be located within introns, nor does their presence necessarily require a very large intron. However, very long introns that accommodate many and diverse regulatory elements may be required for correct regulation of certain types of genes. Vinogradov (2006) suggests that a subset of genes called "intermediately expressed genes" is the most complex in the genome. Because intermediately expressed genes are not expressed in all tissues (unlike housekeeping genes) or one specific type (unlike highly-tissue specific genes), their accurate expression requires greater complexity of regulation. Intermediately expressed genes are often the longest in the human genome, containing many of the regulatory elements mentioned above, as well as more diverse and complex protein-coding domains. This increase in gene length (primarily due to increase in intron length) is proposed to allow more complex transcriptional regulation, including chromatin-mediated epigenetic regulation.

Long introns also provide a delay in the expression of the gene product because of the time required to complete transcription (~1 minute/kb in Drosophila, ~30 sec/kb in mammals), and this can have profound regulatory consequences (Ruden and Jackle 1995; reviewed by Thummel 1992). Additionally, long introns can increase the frequency of recombination between flanking exons, thus allowing selection to operate more efficiently on mutations in those exons (Comeron and Kreitman 2000, 2002).

On the negative side, however, long introns may complicate the accuracy or efficiency of pre-mRNA processing by increasing the probability of cryptic splice sites, cryptic polyadenylation sites, or secondary structures that may affect correct splicing. Furthermore, the association between 5' and 3' splice sites across long introns is delayed due to the required transcription time. Both effects make the pairing of correct splice sites more difficult, so that long introns require higher information content in their processing signals (Weir and Rice 2004, Weir et al 2006, Dewey et al 2006, Papasaikas et al 2010).

Long introns may also pose special challenges to transcript elongation. To begin with, completion of the transcripts may require many hours or even days. Furthermore, long introns may present many barriers to elongation in the form of specific sequences, repeats, protein-DNA interactions, chromatin structure, secondary structure of the nascent RNA, RNA-DNA hybrids and R-loops, or supercoiling induced by transcription (Chavez et al 2001, Huertas and Aguilera 2003, Li and Manley 2005, Voynov et al 2006).

Possible Roles of Recursive Splicing in Long Introns

Recursive splice sites could alleviate one or more of the problems posed by large introns. A possible local function would be to suppress cryptic splice sites and polyadenylation signals in a given neighborhood. Binding of U1 snRNP is known to inhibit 3'-end cleavage of pre-mRNA (Ashe et al 1997) and to suppress premature polyadenylation (Vagner et al 2000). It can also suppress inclusion of cryptic exons, as in the human ATM gene (Pagani et al 2002) as well as other

examples noted in Chapter 2. Co-transcriptional use of RSSs could also reduce the probability of cryptic processing or errors by avoiding the generation of full-length precursors. Reducing the size of the nascent pre-mRNA could help avoid the formation of RNA secondary structures or hnRNP complexes that could impair correct processing.

Transcription initiation, elongation and mRNA processing are coupled physically and functionally (Maniatis and Reed 2002). The process of splicing may thus stimulate gene expression through several mechanisms that influence transcription, polyadenylation and mRNA export. Co-transcriptional use of RSSs would increase the opportunities for interaction between the splicing and transcription machineries, and this could help to recruit or stimulate the activity of elongation factors for RNA Polymerase II. It has been shown that snRNPs binding to TAT-SF1, which binds to elongation factor P-TEFb, can stimulate elongation through a block in an HIV-1 template (Fong et al 2001). In humans, the splicing factor UAP56 (Hel25E in Drosophila) is recruited to the RNA during branch site recognition for splicing and it associates with both the elongating RNA polymerase and the THO complex (Wang et al 2005). Depletion of components of the THO complex impairs elongation. Mutations in elongation and export pathways have been shown to lead to genomic instability (Aguilera et al 2008, Huertas et al 2003), possibly at least in part through the formation of R-loops that render the DNA susceptible to damage or cleavage by endonucleases. Recursive splicing may alleviate this by reducing the size of the nascent pre-mRNA.

A role in transcriptional processivity would require recognition of the RSS during transcription. This seems likely given that most verified or predicted RSSs are followed by uninterrupted intron fragments of 5 kb or more. Co-transcriptional recognition of what we now know to be RSSs has been demonstrated in the *Ultrabithorax* gene (Lopez et al 1996). The experiments took advantage of the large size of the *Ubx* transcription unit and its synchronous activation early in development. In situ hybridization to early embryos was performed using probes that span the junctions created by what we now know are RSSs (the 5' ends of cassette exons mI and mII). Splicing between the upstream exons and these RSSs was detected before transcription of the next exon, and the signal was restricted to two dots in the nuclei, as expected for splicing on nascent transcripts. The timing relative to initiation of transcription was also consistent with splicing close to the time when RNA Pol II would traverse these regions but before it reaches the next exon.

Rationale of the Experimental Approach

In order to study the effects of recursive splicing in Drosophila, we designed experiments to cleanly delete three non-exonic RSSs from three different genes in the Drosophila genome: *Ultrabithorax*, *frizzled* and *polychaetoid*. We chose these genes because they are unrelated in sequence, function and expression but are well characterized genetically, so that loss-of-function alleles are available for complementation analysis. They also offer sensitive visible phenotypes in many tissue types.

Two-step gene replacement by Ends-In Homologous Recombination (Rong et al 2002; Figure 3.1) is one of several techniques available to manipulate genes in Drosophila (Wesolowska and Rong 2010). Although it is very laborious, we preferred this method because it is currently the only one that does not leave behind any exogenous DNA sequence and that allows sequence changes to be introduced at the native chromosomal location. Thus, it is well suited for studies of very large genes with complex regulation. The method starts by creating a donor element that is inserted into the genome of Drosophila by *P*-element-mediated transformation. This donor element carries a large fragment of the target intron and has been engineered to contain the mutation of interest and an *I-SceI* restriction site. The donor element is then excised as a circular DNA with FLP recombinase and cut in vivo within the target intron fragment by the I-*SceI* restriction enzyme. This triggers homologous recombination with the endogenous wild-type gene, duplicating the target region and inserting a *white* marker gene and a *CreI* restriction site between the duplicated elements. One of the duplicated elements should carry the mutant allele, and the other should carry the wild type allele. Finally, reduction of the duplication to a single copy is induced by cutting the DNA between the duplicated elements using the I-*CreI* restriction enzyme. This forces the double-stranded break to be repaired by recombination between the duplicated regions. The intervening *white* marker and other foreign sequences are deleted in the process, and a single copy of the target gene is regenerated that can carry either the mutant allele or the wild type allele. Multiple mutant and wild type reductions can be isolated from the same

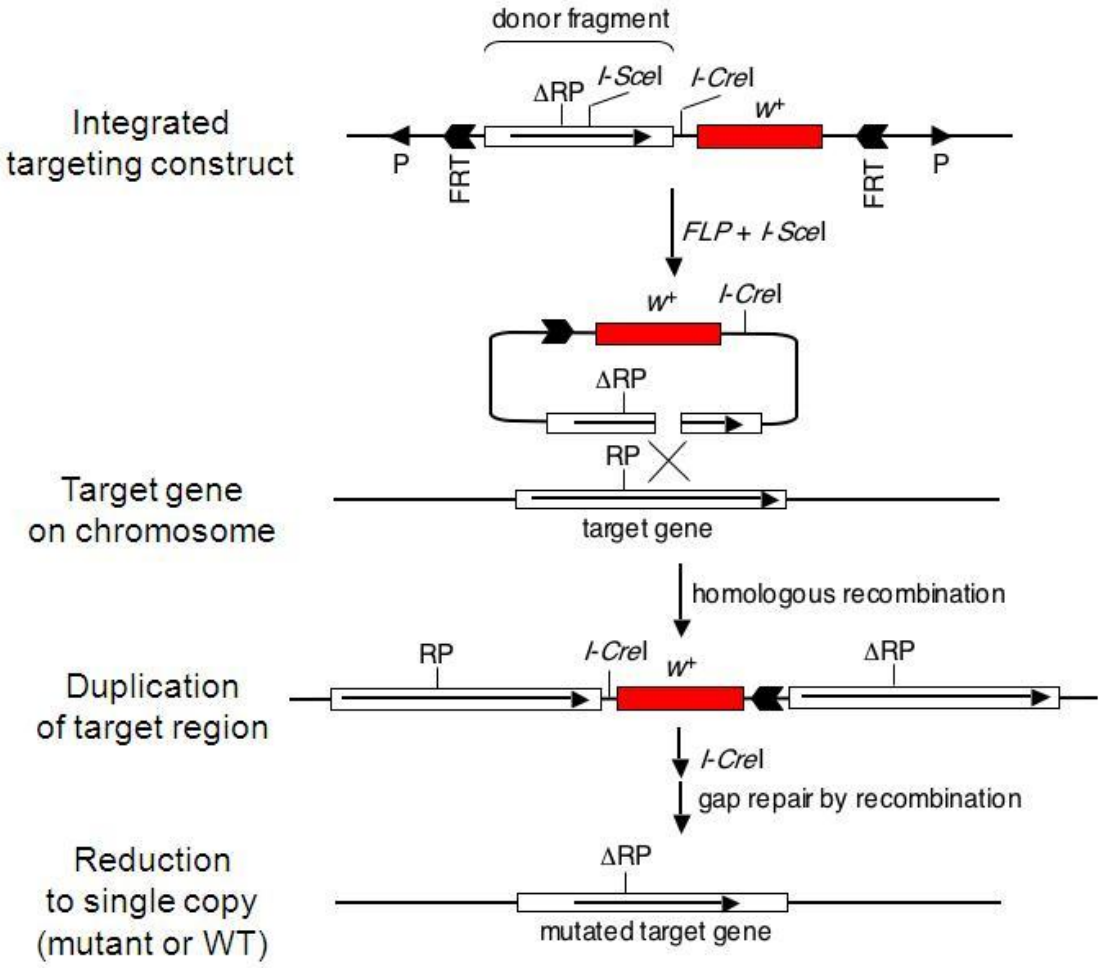


Figure 3.1. Strategy for allele replacement mutagenesis by ends-in homologous recombination.

starting duplication chromosome, so that isogenic mutant and control chromosomes are obtained. This is particularly important because the initial homologous recombination step frequently results in induction of second-site mutations.

These experiments have been performed by myself and three undergraduates under my supervision: Steve Riley for deletion of an RSS in *Ultrabithorax*, Rachel Ehrlich in *frizzled*, and Sherry He in *polychaetoid*. The *Ultrabithorax* deletion has been completed; the first step has been done for *frizzled* and *polychaetoid* and completion is in progress.

The Test Gene *Ultrabithorax*

The first RSS we deleted was the third RSS (RP3) in the *Ultrabithorax* gene. *Ultrabithorax* (*Ubx*) is one of three genes of the Bithorax Complex, along with *Abdominal-A* and *Abdominal*-B. The Bithorax Complex is a set of homeotic genes that control the identity of the segments that comprise the posterior two-thirds of the fly (Review by Maeda and Karch 2006). *Ubx* contains three RSSs (two exonic, one non-exonic) and has been used before for studying recursive splicing in vivo and in cell culture (Hatton et al 1998, Burnette et al 1999, Burnette et al 2005). The *Ubx* transcription unit is 78kb long and is composed of a constitutive exon (E5') followed by two recursively spliced cassette exons (mI and mII), and a final constitutive exon (E3') (Figure 3.2). The two exonic recursive splice sites control the alternative splicing of the cassette exons (Hatton

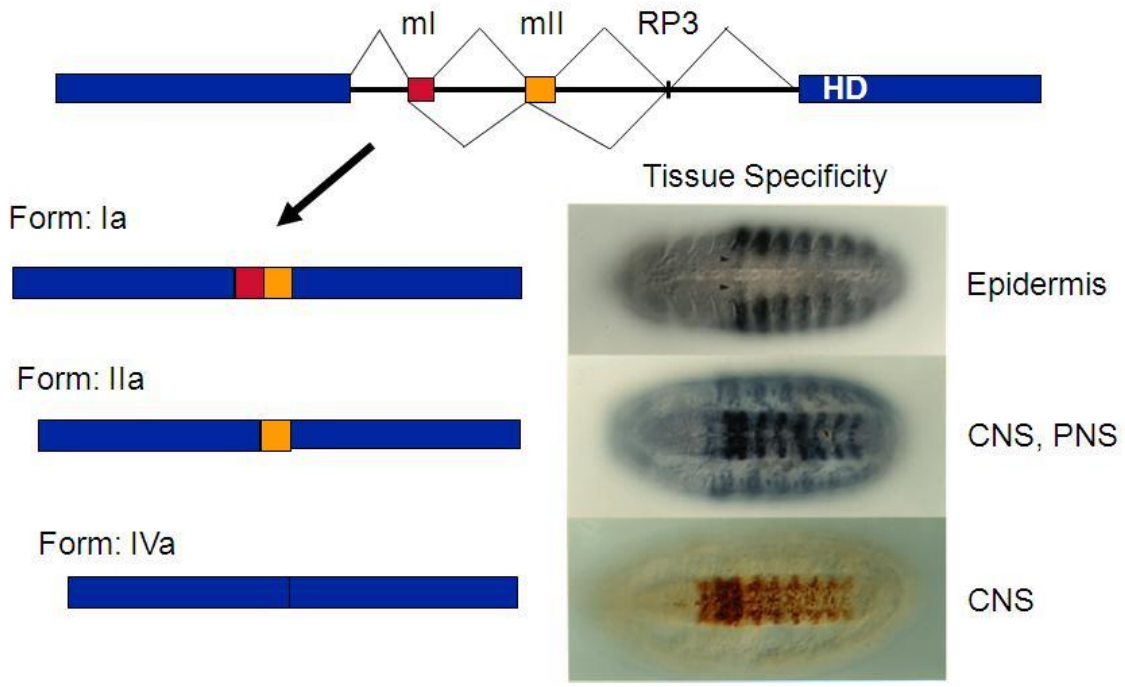


Figure 3.2. Ultrabithorax gene structure and expression. Blue exons are constitutively spliced, red and orange exons are cassette exons. Isoform Ia, IIa, and IVa splicing shown, along with the location of embryonic expression (visualized with isoform-specific monoclonal antibodies: Lopez and Hogness, 1990).

et al 1998). A single non-exonic recursive splice site (RP3) is located in the middle of the 50kb third intron (Burnette et al 2005). *Ubx* produces three primary alternatively spliced mRNA isoforms called Ia, IIa, and IVa (Figure 3.2). Ia contains all of the exons spliced together, while IIa excludes mI, and IVa excludes both mI and mII. A 28bp element between competing 5' splice sites at the end of E5' can also be included in mRNAs, creating the Ib, IIb, and IVb variants, but these are only expressed at low levels during development.

The alternative splicing of *Ubx* RNAs is tissue specific and stage specific (Kornfeld et al 1989, O'Connor et al 1988, Lopez and Hogness 1991). Ia and Ib isoforms are found in the epidermis and mesoderm, while IIa and IIb isoforms are primarily found in the central nervous system, with low expression in the epidermis and mesoderm, and IVa and IVb isoforms are exclusively found in the central nervous system. The different isoforms are functionally distinct (Subramaniam et al 1994, Reed et al 2010). The *Ubx* transcription unit is regulated by cis-acting elements located in two broad regions, the *bxd/pbx* region, which lies upstream of the promoter, and the *abx/bx* region, which lies within the third intron (Figure 3.3A).

Null alleles of *Ubx* are recessive lethal but show a dominant phenotype due to haploinsufficiency. This phenotype is a partial transformation of haltere into wing, manifested as a slight enlargement of the haltere and the appearance of a few wing-like marginal bristles on its surface. Stronger reductions of function in heterozygotes between strong and weak alleles produce progressively stronger transformations of the third thoracic (T3) and first

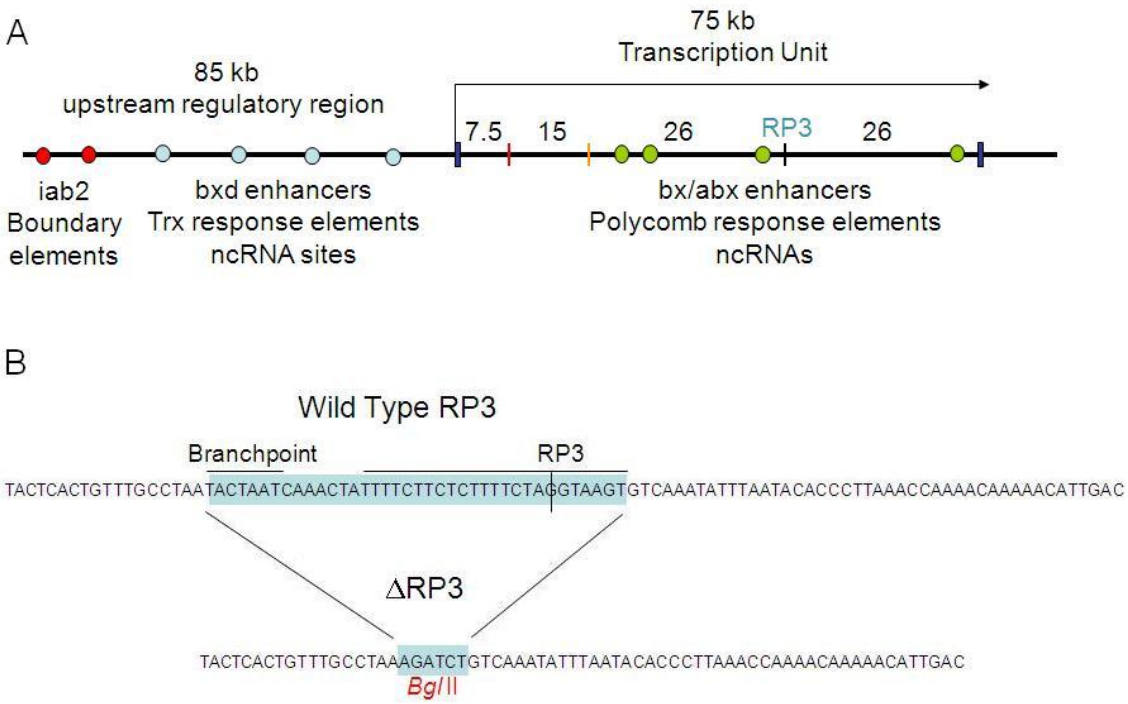


Figure 3.3. The *Ubx* transcription unit and the deletion of non-exonic RSS RP3. (A) *Ubx* gene structure and transcriptional regulatory elements. Red circles denote the iab2 boundary elements; blue circles denote upstream regulatory elements; green circles denote intragenic regulatory elements. Vertical bars denote exons and/or RSSs: dark blue are constituve exons; red and orange are recursively spliced cassette exons; black is non-exonic RSS RP3 (labeled). Numbers indicate length of intron segments in kb. (B) The engineered RP3 deletion removes 44 bp from the branch point sequence through the 5'ss component of RP3, and replaces this with a *Bgl*II restriction enzyme site.

abdominal segment (A1) into second thoracic segment (T2), including the haltere and third leg, the metanotum and pleural regions and internal organs. This allelic series includes rearrangements, nonsense mutations and small insertions and deletions in coding and regulatory elements. Known gain-of-function alleles produce the opposite effect in the dorsal mesothorax, dominantly transforming the wing into a haltere. The strongest of these alleles (*Cbx*^{Hm}) has a simultaneous loss-of-function that recessively transforms posterior haltere into wing and anterior first abdominal segment into third thoracic segment, allowing evaluation of both gain and loss-of-function by the RP3 deletion.

In the presence of RP3, recursive splicing is the predominant processing pathway for removal of the host intron (Burnette et al 2005). The possible requirement for RP3 during the splicing of *Ubx* was tested previously in cell transfection experiments (Burnette et al 2005), where a deletion of RP3 had no effect on the outcome of splicing of *Ubx* although it changed the pathway. However, those tests were done using a minigene system where exon E5' was already spliced to mI and mII, and the RP3 intron had been shortened to either 1 or 10kb. In addition, the minigene was driven by a heterologous promoter from a transfected plasmid in a non-relevant single cell type, and it was not required to provide biological function

In this work, I deleted the non-exonic recursive splice site RP3 from the endogenous *Ultrabithorax* gene using ends-in homologous recombination. The initial mobilization of the donor element into the RP3 region generated a *Ubx* loss-of-function phenotype whose intensity depended on whether the duplicated

target sequences were both wild type, one wild type and one mutant, or both mutant. After reduction to single copy, RP3 deletion alleles produced a weak loss of function haltere phenotype and exhibited altered mRNA isoform ratios during the larval stages. They also exhibited a synergistic phenotypic interaction with alleles of RNA Pol-II that have reduced transcriptional processivity. Additionally, we gained insight into the complications of using ends-in targeting, as the RP3 region has a very strong silencing effect against expression of the *white* marker gene from the targeting vector.

MATERIALS AND METHODS

RP3 Donor Element

All PCR reactions to construct the donor element were done with Platinum Taq Hifi (Invitrogen). Two PCR reactions were used to clone the donor element while deleting RP3. Primer sequences are located in Appendix Table 2. The first reaction was with primers RP3.TK.A.F2 and RP3.del.B.R to amplify the upstream fragment, while the other reaction was with primers RP3.del.B.F and RP3.TK.B.R to amplify the downstream fragment. These two fragments removed 44bp of the RP3 region, from the beginning of the branchsite motif through the end of the 5' splice site motif of RP3. The two amplimers were ligated to each other through engineered *Bgl*II sites on the RP3.del.B.R and RP3.del.B.F primers, and into the pKS Bluescript (Strategene) vector at the *NotI* and *Acc65I* sites. This construct was named pKS.ΔRP3, which was used as a template to amplify two more amplimers, one using PR3.TK.A.F2 with RP3.TK.A.R (4.15kb), and the other using RP3.TK.B.F (which has an I-*SceI* restriction site at the 5' end) with RP3.TK.B.R (4.15kb). Blunt end ligation was use to join the two amplimers, and the ligated fragments were then used as a template to amplify with primers RP3.L.DF and RP3.L.TK.BR (5.53kb). The resulting amplimer was digested with *Bgl*II and *Acc65*I, and ligated into pTV2.ΔRP3 at the *Bgl*II and *Acc65*I sites. This construct was called pKS.ΔRP3.I-SceI. The 8.3kb modified fragment was cloned into the pTV2 vector through NotI and Acc65I sites.

Genetic Procedures

Detailed information about the Drosophila genes and chromosomes mentioned here can be found at http://flybase.bio.indiana.edu/. Strains used in this chapter are listed in Table 1. Strains for Ends-In Homologous Recombination were obtained from the Bloomington Stock Center, along with RNA Polymerase II mutants used to test effects of impaired transcriptional processivity. *Ubx* alleles were from our lab stocks. Mapping and strain construction were accomplished using standard balancer chromosomes. Heat shocks for induction of *FLP* recombinase and homing endonucleases *I-Sce* I and *I-Cre* I were performed as described (Rong et al 2002).

PCR Verification of Homologous Recombination and Reduction

Genomic DNA was obtained by anesthetizing ten flies, and homogenizing them in a disposable microtube with pestle (Fisher) with 100uL of homogenization buffer (80mM NaCl, 60mM EDTA, 5.5% Sucrose, 0.5% SDS, 125mM Tris-HCl). The homogenate was incubated at 65 degrees Celsius for 30 minutes. 22.4uL of 5M potassium acetate) was mixed into the homogenate which was then incubated on ice for 60 minutes. Cellular debris was pelleted and the supernatant removed to a new tube, to which an equal volume of 95% Ethanol was added to precipitate the DNA overnight. The DNA was resuspended in 100uL of TE buffer (10mM Tris, 1mM EDTA, pH 8), and additionally purified by phenol chloroform extraction with final resuspension in 100uL of TE buffer.

Table 3.1. Strains for Gene Replacement

Genotype	Source
y^1 *w**; *P{ry*$^{+t7.2}$*=hs-FLP}*11 *P{v*$^{+t1.8}$*=hs-I-SceI}*2B *sna*^{Sco}*/CyO S*^{2}	Bloomington #6934
w^{1118}; *P{hs-I-CreI.R}*1A *Sb*^{1}*/TM6 Ubx*^{P15}	Bloomington #6937
w^{1118}; *P{ry*$^{+t7.2}$*=hs-FLP}*10	Bloomington #6938
y^1 *w**; *Ubx*$^{abx59.1}$	Recovered during initial screen
y^1 *w**; *TM2 Ubx*^{130}*/TM6B Tb Hu*	Lopez lab
*w**; *TM6B Tb Hu/MKRS*	Lopez lab
*w**; [*CyO;MKRS*]*/T(2;3)ap*^{Xa}	Lopez lab
w^{1118}; *P{hs-FLP}*10; *TM2 Ubx*^{130}*/TM6B Tb Hu*	This study

hs- denotes a cDNA fusion to the heat-inducible promoter from *hsp70*
* denotes an unspecified null allele

PCR reactions using AccuPrime Taq (Invitrogen) were used to verify correct targeting by using primers that amplified three different sections of the expected targeting product. The first primer set (upst.LF and pTV2.CreI.R) amplified from upstream of the duplicated targeting sequence to the *CreI* site (designated the left amplimer). The second set (wtRP3.downst.F and dRP3.upst.R) amplified from upstream of the *CreI* site to downstream of the FRT site (designated the middle amplimer). The last set (pTV2.FRT.F2 and downst.RR2) amplified from the *CreI* site to downstream of the duplicated targeting sequence (designated the right amplimer). Four more primer sets (upst.LF2 and pTV2.CreI.R2, upst.LF3 and pTV2.CreI.R3, pTV2.FRT.F3 and downst.RR3, and pTV2.FRT.F4 and downst.RR4) were designed to validate the left and right amplimers independently, with two sets of primers for each amplimer. *Bgl*II digests of the left and right amplimers verified the presence or absence of the deletion. For the reductions, the primers flanking the targeting sequence (upst.LF and downst.RR2) were used to produce an amplimer which was digested with *Bgl*II to verify the reduction to either wild-type or deletion mutant.

Analysis of mRNA, Predicted Intermediates, and Lariats

Both mutant and wild-type reduction strains were raised at 25°C. For both strains, embryos were collected for developmental stage intervals 0-4, 4-8, 8-12, 12-16, 16-20, and 20-24 hours after egg laying. The first, second, and third instar larval stages were also collected. Embryos and larvae were crushed in 40uL

Trizol (Invitrogen) using micro-mortars and pestles (Fisher) before bringing the total volume of crushed tissue in Trizol up to 1mL. The rest of the extraction was performed according to the manufacturer's protocol. RT-PCR reactions were done with the Superscript II Reverse Transcription System with Platinum Taq (Invitrogen) using random hexamers. cDNA synthesis was done at 25 degrees for 10', followed by 42 degrees for 50', and finally heat inactivation at 65 degrees for 15'. PCR for mRNA was done for 33 cycles (95 degrees for 30", 57 degrees for 30", 72 degrees for 30") in a 25uL reaction using primers *Ubx*.5S1 and *Ubx*.3A1 (from the first to the last exons of *Ubx*). RT-PCR to detect recursive intermediates used a similar protocol as above except for 38 cycles, with *Ubx*.5S1 and I3A2 primers. Rp49 was used as a control and quantitation reference, with a similar RT-PCR protocol but only amplifying for 22 cycles with primers Rp49.F1 and Rp49.R1.

RESULTS

Homologous Recombination of an RP3 Deletion into the *Ubx* Locus

To test the function of a RSS in the context of the endogenous gene, we deleted the non-exonic RSS RP3 from the *Ultrabithorax* gene using allele substitution by Ends-In Homologous Recombination (Rong et al 2002), which allows us to delete RP3 cleanly from the genome. The deletion spans 44bp, from the branch point upstream of the 3'ss component of RP3 through the end of the 5'ss component of RP3, and replaces the RSS with a *Bgl*II restriction enzyme site (Figure 3.3B).

Excision of the donor element by *FLP* recombinase produces white-eyed flies due to loss of the w^+ marker and is very efficient (>99%; Rong and Golic 2001). Flies that have integrated the donor element into the target site after *FLP* excision and homologous recombination should have pigmented eyes. Flies that have not excised the donor during *FLP* expression or have re-integrated the donor by non-homologous recombination can also have pigmented eyes. Eye pigmentation resulting from integrations by homologous recombination into the target site can be distinguished from the other sources of eye pigmentation by testing for change of linkage of w^+ to the target chromosome, followed by molecular analysis.

We started with a donor strain (M59.F1) that had pTV2.ΔRP3 inserted into the third chromosome by *P*-element transposition. This location was not optimal for screening, given that *Ubx* is also located on the third chromosome, but this was the first donor strain available. Strain M59.F1 was mated to *w;P{hs-FLP},*

P{hs-I-SceI}, Sco/Cyo, which contains *FLP* recombinase and I-*SceI* restriction enzyme under control of a heat shock promoter. At age 1-3 days, the progeny *w; P{hs-FLP}, P{hs-I-SceI}, Sco; M59.M1A/+* larvae were heat shocked at 38.5 degrees Celsius for one hour. 2000 white-eyed or mosaic-eyed progeny were crossed to *y,w;P{hs-FLP10}; TM2/TM6B*, which constitutively expresses FLP recombinase and introduces third choromosome balancers (Figure 3.4A). This allowed screening progeny for eye color that remains stable in the presence of FLP recombinase. This is expected to be associated with marker reintegration by homologous recombination because the w^+ marker is no longer flanked by FLP sites, although it can also result from non-targeted integrations. About 60000 flies were screened for pigmented eyes. Five flies with red eyes were obtained, but mapping showed that they were non-targeted insertions. A light yellow-eyed fly was recovered with a weak *Ubx* phenotype. RT-PCR analysis showed that it did not carry the RP3 deletion. However, there was an aberrant insertion upstream of RP3 that probably disrupted the *abx* regulatory region, causing the weak *Ubx* phenotype (data not shown). This new homozygous viable, recessive *Ubx* allele was named $Ubx^{abx59.1}$. We also recovered a fly that had a weak dominant loss-of-function *Ubx* phenotype, but no pigment in the eyes. Further analysis revealed that this fly carried a correct integration of the donor element into the RP3 region; with the expected duplication of the targeted region (Figure 3.5B, lane 1 and 2). This correctly targeted *Ubx* allele was named Ubx^{DF35}.

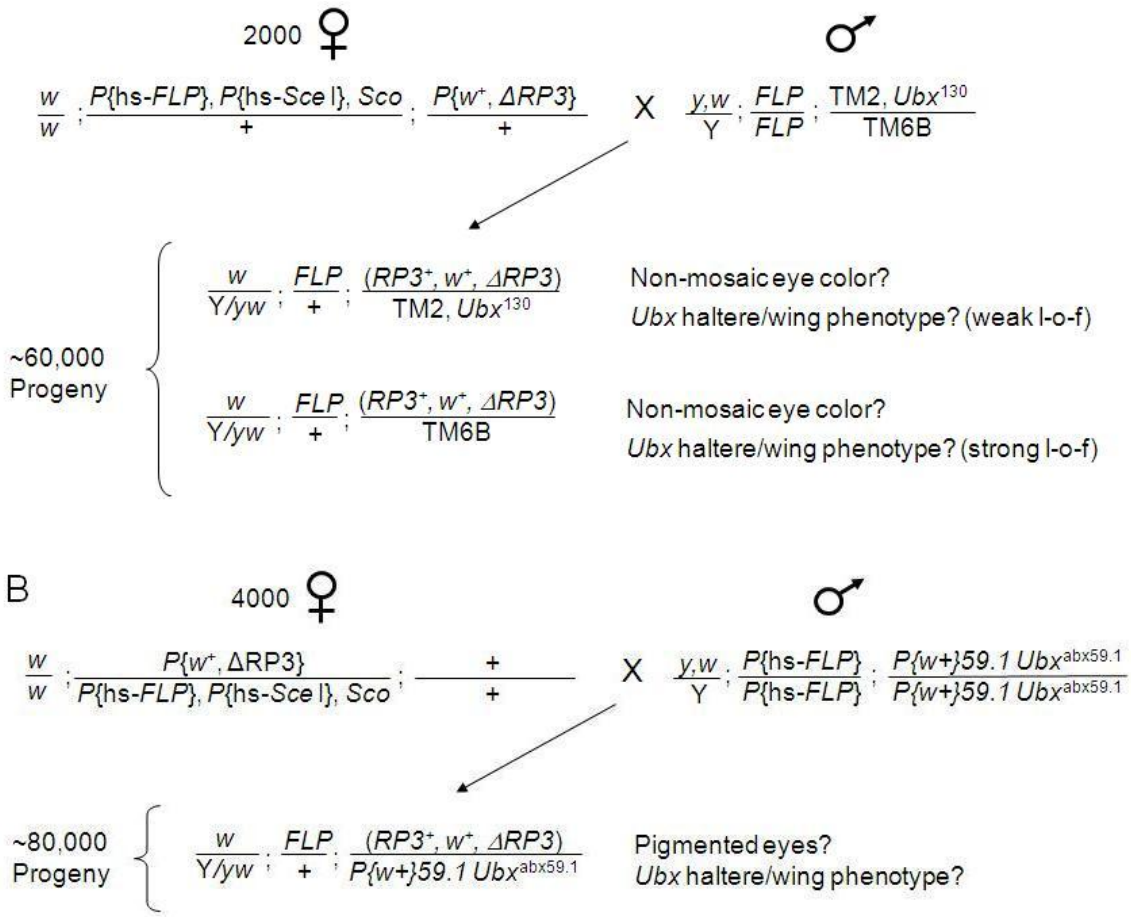


Figure 3.4. Screens for homolgous recombination targeting events. (A) Initial screening strategy. About 2,000 heat-shocked donor females were crossed to males that introduce third-chromosome balancers and a hs-*FLP* transgene with high basal expression (*P{hs-FLP}10*, abbreviated here as *FLP*). 60,000 progeny were screened for non-mosaic eye color and/or possible *Ubx* phenotype. A weak loss of *Ubx* function would be detected more easily over *TM2, Ubx*^{130} due to non-complementation by the null allele; a strong loss of function would be lethal over *TM2, Ubx*^{130} but viable over *TM6B* and should produce a dominant, weak but detectable haltere-to-wing transformation due to haploinsufficiency of *Ubx*. (B) Final modified screening strategy. We used non-complementation by a novel hemizygous viable weak *Ubx* allele (*abx*$^{59.1}$) to facilitate detection of weak or strong loss of *Ubx* function by the targeting event. In addition, *abx*$^{59.1}$ is associated with a *w*$^{+}$ insertion that produces very light eye color; due to the additive nature of *w*-dependent eye pigmentation, this was expected to facilitate detection of targeting events if they resulted in weak expression of the *w*$^{+}$ marker. Finally, a donor insertion on the second chromosome was used.

These results suggested that correct homologous recombination of the donor fragment into *Ubx* might both silence the w^+ marker and disrupt *Ubx* expression. However, the *Ubx* phenotype of Ubx^{DF35} could also be due to a second-site *Ubx* mutation, which would invalidate any further analysis based on this homologous recombination chromosome. Because we were screening for eye color and not for dominant *Ubx* phenotypes (which are relatively subtle), we might have missed many homologous recombination events, so we could not determine whether correct recombinations are obligatorily associated with loss of *Ubx* function and silencing of the w^+ marker. For this reason, we modified the screening strategy (Figure 3.4B). We changed the donor strain to M59.M1A, which had pTV2.ΔRP3 inserted on the second chromosome. This was chosen as the donor because *Ubx* resides on the third chromosome; thus successful homologous recombination should result in a change of linkage of the w^+ marker to the third chromosome. Additionally, the white-eyed or mosaic-eyed progeny following heat shock were mated to *y,w;P{hs-FLP};* $Ubx^{abx59.1}$. This allowed doubly sensitized screening for stably pigmented eyes (due to the weak eye color contribution by $Ubx^{abx59.1}$) and/or for a haltere-to-wing transformation (due to failure to complement *Ubx* $^{abx59.1}$). Because $Ubx^{abx59.1}$ is viable and fertile over a null *Ubx* allele (e.g. Ubx^{130} on TM2), we anticipated that even a strong reduction of function due to homologous recombination could be recovered over $Ubx^{abx59.1}$. Progeny with pigmented eyes and/or *Ubx* phenotypes were crossed to *w;TM2/TM6B* to balance the third chromosome and re-test for loss of *Ubx* function.

About 4000 white or mosaic eyed flies obtained after heat shock were mated for screening, producing about 80000 progeny. We recovered 20 independent candidate targeting events, with 6 candidates identified by a stably pigmented eye phenotype and 14 candidates identified by a *Ubx* phenotype. None of the stably pigmented phenotypes mapped to the third chromosome, so they were all due to non-targeted integrations. In contrast, the candidates based on a *Ubx* phenotype were all found to contain the correct duplication of the RP3 region due to homologous recombination with the donor fragment. All of these were recessive lethal and exhibited unpigmented eyes when first isolated, but in subsequent generations occasional heterozygous balanced progeny exhibited variable orange eye pigmentation. Most progeny of these pigmented-eyed flies also lacked eye pigmentation, which reappeared in subsequent generations. These results confirmed that correct homologous recombination results in strong but unstable silencing of the w^+ marker and in reduction of *Ubx* expression or function.

In summary, about one correct homologous recombination targeting event was recovered per 40 vials screened by the second strategy, similar to results reported by Rong and Golic 2002 for more straightforward targets where silencing of the w^+ marker was not observed and did not complicate the screens. The 14 targeting events from this strategy plus Ubx^{DF35} resulted in 15 correct targeting events from 120000 progeny screened.

PCR amplification from genomic DNA was used to verify correct recombination by amplifying the targeted locus in three sections spanning the

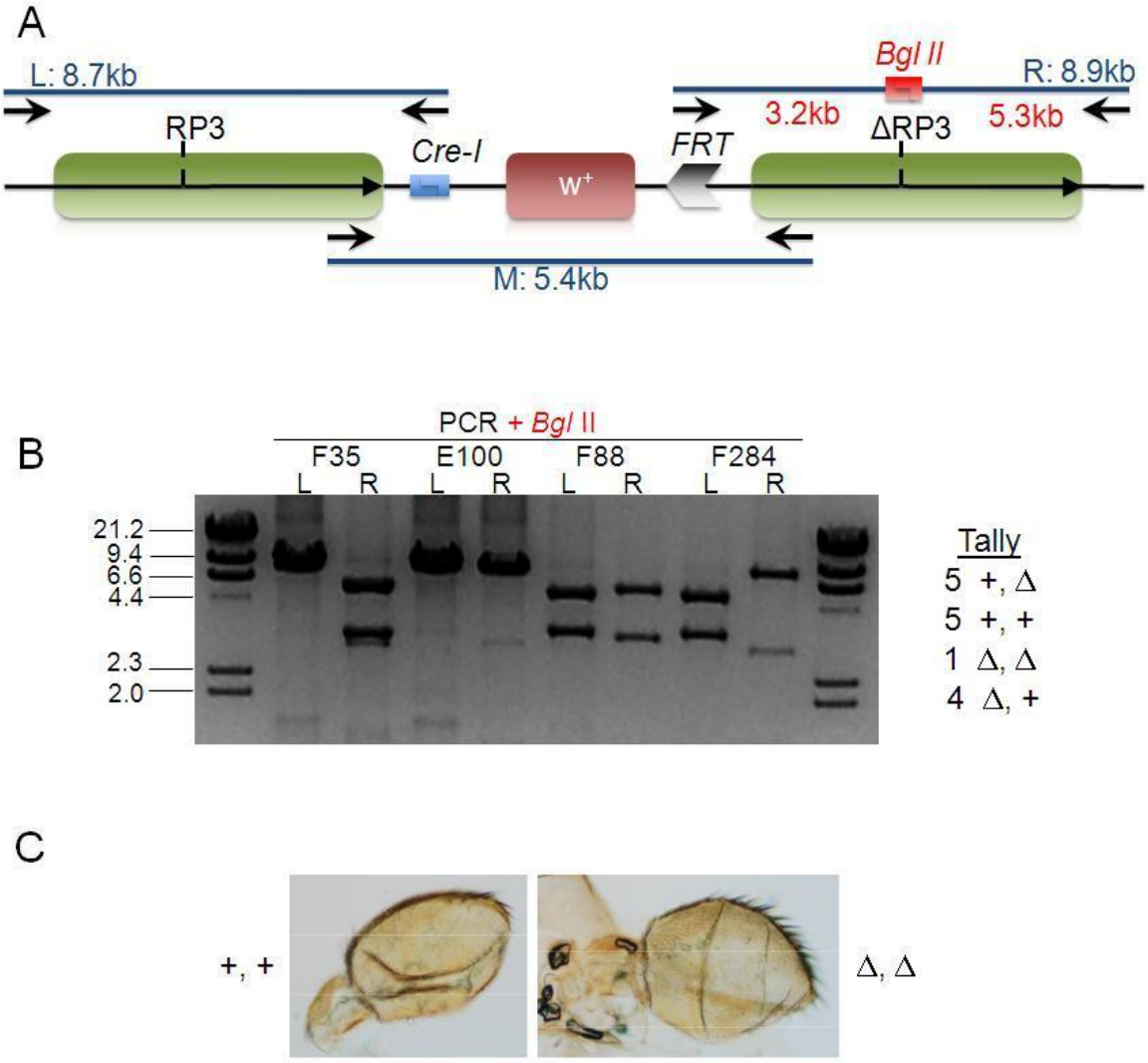


Figure 3.5. Molecular Verification of Homologous Recombination Targeting Events. (A) Schematic of PCR strategy to verify the targeting event. (B) Examples to illustrate the recovery of different RP3 genotype combinations in the left (L) and right (R) duplicated elements. (C) Comparison of a haltere from a (+,+)/+ fly against a haltere from a (Δ,Δ)/+ fly.

duplicated region (Figure 3.5A). The amplimers containing the left or right duplication element were digested with *Bgl*II to identify which element (if any) had the mutant (i.e. ΔRP3) allele. Because the I-*SceI* site lies downstream of RP3, we predicted that a wild-type RP3 (wt) site would lie in the upstream duplicated sequence, while the deleted RP3 site would lie in the downstream duplicated sequence (i.e. orientation should be WT-ΔRP3). However, out of the 15 correct targeting events, five were WT-ΔRP3, five were WT-WT, one was ΔRP3-ΔRP3, and four were ΔRP3-WT (Figure 3.5B). Table 2 lists all the correct targeting events and the orientation of the duplication with respect to mutant or wild-type RP3 alleles. In the WT-WT homologous recombinants, the *Ubx* phenotype was weaker, while the ΔRP3-ΔRP3 homologous recombinants had a stronger *Ubx* phenotype (Figure 3.5C). This suggested that RP3 contributes to *Ubx* function, at least in the context of these duplication/insertion alleles. On the other hand, the genotypes of the duplication regions had no effect on silencing of the w^+ marker.

Reduction of the Duplicated Sequence to Single Copy

Following verification of homologous recombination into the target locus, the targeted strains *w;Ubx*^{DF35}*/TM6B*, *w;Ubx*^{F456}*/TM6B*, and *w;Ubx*^{F641}*/TM6B* were used in the reduction step. *Ubx*^{DF35} is WT-ΔRP3; *Ubx*^{F456} is ΔRP3-WT; and *Ubx*^{F641} is also ΔRP3-WT. Reduction was induced by crossing these strains to *w;P{hs-CreI},Sb/TM6B*, which provides expression of *CreI* restriction enzyme under a heat shock promoter. The resulting *w;Ubx/P{hs-CreI},Sb* progeny

Table 2. List of correct homologous recombination targeting events and the orientation of the duplication genotypes with respect to the direction of *Ubx* transcription (left to right).

DF35	WT-ΔRP3
E100	WT-WT
F88	ΔRP3-ΔRP3
F109	WT-WT
F284	ΔRP3-WT
F567	WT-ΔRP3
F124	ΔRP3-WT
F40	WT-ΔRP3
F475	WT-WT
F456	ΔRP3-WT
F819	ΔRP3-WT
F641	WT-ΔRP3
F208	WT-WT
F411	WT-ΔRP3
F675	WT-WT

underwent heat shock at 36 degrees Celsius for one hour after aging 1-3 days. Adult progeny were then mated to *w;TM6B/MKRS* to establish balanced lines over *TM6B*. DNA was harvested from these balanced lines, and analyzed by PCR amplification over the target region and digestion with *Bgl*II (Figure 3.6, top). Depending on where the mutant allele is placed within the target sequence, reduction to the mutant allele can occur more than half of the time. Reduction to single copy was very efficient. Currently we have a ΔRP3 and a wild-type reduction from F456 (R15 and R16, respectively) and three ΔRP3 and three wild-type reductions from F641. The reduction chromosomes from F456 share recessive phenotypes consisting of rough eyes and disorganized triple row bristles along the anterior wing margin; these are unrelated to *Ubx* and must result from one or more mutations at other loci on the third chromosome. The reduction chromosomes from F641 share a recessive lethal phenotype also unrelated to *Ubx*.

We also obtained aberrant reductions that retained but reactivated the *white* marker. All of these exhibited loss of *Ubx* function and have dominant (i.e. haploinsufficient) *Ubx* haltere phenotypes. They result from deletions extending upstream of the integrated w^+ marker into the flanking duplication element and the *abx* region (not shown), but they have not been characterized further.

Effect of RP3 Deletion on *Ubx* splicing

Our first reductions were from the Ubx^{F456} parent chromosome, from which we isolated $Ubx^{\Delta RP3.R15}$ (R15) and its isogenic control Ubx^{+R16} (R16). R15 and R16

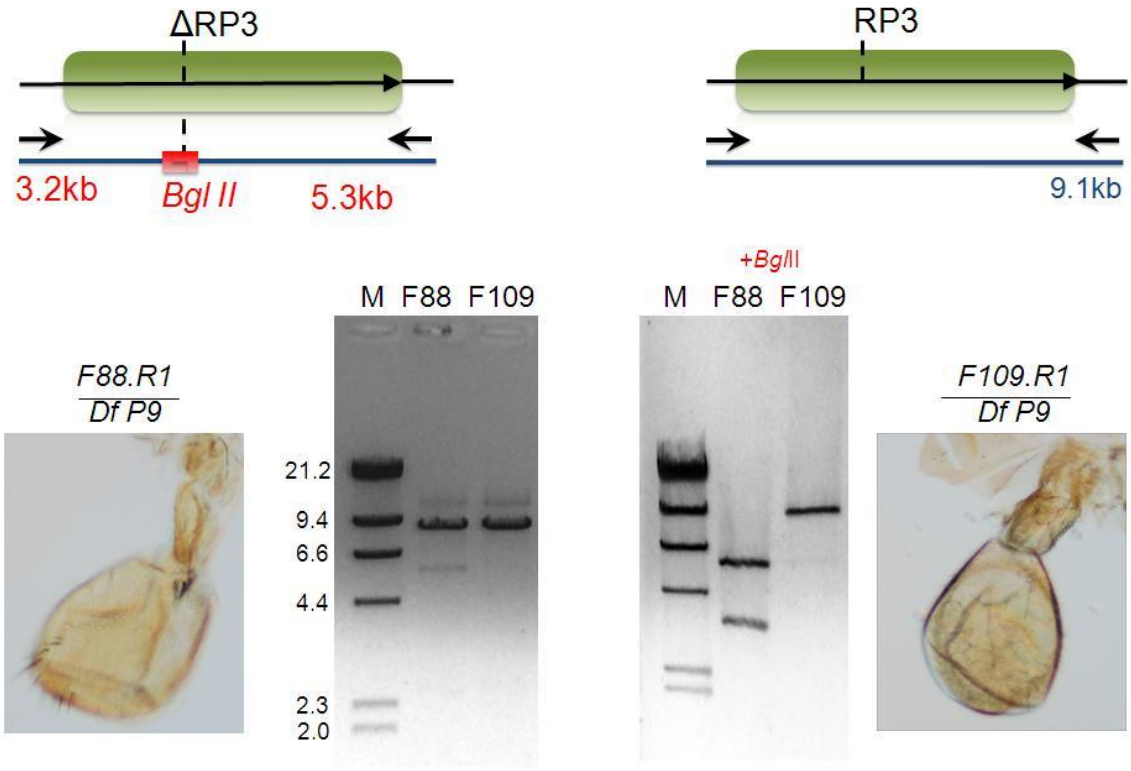


Figure 3.6. Molecular verification of duplication reduction to single copy. The left gel shows the PCR amplification of DNA from F88 reduction hemizygotes (ΔRP3/*Df P9*) and F109 reduction hemizygotes (+/*Df P9*).The right gel shows the same amplimers from the left gel after digestion with *Bgl*II. The haltere from a reduction to ΔRP3 (left) is compared to the haltere from a reduction to RP3+ (right).

are isogenic except for RP3. We have used these two strains as mutant/control pairs in subsequent experiments. An obvious question is whether RP3 is essential for accurate or efficient splicing *Ubx* transcripts. To test this, I analyzed the *Ubx* mRNA isoform ratios from embryos and larvae of R15 and R16 homozygotes using semi-quantitative RT-PCR. As expected, R15 did not produce an RP3 recursive intermediate at any stage, since it lacked RP3 (not shown). Both lines showed similar mRNA isoform ratios throughout embryonic development. However, there was a reproducible reduction in the ratio of isoforms Ia and IIa relative to isoform IVa during the first and second larval stages in R15 homozygotes (ΔRP3) compared to R16 homozygotes (wild type) (Figure 3.7). Qualitatively similar results were obtained in three technical and two biological replicates, although the shift in isoform ratios during the larval stages was even stronger than shown here in some replicates. The moderate change in mRNA isoform ratios was consistent with the very weak haltere phenotype (see below), because a strong disruption of this ratio in the same direction (as in Ubx^{MX17}homozygotes, which produce only isoform IVa) causes a strong transformation of haltere into wing (Subramanian et al, 1994). It has been shown previously that subtle changes in alternative splicing (10%) caused by mutations in trans-acting factors can be detected by semi-quantitative RT-PCR even though they produce haltere effects that are only detectable in double heterozygotes with null *Ubx* alleles (Burnette et al 1998).

Deletion of RP3 causes a weak loss of function *Ubx* Phenotype

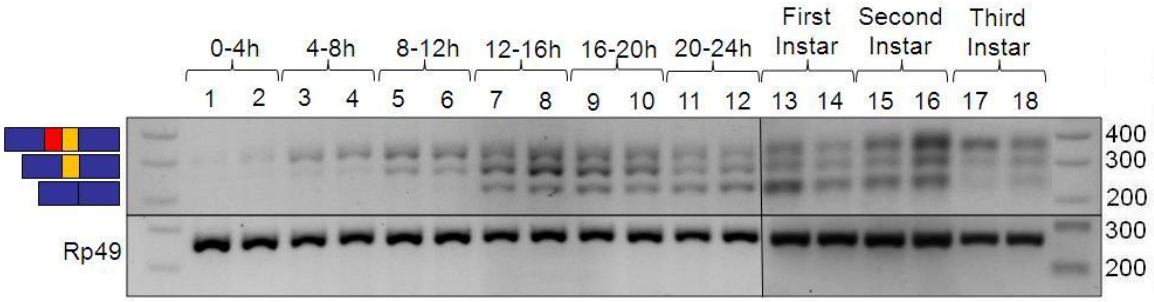


Figure 3.7. Expression of *Ubx* mRNA isoforms during development in *Ubx*$^{\Delta RP3\text{-}R15}$ homozygotes and *Ubx*$^{+R16}$ homozygotes. The R15 and R16 chromosomes are isogenic except for the ΔRP3 mutation in R15. *Ubx* mRNA was assayed every 4 hours for the first 24 hours of development, followed by the three instar stages collected at 32 hours, 54 hours, and 96 hours. For each developmental stage, R15 and R16 are displayed side-by-side, such that lane 1 is R15, lane 2 is R16, lane 3 is R15, lane 4 is R16, etc. The major *Ubx* isoforms Ia, IIa, and IVa are identified at the left. The *Rp49* mRNA was used as a quantitation standard. Qualitatively similar results were obtained in three technical and two biological replicates, although the shift in isoform ratios during the larval stages was even stronger than shown here in some cases.

ΔRP3 chromosomes exhibit a weak recessive *Ubx* haltere phenotype seen in transheterozygotes with diverse *Ubx* loss-of-function alleles. Unlike their isogenic wild-type control chromosomes, the ΔRP3 chromosomes enhance slightly the dominant haltere-to-wing transformations of null alleles *Df(3R)P9*, Ubx^{130}, and $Ubx^{9.22}$. *Df(3R)P9* is a deletion of the entire bithorax complex so that *Ultrabithorax*, *Abdominal-A*, and *Abdominal-B* are absent. Ubx^{130} is a null allele caused by a chromosomal inversion breakpoint within *Ubx*. $Ubx^{9.22}$ is a 1.6kb deletion that removes the last part of intron 3, the 3'ss, and first 48 codons of the homeodomain, inactivating all UBX protein isoforms. R15 and its isogenic wild type control R16 were also tested against $Ubx^{abx\text{-}2}$ (a 1.5kb deletion within the *abx* regulatory region in intron 3 about 1 kb upstream of RP3), $Ubx^{bx\text{-}34e}$ (an insertion of a *gypsy* element in inverse transcriptional orientation within the *bx* regulatory region in intron 3, which overlaps the *abx* region about 11 kb upstream of RP3), and Ubx^{MX17} (an 18kb inversion including mII that only expresses isoform IVa). R15 (ΔRP3) also produced weak haltere-to-wing transformations over these recessive partial loss-of-function *Ubx* alleles.

Unexpected phenotypes associated with deletion of RP3

R15, R16 and the F641 reductions 36-8 (ΔRP3) and 15-21 (WT) have also been tested against $Ubx^{Cbx\text{-}Hm}$, a complex rearrangement that has both gain of function in the mesothorax and loss-of-function in the metathorax and first abdominal segments. The gain-of-function component is due to ectopic expression of *Ubx* in wing precursor cells, transforming the wing blade into a copy of the haltere

capitellum (Gonzalez-Gaitan et al 1990), although the resulting ectopic haltere is about twice as large as the normal haltere. The loss-of-function component is due to reduction of *Ubx* expression in the posterior metathorax and first abdominal segment. This results in a recessive transformation of structures in this region towards a mesothoracic identity, including transformation of posterior haltere into posterior wing (Bender et al 1985). The $Ubx^{Cbx\text{-}Hm}$ allele thus provides a sensitized background to test for both loss and gain of *Ubx* function by ΔRP3 alleles. As in tests against other loss-of-function *Ubx* alleles, the ΔRP3 alleles enhanced the haltere-to-wing transformation slightly, but they also partially suppressed the gain-of-function phenotype, increasing the size of the ectopic haltere and restoring a more wing-like morphology in some cases. These effects were not seen with the corresponding wild-type isogenic controls. Both effects are consistent with loss of *Ubx* function by the ΔRP3 alleles. The partial suppression of the wing-to-haltere transformation suggests that this phenotype results not only from ectopic expression of the $Ubx^{Cbx\text{-}Hm}$ allele, but also from trans-activation of the wildtype allele in heterozygotes, which could occur by transvection or by positive feedback. Transvection can occur as a consequence of chromosome pairing-dependent interactions between promoters and regulatory elements on homologous chromosomes (Lewis 1954, reviewed by Duncan 2002). Transvection has been demonstrated for weaker Ubx^{Cbx} alleles although not for $Ubx^{Cbx\text{-}Hm}$ (Micol and Garcia-Bellido 1988). Positive feedback mediated by cell interactions has been proposed to explain non-clonal

Table 3.3 Viability of *Ubx*$^{\Delta RP3}$/*Ubx*$^{Cbx\text{-}Hm}$ *

	Paternal Genotype#			
Progeny Genotype	R15/R15	R16/R16	36.8/MKRS	21.5/TM6B
*/MKRS	183	192	108	97
*/Cbx-Hm	109	182	56	89
MKRS/Cbx-Hm	NA	NA	90	NA
TM6B/Cbx-Hm	NA	NA	NA	74
TM6B/MKRS	NA	NA	NA	68

* Males with the indicated genotypes were mated to virgin females of genotype *Ubx*$^{Cbx\text{-}Hm}$/*MKRS*. 36 males and 36 females were mated for each test; 8 pairs were placed in each of four food vials and allowed to lay eggs for 4 days. Progeny were removed and scored every eight hours until all pupae had eclosed)
# R15 and R16 are isogenic chromosomes, except R15 carries ΔRP3; 36.8 and 21.5 are isogenic chromosomes, except 36.8 carries ΔRP3.

inheritance of *Ubx* ON-OFF states in *Ubx* gain-of-function and loss-of-function mutants (Botas et al 1988).

Unexpectedly, the ΔRP3/Cbx^{Hm} genotypes exhibited strongly reduced viability (Table 3). This was surprising, since no reduction of viability was observed in heterozygosis over amorphic or hypomorphic *Ubx* alleles, nor even in hemizygosis over *Df(3R)P9*, which provides no *Ubx* expression or transvecting regulatory functions. Surviving ΔRP3/Cbx^{Hm} heterozygotes exhibited only weak haltere phenotypes that do not interfere with viability in other genotypes with similar haltere transformations, and the partially suppressed wing-to-haltere transformation did not affect pupal eclosion (all pupae eclosed), suggesting that lethality is due to defects in development or function of internal organs or structures. This would also be consistent with the observation that most ΔRP3/Cbx^{Hm} survivors exhibited poor mobility and had to be rescued from the food surface; this was not the case for the isogenic +/*Cbx*^{Hm} controls or even for *MKRS*/Cbx^{Hm}, which appeared nearly as robust as homozygous +/+ or +/MKRS. There are three possible explanations: (1) The reduced viability reflects a loss of function associated with ΔRP3 that synergizes with the gain-of-function phenotypes of Cbx^{Hm} to reduce survival (other *Ubx* loss-of-function alleles are known to be semi-lethal over Cbx^{Hm}). (2) It reflects a hypermorphic or neomorphic gain-of-function associated with the ΔRP3 alleles that enhances a previously undescribed gain-of-function effect by Cbx^{Hm}. (3) It reflects a loss or gain of function associated with ΔRP3 that synergizes with a non-*Ubx* mutation on the Cbx^{Hm} chromosome to produce synthetic lethality. It may be possible to

distinguish among these possibilities by testing against other *Ubx*^{Cbx} –type alleles, examining effects of derepression by mutations in negative regulators such as *Polycomb*, or by determining whether an interacting locus can be separated from Cbx^{Hm} by recombination.

Interaction with *RpII-215*^{C4}

Non-exonic RSS may also influence other processes that can be associated with splicing, for example promoting elongation by RNA Polymerase II, or minimizing genomic instability at long transcription units. To begin to test this, we examined ΔRP3 interactions with two alleles of the RNA Polymerase II 215kD subunit gene (*RpII215*) that have been reported to exhibit weak *Ubx* phenotypes. The C4 allele contains a substitution of arginine by histidine at position 741 (R741H), which is located in the funnel region that normally provides access to substrate nucleotides and to the TFII-S endoribonuclease, a factor that helps overcome blocks to elongation. The RPII215 subunit encoded by this mutant allele has reduced processivity, reduced ability to overcome elongation blocks and reduced ability to respond to TfIIS (Coulter and Greenleaf 1985). The *C4* allele confers resistance to alpha-amanitin, a molecule that inhibits processivity of RNA polymerase II by binding to the bridge helix in the funnel region (Chen et al 1993). *RPII215*^{C4} is homozygous and hemizygous viable, but in heterozygotes with a wildtype subunit it produces a neomorphic weak *Ubx*-like phenotype consisting in a transformation of haltere towards wing similar to that produced by heterozygotes for null *Ubx* alleles (Greenleaf et al 1980, Mortin and Lefevre

1981). *RPII215*^{C4} also enhances the haltere phenotypes of many *Ubx* loss-of-function genotypes, particularly when heterozygous over a wild-type *RP215* allele. The requirement for a wild-type allele has been interpreted in terms of interference between enzymes with normal and compromised processivity as they transcribe the same template.

The Ubl allele of *RpII215* contains a nucleotide substitution (G4471A) that changes the amino acid at position 886 from aspartate to asparagine (D886N). The biochemical effect of this mutation is not known and it does not cause an elongation or processivity defect (Coulter and Greenleaf, 1985), but it produces a dominant *Ubx*-like phenotype (hence its name) and also enhances the phenotypes of *Ubx* mutants. However, its genetic behavior is distinct from *RIIP215*^{C4} and appears to be a straightforward antimorph: Ubl is lethal as a homozygote or hemizygote but behaves as a dosage-sensitive enhancer of *Ubx*, with more copies of Ubl relative to wild-type creating a stronger effect. In contrast, null alleles of RPII215, which are recessive lethal, do not have dominant *Ubx*-like phenotypes nor do they enhance the phenotypes of *Ubx* mutants.

We crossed R15 (ΔRP3) and R16 (isogenic WT) strains with *RPII215*^{C4}, as well as with a strain containing *RPII215*^{C4} and one of several other *Ubx* alleles to sensitize for *Ubx* phenotypes. The C4/+; R15/*Ubx* flies exhibited synergistic enhancement of the haltere phenotype that was not observed with R16 flies (Figure 3.8). Similar tests with the *RPII215*^{Ubl} allele did not show a difference between R15 and R16. These results suggested a differential sensitivity of R15 (ΔRP3) to the processivity defect of *RPII215*^{C4}. However, the effects of the Ubl

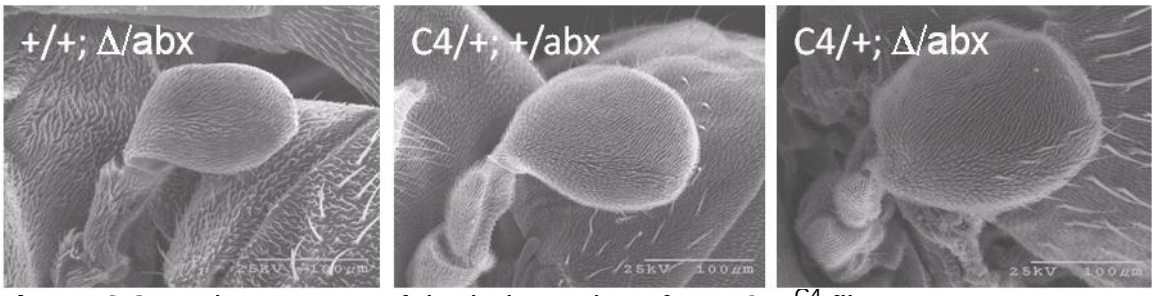


Figure 3.8. Enhancement of the haltere size of *RPII215*^{C4} flies.

allele were very weak in all cases, which was unexpected given that literature reports (Mortin and Lefevre 1981) have described the Ubl phenotype and Ubl x *Ubx* interactions as much stronger than the C4 phenotype and C4 x *Ubx* interactions.

We re-examined the comparison between C4 and Ubl with additional *Ubx* alleles, including hypomorphs Ubx^{MX17}, and Ubx^{abx2} as well as deletion null allele *Df(3R)P9* and inversion null Ubx^{130}, which had originally been used to describe the Ubl effect as stronger than C4. In all cases, Ubl produced a much weaker interaction than C4, and C4 a stronger effect than expected. The same was true for two C4 and two Ubl strains obtained from two different sources that have been separated by nearly two decades, including the Bloomington Stock Center. The identities of the C4 and Ubl strains were verified by associated markers and by complementation analyses against null or temperature-sensitive *RPII-215* alleles: C4 is viable over all alleles except Ubl and produces a weak *Ubx*-like phenotype when heterozygous over WT but not when homozygous or hemizygous; Ubl is lethal over all except WT). Recombination against a wild type chromosome, verified by exchange of flanking markers, did not enhance the effect of Ubl, suggesting that the difference from expected behavior is not due to accumulation of an extragenic suppressor. Currently we have no explanation for the discrepancy.

R15 (ΔRP3) and R16 (isogenic wild type) were also tested against a TFIIS loss-of-function allele ($TfIIS^{2}$). TFIIS is a processivity factor that cleaves the nascent RNA to generate a new 3'OH at the active site, allowing paused

polymerases to resume elongation (Chen et al 1996). As noted above *RPII215*^{C4} has reduced responsiveness to TfIIS. *TFIIS*^{2}*/+;R15/R15* flies showed a slight enhancement of the haltere, compared to *TFIIS*^{2}*/+;R16/R16* flies (not shown).

DISCUSSION

We have used two-step targeted gene replacement by ends-in homologous recombination successfully to delete non-exonic RSS RP3 from the *Ubx* gene in Drosophila. Although the phenotypic and molecular analysis of these mutants is still ongoing and will require much more detailed experiments, the results so far indicate that deletion of RP3 results in a modest reduction of *Ubx* function with respect to haltere development (which represents a convenient but limited aspect of *Ubx* function) and a more significant alteration with respects to other *Ubx* functions that are still uncharacterized. The first steps for deletion of non-exonic RSS in frizzled and polychaetoid have been accomplished and completion of the gene replacements is under way.

Interestingly, the initial targeting step in *Ubx* itself had the effect of impairing *Ubx* function significantly due to the insertion of *w*+ and/or the duplication of the target region, even though all of these alterations were within a 50kb intron. Although the homologous recombination method would predict that the duplicated element with a wild-type RP3 would be upstream of the element with deleted RP3 (Figure 3.1), we obtained all possible permutations of deleted and wild-type RP3 among the duplicated elements. This has been seen before in other cases of ends-in homologous recombination (Dolezal et al 2003, Elmore et al 2003, Lankenau et al 2003, Xie and Golic 2004). These different arrangements probably arise by gene conversion occurring during or after double stranded break repair by homologous recombination. Recovery of these different

genotypes provided an early indication that RP3 plays a role in *Ubx* expression, as homologous recombinants that had RP3 deleted in one or both duplication elements had a stronger haltere-to-wing transformation phenotype than recombinants with intact RP3 in both duplication elements.

Analysis of the final ΔRP3 mutants failed to reveal any obvious qualitative or quantitative defect in splicing of the *Ubx* transcripts during embryonic development. However, there is a decrease in the steady-state ratio of isoform Ia and IIa relative to IVa during larval development. This could be due to a defect in the accuracy of splicing caused directly by the deletion, or to a change in transcription efficiency of *Ubx* in specific tissues leading indirectly to a change in steady-state isoform ratios in the whole animal, as IVa is specific to the central nervous system (Lopez and Hogness, 1991; Bomze and Lopez, 1994; Subramaniam et al, 1994; Lopez et al 1996). More detailed anlayses of tissue specific transcription and isoform ratios will be necessary to distinguish these possibilities, particularly as other hypotheses can explain how organisms can splice accurately and efficiently over long distances, for example through the formation of long-range RNA secondary structures mediated by pairing of repetitive elements (Shepard et al. 2009), and association of exons to the C-terminal domain of transcribing RNA Polymerase II (Dye et al. 2006).

Complications encountered during the homologous recombination screen and the effects of aberrant reductions to single copy revealed profound silencing of marker gene expression around the RP3 locus, presumably as a consequence of repressive chromatin structures that may be related to the function of RP3.

Every one of the 16 insertions of the donor fragment by correct homologous recombination into the RP3 exhibited profound silencing of the *white* marker and exhibited a dominant *Ubx* phenotype, and this silencing could be abolished by aberrant reductions that retained the white marker but generated deletions extending upstream. This is notable since a number of mobile element insertions into other areas in the *Ubx* gene and promoter region are not silenced (Bender and Hudson 2000, McCall et al 1994), suggesting that the effect is localized to the RP3 region. The *abx* regulatory domain is known to be about 1kb upstream of RP3; this element enhances the expression of *Ubx* in imaginal discs and the embryo. Deletions in the *abx* region cause transformations of the third leg towards the second leg and the haltere towards wing, with varying intensity (Peifer and Bender 1986). This region also contains sequences that serve as insulators and as targeting elements for negative regulation of *Ubx* by the Polycomb group genes. The targeting step of ends-in homologous recombination duplicates this region so that two copies flank the *white* marker. No obvious gain-of-function *Ubx* phenotypes seem to accompany the activation of the *white* marker by the aberrant reductions, although the reductions retain a loss-of-function haltere-to-wing transformation. This suggests that silencing of *white* results from the duplicated element (either because of its position relative to *white* or because of interaction with the element on the other side of *white*) and that the single remaining copy after aberrant reduction maintains proper negative regulation of *Ubx*. Fine mapping of the deletions in the aberrant reductions using

PCR would elucidate the nature and location of the cis-elements controlling the silencing.

Given the interaction with *RPII215*^{C4}, deletion of RP3 may impair transcriptional elongation through the long intron. Splicing factors have been shown to associate with the C-terminal domain of RNA Polymerase II (review by Muñoz et al 2010). Splicing can be coupled to transcription (Das et al 2007) and can stimulate transcription (Fong and Zhou 2001). It will be important to examine *Ubx* RNA levels using quantitative assays in ΔRP3 homozygotes in the presence and absence of a C4 allele.

In contrast to C4 the Ubl allle of *RpII-215* did not exhibit an interaction with ΔRP3, suggesting that RP3 is sensitized specifically to a compromise in transcriptional processivity. Ubl and C4 both interact with known loss-of-function alleles of *Ubx*, such as *Ubx*^{130}. The mechanism by which Ubl produces *Ubx*-like phenotypes and enhances *Ubx* phenotypes is unknown (Chen et al 1993), but it is hypothesized that the effect is due to a reduced response of RNA polymerase to UBX proteins, which are transcriptional regulators. For C4, the deficient processivity of the polymerase is thought to cause the *Ubx* effect, but the mechanism has not been investigated. The specific enhancement of the R15 *Ubx* phenotype by C4 as opposed to Ubl suggests that the interaction is due to interplay between the processivity defect of the polymerase and some elongation impairment caused by ΔRP3, rather than a reduction in UBX levels directly by ΔRP3. However, it is also possible that the processivity defect of C4 enhances the splicing defect in ΔRP3 by altering the kinetics of splicing across long intron

sections. Many factors contribute to efficient transcriptional elongation (reviewed by Selth et al 2010 and Sims et al 2004), including the association of elongation factors with splicing factors and chromatin remodelers. Additionally, chromatin remodeling also contributes to efficient elongation, but it can also recruit splicing factors (reviewed in Luco and Misteli 2011). Chromatin-binding adaptor proteins associate with splicing factors, and can affect alternative splicing.

The deletion of a RSS is the first step to elucidating the function of recursive splicing. Further analysis of *Ubx* ΔRP3 alleles and additional RSS deletions in other genes should clarify the significance of our initial observations.

Chapter 4: Recursive Splicing in Mammalian Introns

ABSTRACT

Recursive splicing has not been demonstrated in humans or other mammals. However, bioinformatic analyses of the human genome identify elements that resemble Drosophila recursive splice sites (RSSs), and there is partial evidence for recursive splicing of an NMD-triggering cassette exon in rat alpha-tropomyosin. I tested the function of a sample of predicted RSSs from human using the same approaches as in Drosophila: (1) verify the use of the RSS as a 3'ss by assaying for the recursive intermediate, (2) test for the use of the RSS as a 5'ss by assaying for the corresponding splicing lariat, and (3) test for the use of the regenerated 5'ss by mutating the 5'ss component of the RSS and assaying for a change in exon junction as a consequence of a shift to use of alternative or cryptic 5' splice sites. These tests made use of a recursive splicing reporter system or minigenes transfected into human cell lines. For three out of eight human RSS candidates tested, I was able to detect the predicted recursive intermediates and a shift to use of an alternative 5'ss after mutation of the RSS 5'ss motif. Analysis of lariats has not succeeded in this experimental system, where I have been unable to detect lariats for either recursive or direct splicing. I also validated a RSS associated with a novel ORF-truncating cassette exon (E3b) located in intron 3 of the human dopamine reuptake transporter gene *SLC6A3*. We verified alternative splicing of exon E3b in endogenous *SLC6A3*

transcripts in the *substantia nigra* of adult human brain and in minigene transcripts in a transfected neuronal cell line. E3b is flanked by single-nucleotide polymorphisms (SNPs) that appear to be associated with differential risk for schizophrenia. We found that the risk-associated haplotype increases the inclusion of E3b in cell transfections assays. Thus, the risk-associated haplotype might be associated with reduced expression of the dopamine reuptake transporter in vivo as a consequence of altered splicing, and this might underlie or exacerbate dopaminergic dysfunction.

INTRODUCTION

The existence of recursive splicing in mammals remains an open question. What we now know as recursive splicing in Drosophila was first proposed as the mechanism of alternative splicing for a cassette exon in the Adenosine Monophosphate Deaminase 1 (*AMPD1*) gene of rat (Mineo et al 1990). This exon held particular interest because variation in its inclusion was linked to variation in severity of muscle degeneration caused by mutations in human *AMPD1*. However, the hypothesis of recursive splicing was based solely on the observation that the pre-formed exon-exon junction at the 5' end of the cassette exon in an *AMPD1* minigene could function (albeit weakly) as a 5'ss in cell transfection experiments. Subsequently, it was shown that this is not the major mechanism for alternative splicing of native *AMPD1* transcripts in vivo, which instead involves exon skipping as a consequence of a weak 3'ss for the cassette exon (Mineo et al 1991).

More recently, stronger evidence of mammalian recursive splicing has been reported for the alpha-tropomyosin gene of rat (Grellschied and Smith, 2006). In this case, a novel cassette exon was discovered between exons E3 (itself a cassette exon) and E4. The novel cassette exon is included in mRNA only when it is spliced to exon E2 (i.e. when E3 is skipped); when it is spliced to E3, it appears to undergo recursive splicing and is removed from the mRNA along with the downstream intron. In this case the evidence for recursive splicing

included detection of the predicted recursive intermediate in vivo as well as the ability of the preformed junction to function as a 5'ss, but the data did not rule out the possibility that the putative recursive intermediate might actually be a dead-end product.

The evidence for recursive splicing in Drosophila examples is far more robust and involves three types of tests (Hatton et al 1998; Burnette et al 2005; Papasaikas 2010): (1) in vivo detection of the predicted intermediates for both the first and second steps of recursive splicing (together with absence of the intermediates predicted for direct splicing); (2) 5'ss activity of the preformed exon-exon junctions in minigenes, and (3) the effects of mutations that block regeneration of the functional 5'ss. No proposed example of recursive splicing in mammals has yet been submitted to all three tests. In part this is the result of the technical difficulty of some tests, particularly the analysis of lariats and recursive intermediates in the context of the larger and more complex mammalian genomes. This context also makes it difficult to generate high-confidence predictions of recursive splice sites in mammals, particularly in the absence of a reliable model for the core motif and associated elements based on well-validated examples, as in Drosophila.

Arguments For and Against Mammalian Recursive Splicing

The distinctive association of recursive splicing with long introns in Drosophila (Chapter 1; Burnette et al 2005; Papasaikas et al 2010) would suggest that recursive splice sites should also be found in mammalian genomes, which

contain a higher proportion of very large introns. About 10% of human introns span more than 10kb (Deutsche and Long 1999), and over 3000 are larger than 50kb. In Drosophila, every intron over 50kb contains at least one recursive splice site (Papasaikas et al 2010). In contrast, Singh and Padgett (2009) analyzed a 107 kb human intron and failed to find evidence of recursive splicing in the kinetics of accumulation of nascent RNA during transcription. On a larger scale, published bioinformatic analyses have failed to detect an overrepresentation of predicted recursive splice site motifs in large introns of mammals above random expectation (Shepard et al 2009). The authors concluded that recursive splicing does not occur in mammals, and they argued further that they are rendered unnecessary by proposing that the accurate pairing of splice sites across large mammalian introns is facilitated by long-range secondary structures formed by repetitive elements.

There are several problems with these arguments. First, the analysis of Shepard et al (2009) used a model based on juxtaposition of standard 3' and 5' splice site motifs; even in Drosophila this *ad hoc* model has much lower information content than the real recursive splice site core motif (Papasaikas et al 2010), and the problem of distinguishing real sites from background would be even greater in mammals given the larger genomes and introns. Furthermore, the role of auxiliary cis-acting elements might also be expected to be even more important than in Drosophila (Papasaikas et al 2010). The second problem resides in taking for granted that the function of recursive splice sites is to facilitate the accurate splicing of long introns. This was an obvious early

hypothesis (Burnette et al 2005), but it is not yet supported by experimental data in Drosophila (see Chapter 2). The specific association of RSSs with long introns of Drosophila might be dictated by other features of gene function, regulation, or chromatin structure that might correlate strongly with intron size in flies but not in mammals.

More sophisticated bioinformatic analyses with a mammalian recursive splice site model generated by semisupervised learning (EMSS model) did reveal an overrepresentation of RSSs in mammalian introns longer than 10 kb, but the magnitude (~2-fold) was much smaller than in Drosophila (Papasaikas et al 2010). The difference may be due to lower sensitivity and/or discrimination by the mammalian EMSS model, which was initiated with a hypothetical model rather than known examples and had not been refined by experimental validation. In addition, the core motif may, as for regular splice sites, not be sufficient for definition of authentic RSSs without additional cis-auxiliary elements. An additional possibility is that mammalian intron size might be correlated less strongly with the features that confer a selective advantage on recursive splicing. Comparing the frequency, distribution and context of recursive splicing in mammals and Drosophila could thus shed light on its biological functions.

Functionally Distinct Elements that Resemble RSS Sequences

A further complication in the prediction and analysis of recursive splicing in mammals is the existence of another class of elements that also resemble juxtaposed 3' and 5' splice sites. These are "dual specificity" splice sites (Zhang

et al 2007). They differ from RSSs in that they can function as either 3' or 5' splice sites to mediate alternative splicing, but they do not function sequentially as both. However, this functional difference from Drosophila RSSs is reflected in the sequences of dual specificity splice sites, which correspond to weak versions of both the regular 3' and 5'ss motifs and define a loose consensus (Zhang et al 2007). Thus, they are ambiguous, unlike Drosophila RSSs where both the 3' and 5'ss motif components are very strong and the 3'ss component and branch site are further enhanced (Papasaikas et al 2010). The ambiguity of dual specificity splice sites probably makes them inherently inefficient and allows them to be modulated easily by trans-acting factors to control alternative splicing. They appear to be only weakly conserved. Zhang et al (2009) reported the prediction of numerous dual-specificity splice sites in Drosophila but did not provide data. It is possible that some weak recursive splice site predictions might function as dual-specificity splice sites. However, these would be located at the boundary between alternative exons that are immediately adjacent in the genome sequence.

Testing for Recursive Splicing in Human Genes

In this chapter, I test a number of recursive splice sites that were predicted with the human EMSS model. I verify that several of these function as predicted, first as a 3'ss and then as a 5'ss. In the course of validating the RSS predictions, I was also able to show that a recursive splice site is associated with a novel cassette exon (dubbed "E3b") of the human *SLC6A3* gene, which encodes the

dopamine reuptake transporter DAT (Talkowski et al 2011). E3b is predicted to trigger nonsense-mediated decay of *SLC6A3* mRNA, and this could have important consequences for modulation of dopaminergic signaling with implications for susceptibility to schizophrenia and other neurological disorders. In previously published collaborative work (Talkowski et al (2011) we have found that retention of E3b in *SLC6A3* mRNA is altered by flanking single-nucleotide polymorphisms (SNPs) that have been associated with differential risk for schizophrenia in two human population samples (Talkowski et al 2008; Talkowski et al 2011).

MATERIALS AND METHODS

Prediction of Mammalian Recursive Splice Sites

A mammalian recursive splice site motif was constructed using the same previously described semi-supervised learning methods that were used for Drosophila (Papasaikas et al, 2010), except that the initial round of semi-supervised learning used an *ad-hoc* model consisting of the juxtaposed motifs for human 3' and 5' splice sites rather than experimentally identified RSSs. In addition, repetitive elements such as *Alu* were masked from the genome sequence during semi-supervised learning. The final EMSS model was then used to predict and rank mammalian recursive splice sites in the genome.

Construction of pMRSR and RSS minigenes

The Mammalian Recursive Splicing Reporter (MRSR) vector was created by cloning a fragment from the human Coagulation Factor VII Precursor gene (*hF7*), containing the last 156 bp of exon 7, all of intron 7, and the first 194 bp of exon 8 and fusing it to the constitutive CMV promoter and SV40 polyadenylation site in pCMV-Script (Invitrogen). For this purpose an 1117 bp fragment of *hF7* was amplified from HeLa genomic DNA by high-fidelity PCR with Phusion DNA Polymerase (New England Biolabs) using primers hF7.F1 and hF7.B1 (Supplementary Table 1). After gel purification the amplimer was digested 20 bp from the 5' end with *Sac* I and ligated between the *Sac* I and *Eco* RV sites of pCMV-Script. The 5' splice site for *hF7* exon 7 was converted into a perfect

consensus site (CAG/GTGAGT) by PCR-mediated site-directed mutagenesis. Unique sites for *Xba* I and *Eco* RV were introduced at 316 bp and 328 bp downstream of the 5' splice site for *hF7* intron 7 by PCR -mediated site-directed mutagenesis using primers MRSR.Xba.R1 with MRSR.F1, and MRSR.RV.F1 with hF7.B1. These amplimers were digested with *Sac* I and *Xma* I, and used in a three-part ligation between the *Sac* I and *Xma* I sites in pMRSR.

Plasmids pMRSR.CDH4, pMRSR.BACH2, pMRSR.MAPKAPK2, pMRSR.GRK5, pMRSR.SLIT3, pMRSR.ACTN4, pMRSR.RSU1, and pMRSR.SLC6A3.RSS2 were constructed by amplifying a fragment containing the predicted recursive splice site using primers containing *Xba*I and *Eco*RV sites and ligating it into the *Xba*I and *Eco*RV sites in pMRSR. Each RSS was amplified from HeLa genomic DNA by high-fidelity PCR with Phusion DNA polymerase using primers listed in Appendix Table, followed by gel purification and digestion with *Xba*I and *Eco*RV. pMRSR.CDH4 contains a 889 bp fragment extending 390 bp upstream and 495 bp downstream of the RSS. pMRSR.BACH2 contains a 700 bp fragment extending 375 bp upstream and 325 bp downstream of the RSS. pMRSR.MAPKAPK2 contains a 824 bp fragment extending 499 bp upstream and 325 bp downstream of the RSS. pMRSR.GRK5 contains a 787 bp fragment extending 278 bp upstream and 509 bp downstream of the RSS. pMRSR.SLIT3 contains a 983 bp fragment extending 491 bp upstream and 492 bp downstream of the RSS. pMRSR.ACTN4 contains a 700 bp fragment extending 298 bp upstream and 402 bp downstream of the RSS. pMRSR.RSU1 contains a 901 bp fragment

extending 362 bp upstream and 539 bp downstream of the RSS. pMRSR.SLC6A3.RSS2 contains a 925 bp fragment extending from 444 bp upstream through 481 bp downstream of predicted exon E3b from *SLC63A*.

Construction of pCMV.DAT.E3-E4 risk and non-risk variants

To generate an *SLC6A3* minigene to study splicing of predicted recursive cassette exon E3b in its native intron context, I amplified a 9 kb fragment of *SLC6A3* containing the last 57 bp of exon 3, all of intron 3 (including predicted exon E3b), and the first 232 bp of exon 4 and cloned this into plasmid pCMV, fusing Exon 3 to the constitutive CMV promoter and Exon 4 to the SV40 polyadenylation site in the vector. The *SLC6A3* fragment was amplified from genomic DNA by high-fidelity PCR with Accuprime HiFi DNA Polymerase (Invitrogen) using primers hDAT.E3.R1.F1 (containing an *Eco* RI site) and hDAT.E4.X1.B1 (containing an *Xho* I site). The purified fragment was digested with *Eco* RI and *Xho* I and ligated between the corresponding sites of pCMV-Script. Separate constructs were made with equivalent *SLC6A3* fragments derived from non-risk and risk-associated haplotype sources. Non-risk haplotype DNA sources were HeLa and CEPH DNA sample GM-12155. Risk haplotype sources were CEPH DNA samples 11881, 12154, 12249, 12760 and 12892. CEPH DNA samples were obtained from Coriell Cell Repositories. Constructs were verified by sequencing.

Analysis of construct splicing by transfection into HeLa cells

HeLa cells were transfected at 70% confluence in 35 mm dishes with 0.4 ug supercoiled plasmid DNA using Effectene Transfection Reagent (Qiagen). Total RNA was extracted after 48 hours using Trizol (Invitrogen) and resuspended in 100 uL of RNase-free water. Reverse transcription was performed with 50 units of Superscript II (Invitrogen) using 50 ng of random hexamer primers and 1 ug of RNA in a total volume of 20 uL. After treatment with RNase H, $1/10^{th}$ of the cDNA was amplified for 20-30 cycles (95°, 30 sec; 55°-58°, 30 sec; 72°, 30 sec) in a 25 uL reaction containing 10 pmol of each primer, 1.5 mM $MgCl_2$, and 0.5 units of Platinum Taq. To analyze the mRNA from pMRSR and derivatives, primers MRSR.F1 and hF7.B1 were used in a 30-cycle reaction. To analyze intermediates, primers MRSR.F1 and a reverse primer from Appendix Table were used in a 35-cycle reaction. To analyze mRNA from the pCMV.SLC6A3.E3-E4 constructs, primers pCMV.F1 and pCMV.R1 were used in a 30-cycle reaction, followed by dilution of the amplimers 1:164 in water and subjecting 2 uL of this to 20 cycles of reamplification using hDAT.E3F1 and hDAT.E4R2. A 5 uL sample of the PCR reaction was analyzed by electrophoresis through 2% agarose with GelStar fluorescent stain (Cambrex).

RESULTS

Generation of a Reporter Minigene to Study Mammalian Recursive Splicing

In order to test mammalian RSSs, we generated a mammalian equivalent of the minigene reporter system used previously in Drosophila (Burnette et al 2005, Papasaikas et al 2010). The minigene system allows the detection of processed RNAs and intermediates that are produced during recursive splicing, and it also provides a suitable platform for mutagenesis tests. As the Drosophila minigene system itself is unsuitable for testing recursive splicing in mammalian cells, we constructed a mammalian equivalent by taking advantage of the unique gene structure of the human coagulation factor 7 (*hF7*) gene. The *hF7* gene contains six 37 bp repeats at the 5' end of intron 7. Each repeat contains a weak 5'ss, and all of these are reportedly in pure competition for splicing of exon 7 to exon 8, without modulation by additional cis-acting elements (Borensztajn et al 2006). We amplified a fragment spanning the 3' portion of exon 7 through the first two repeats of intron 7 and ligated this to a fragment containing the 3' end of intron 7 and the 5' end of exon 8 (Figure 4.1A). This reconstructed region was then inserted into pCMV-Script (Figure 4.1B). The plasmid now had a reduced intron between exons 7 and 8, with the upstream exon containing two competing 5' splice sites. We mutated the distal 5'ss to match the consensus sequence (GTRAGT). The resulting plasmid (named pMRSR) now primarily splices using the distal 5'ss, only using the proximal 5'ss 10% of the time (Figure 4.3 lane 1).

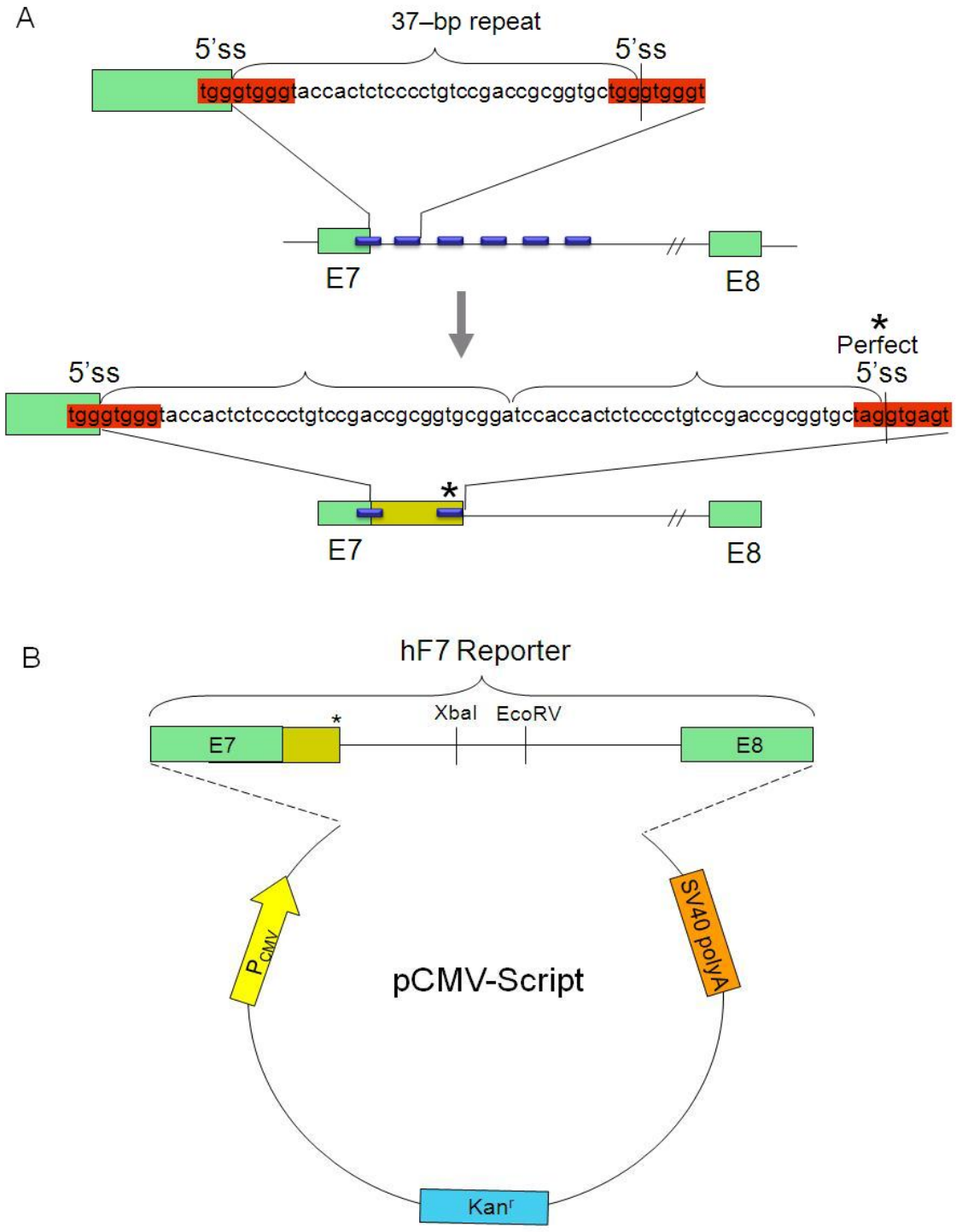


Figure 4.1. Generation of mammalian recursive splicing reporter vector pMRSR. (A) Diagram and sequence of the hF7 repeat used in creating the competing 5' splice sites mimicking the Drosophila minigene. (B) Diagram of the hF7 fragment ligated into pCMV-Script to create pMRSR.

RNAs from this construct are not translatable, so alternative splicing should not be confounded by nonsense-mediated decay.

Selection of Human Predicted RSSs for Testing

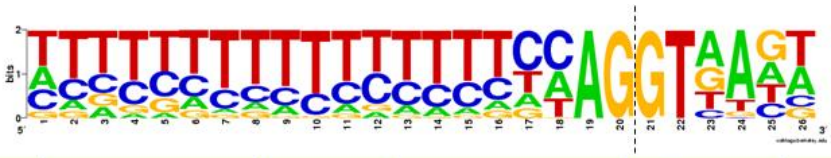
A human recursive splice site model was generated using similar semi-supervised learning methods as were used to generate the EMSS-RSS model for Drosophila (Papasaikas et al 2010) (Figure 4.2). We used this model to search and rank possible RSSs in the human genome. We chose six predicted RSSs from different genes to test experimentally. They were located in *ACTN4, BACH2, SLIT3, GRK5, RSU1*, and *MAPKAPK2*. These six RSSs were chosen because they were highly similar to the predicted consensus motif, they were conserved in other mammals (Figure 4.2), and they were located within introns larger than 10 kb. These six RSSs were also chosen because they were the only predicted RSSs in their respective introns, which facilitates their analysis. A seventh RSS (in *CDH4*) was chosen although it was not the only predicted RSS in its intron because previous data suggested that it might function in mouse (data not shown).

Experimental Tests of Human RSSs

Each predicted RSS was cloned into pMRSR and tested for the use of the RSS using the methods previously described by Burnette et al (2005). First, I tested the effect on mRNA production of inserting the candidate RSSs into the pMRSR intron (Fig. 4.3). Using the primers MRSR.F1 and HF7.B1 for RT-PCR of pMRSR

RNA amplifies a major product of 360 bp corresponding to removal of the intron using the distal (i.e. downstream) 5'ss. The major mRNA product from each RSS construct matched this major product of pMRSR. Thus, the candidate RSSs are not defining strong novel exons or altering the splicing of pMRSR strongly. Minor additional bands were observed for the *CDH4, MAPKAPK2*, and *GRK5* RSS constructs. Most of these amplimers are not observed consistently, but in the case of *CDH4* a band of 466 bp corresponds to use of a cryptic 5'ss (/GUGAGG) 106 bp downstream of the RSS, defining a cryptic cassette exon. Activation of a cryptic splice site is not unexpected, as moving a RSS out of its native context can interfere with regulation of the RSS and derepress surrounding cryptic splice sites. This has been observed previously for some Drosophila RSSs (Papasaikas et al 2010). The *MAPKAPK2* construct uses the upstream competing 5'ss of hF7 more frequently than pMRSR, but this still generates only a minor amplimer of 309 bp in addition to the 360 bp amplimer.

Next, I tested for the formation of the recursive intermediate, which would verify the use of the RSS as a 3'ss (Figure 4.3). For this purpose I used the same forward primer as for the mRNA, but in combination with a reverse primer targeting downstream of the RSS. Amplimers of the predicted sizes could be detected for the *CDH4*, *MAPKAPK2*, and *BACH2* constructs, confirming that the RSS motif in these cases functions as a 3'ss. These recursive intermediates were detected using 5 more PCR cycles than the corresponding mRNA, which is consistent with previous results in Drosophila showing that recursive



Gene	Sequence	Score	Conserved	Intron Size	Position
ACTN4	ACCCTCCCTGTTCT CCAGGTGAAGGG	0.99999967	All Mammals	52,692	38,573
BACH2	TTACGTTTTTTTTTA ACAGGTGAGCAT	0.99998932	All Mammals	89,803	55,239
SLIT3	CCCATTCTTCTCCC TGCAGGTAGCTGT	0.99999487	All Mammals	49,053	35,497
GRK5	TCCCCTCTTCTCCC TCCAGGTGACTCA	0.99999999	Primates	118,546	98,854
RSU1	TTTTTTTTCCTCTC CCCAGGTGGGGAT	0.99999998	Chimp	101,526	61,548
MAPKAPK2	CCTCTCCTTTGTCC TCCAGGTGAGAGA	0.99999996	Primates	43,201	14,803
CDH4	TGTTCTCTCCTTTC TAGGTAAGTGAT	0.99999926	All Mammals	488,625	246,908

Figure 4.2. The seven predicted human RSSs studied in this chapter. The provisional human EMSS RSS motif is shown at the top.

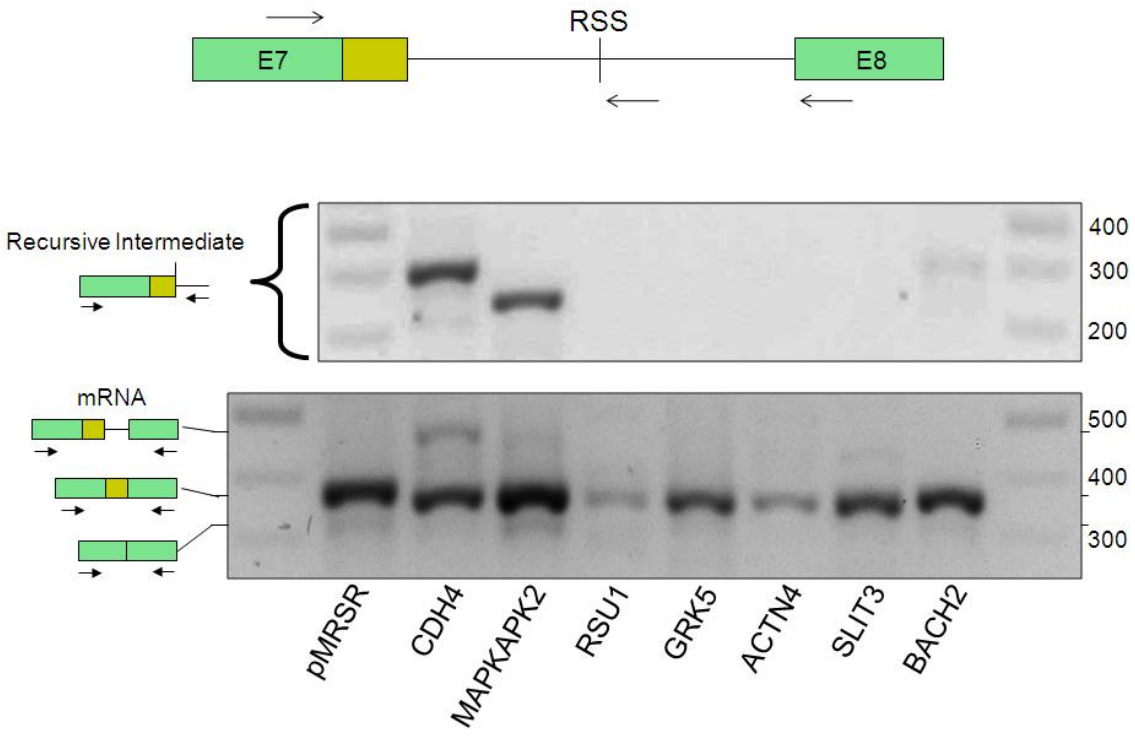


Figure 4.3. Splicing of the various RSSs in pMRSR. The diagram at the top shows the placement of the primers to detect the mRNA and recursive intermediate.

intermediates accumulate to about 3% of the level of its corresponding mRNA (Burnette et al 2005, Conklin et al 2005, Papasaikas et al 2010).

In my second set of experiments I attempted to verify the use of the RSS as 5'ss by assaying the formation of the recursive lariat which should be formed during the second stage of recursive splicing. To facilitate the detection of splicing lariats, I attempted to knock down the expression of RNA lariat debranching enzyme (DBR1) using siRNAs (Ambion), and a plasmid expressing siRNA (a gift from the Camerini Lab). I also tested a human version of the dominant negative allele of Drosophila DBR1 that substitutes a histidine with a tyrosine (H85Y), inactivating the debranching activity (Conklin and Lopez, personal communication). Unfortunately, none of the methods allowed me to detect construct lariats in HeLa cells, so I was unable to verify the use of the RSS as a 5'ss by this approach. As an alternative approach, I used mutagenesis to test whether the RSS functions sequentially as a 3'ss and regenerated 5'ss. This involved mutating the 5'ss component of the RSS motif so as to weaken the regenerated 5'ss. In this situation, completion of recursive splicing should become inefficient, resulting in retention of the downstream intron fragment or use of a competing 5'ss to complete its removal. This could be the upstream 5'ss in exon 7 of pMRSR or a cryptic 5'ss downstream of the RSS in the inserted fragment. Only *CDH4, MAPKAPK2* and *BACH2* were tested, because a recursive intermediate was only detected for these constructs in the previous experiments. *CDH4* and *MAPKAPK2* showed a dramatic effect when mutated, with a nearly complete or very strong shift to use of a 5'ss downstream of the

RSS (Figure 4.4). For *CDH4* this was the same as the weakly activated cryptic 5'ss in the wildtype construct (Fig. 4.3); this site is predicted to be stronger than the proximal competing 5'ss in exon 7 of pMRSR (GTGAGG versus GTGGGT). For *MAPKAPK2* this was a weaker and previously undetected cryptic 5'ss (/GTCTGT) located 126 nt downstream of the RSS. *BACH2* has a potential weak 5'ss 38 bp downstream of the RSS (GTAATA), but no splicing was detected at this site. All three mutants produced recursive intermediates at the same level as their corresponding wildtype minigene, indicating that the mutation did not affect use of the RSS as a 3'ss. The *BACH2* mutant construct showed no interference of recursive splicing. Restriction digests and sequencing have confirmed that the mutation is in the correct location. However, regulatory elements could overcome the weakened splice site, as the first two nucleotides of the 5'ss (GT) are still intact.

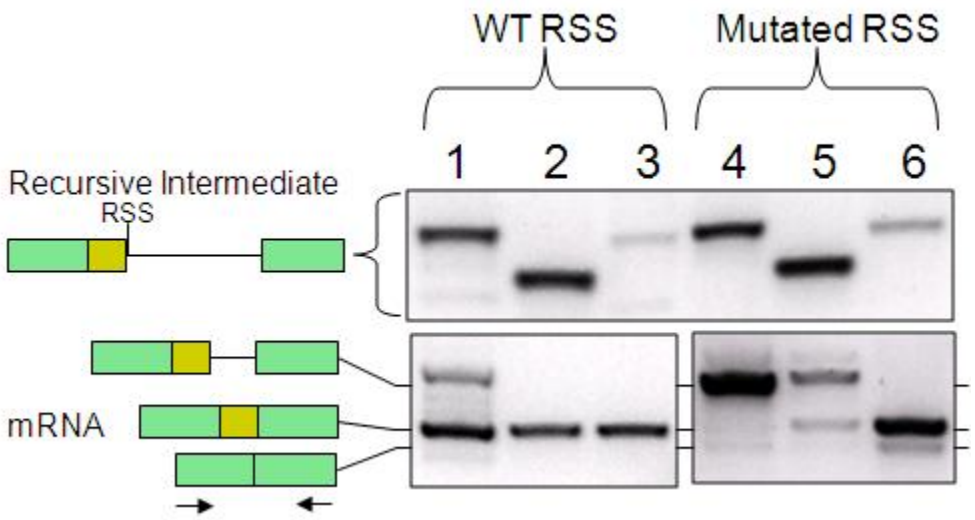


Figure 4.4. Effect of mutating the 5'ss component of RSSs. Lane 1, 4: *CDH4*, Lane 2, 5: *MAPKAPK2*, Lane 3, 6: *BACH2*.

Prediction of a Recursively Spliced Cassette Exon in the human Dopamine Transporter Gene (*SLC6A3*)

At the same time that I was performing the above experiments, Dr. Vishwajit Nimgaonkar's group in the Department of Human Genetics at the University of Pittsburgh: School of Public Health was identifying single nucleotide polymorphisms associated with differential schizophrenia risk in two human populations (Talkowski et al, 2008, Talkowski et al 2011). The strongest associations involved a haplotype whose component SNPs were located within the large intron 3 and intron 4 of the *SLC6A*3 gene, which encodes the dopamine reuptake transporter (DAT). The intronic regions containing the SNPs spanned 12.5kb. As the SNPs were all intronic and no alternative splicing had been observed in *SLC6A3*, Dr. Nimgaonkar requested our help to determine whether the intronic SNPs might be causal by affecting one or more recursive splice sites.

A search of the *SLC6A3* introns with our mammalian RSS model revealed a region located ~6 kb downstream of exon 3 that stood out as a possible RSS associated with a candidate cassette exon (which we named E3b; Figure 4.5). E3b is defined by a RSS with a suboptimal 3'ss (AAG; the A at -3 is infrequent but observed occasionally in regular 3' splice sites) followed by multiple downstream 5'ss motifs spanning a region of 363 bp. There is a potential alternative tandem 3'ss (TAG) four nucleotides downstream of the AAG, and this overlaps with the 5'ss component of the RSS. In the human sequence, the closest of the exon-defining 5' splice sites is located 108 bp downstream of the

RSS (Figure 4.5). It was particularly interesting that four SNPs shared by the haplotypes associated with schizophrenia (Talkowski et al, 2011) are located within 600 bp of E3b (Figure 4.5). Another notable feature is that use of either tandem 3'ss would introduce in-frame stop codons if E3b is retained in mRNA (Figures 4.5); this would truncate the ORF after ~100 codons (out of 620) and preclude expression of functional DAT.

Our phylogenetic analysis revealed conservation of an ORF-truncating exon E3b in all simian species with sequenced genomes and in almost all sequenced representatives of the nine Eutherian orders that have been analyzed (Talkowski et al, 2011). The exceptions are rat,mouse and rabbit, where a large portion of the intron, including E3b, is deleted. Thus, alternative recursive splicing of exon E3b could play a negative regulatory role in dopaminergic signaling, and differences in retention of the exon in schizophrenia-associated haplotypes could play a role in enhancing risk for schizophrenia.

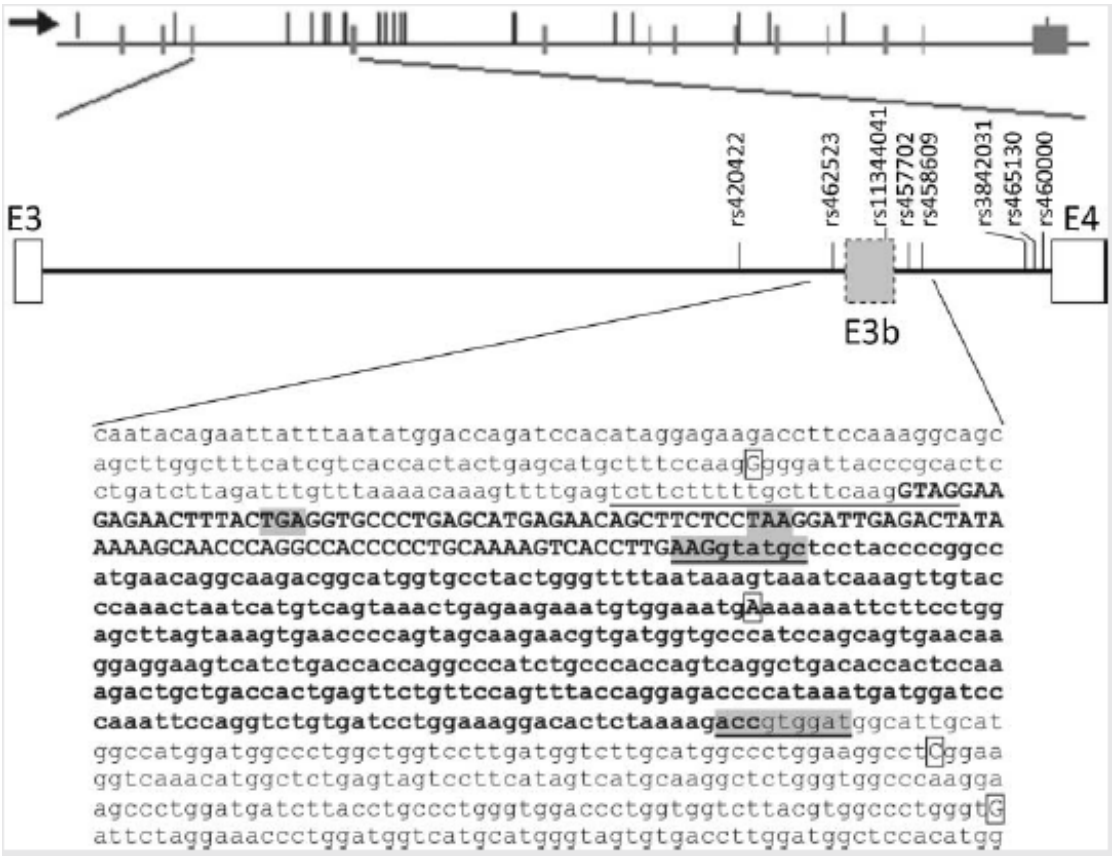


Figure 4.5. Computational prediction of E3b. Top panel: Human *SLC6A3*. The horizontal line represents the transcription unit, with exons as vertical bars or rectangles that cross the line. Thin vertical bars above the line represent SNPs with nominally significant associations with differential risk for schizophrenia (Talkowski et al 2011). Middle panel: Diagram of the genomic region from Exon 3 through Exon 4 of human *SLC6A3*. Labeled vertical ticks indicate indicate the identified schizophrenia-associated SNPs surrounding Exon E3b and upstream of E4. Bottom panel: Sequence of E3b and flanking intron regions. The exon sequence is shown in bold. The 3'ss motif is underlined; two potential tandem 3' splice sites (AAG/ and TAG/) are four nt apart within the underlined motif. Two alternative 5' splice site motifs are underlined and shaded. Bold uppercase corresponds to the exon region included if the first underlined 5'ss is used; bold lower case corresponds to the additional exon region included if the downstream underlined 5'ss is used. The first in-frame stop codons in each of the two relevant reading frames are highlighted in gray: TGA is in frame when the downstream 3'ss (TAG) is used; TAA is in frame when the upstream 3'ss (AAG) is used. Schizophrenia-associated SNP positions identified in two population samples (Talkowski et al 2011) are boxed; the risk-associated allele is shown in each case. Note that the 5' end of E3b conforms to a recursive splice site motif (Y_nNAAG/GTAGGA).

Experimental verification of E3b Splicing

Inclusion of E3b in *SLC6A3* mRNAs was not described in the literature or in current mRNA or EST databases. However, inclusion of E3b would truncate the ORF early and more than 55 nt upstream of the E3b/E4 exon junction, and this would be expected to trigger mRNA degradation by the nonsense-mediated decay pathway (NMD; reviewed in: Stalder and Muhleman, 2008). Rapid degradation of E3b(+) mRNAs could prevent their accumulation to significant steady-state levels, which could explain the absence of EST and cDNA evidence for E3b inclusion even if this splicing event were relatively frequent. Our RT-PCR analyses of mRNA in the *substantia nigra* of human post-mortem brain samples have confirmed that E3b is alternative spliced *in vivo* in the context of full-length *SLC6A3* mRNAs and that the resulting E3b(+) *SLC6A3* mRNAs accumulate to very low steady-state levels, as predicted (Talkowski et al 2011; these experiments were carried out in collaboration with Carnegie Mellon undergraduate researcher Kathleen McCann). In conjunction with sequencing of the amplimers, these RT-PCR analyses also confirmed use of the alternative tandem 3' splice sites as well as alternative 5' splice sites for E3b (Talkowski et al 2011).

For initial tests of whether E3b iis recursively spliced, I cloned a 925-bp fragment of human genomic *SLC6A3* DNA surrounding E3b into pMRSR (Figure 4.6). Analysis of construct-derived RNA by RT-PCR after transient transfection into HeLa cells revealed that E3b behaved as an alternatively spliced exon that was included in a substantial proportion of the processed RNAs (Figure 4.6). To

test whether inclusion of E3b is controlled by recursive splicing at its 5' end, I mutated the 5'ss component of the RSS so it would not regenerate an efficient 5'ss (Fig. 4.6). The result was that a higher proportion of the mRNA retained E3b, as expected if E3b(-) mRNA results from recursive splicing of E3b.

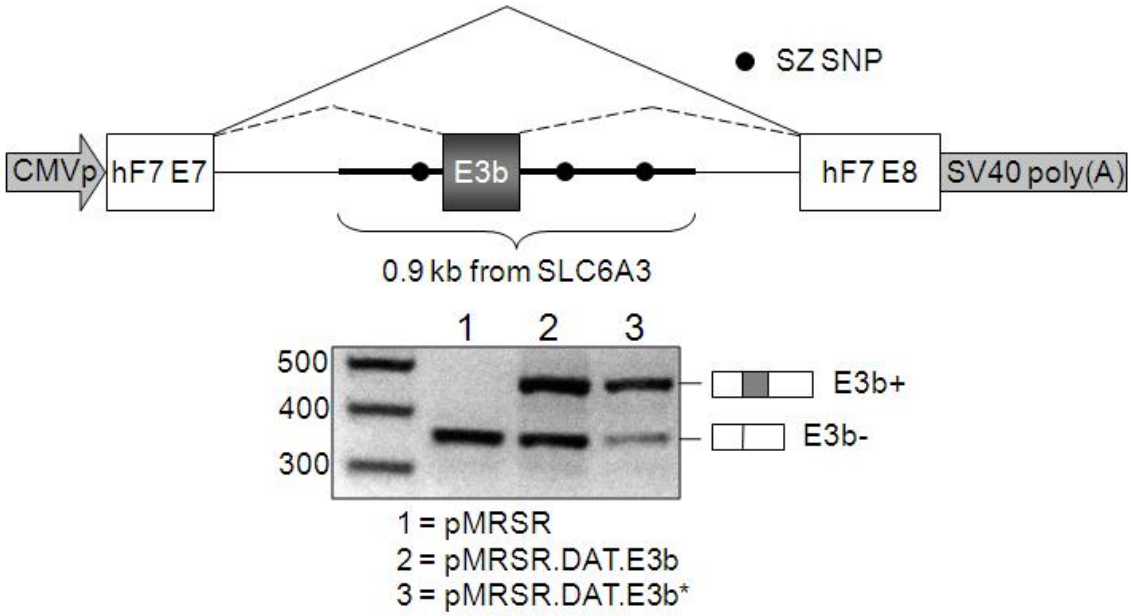


Figure 4.6. Alternative splicing of E3b in a chimeric splicing construct. Top: diagram of the chimeric transcription unit in pMRSR. Bottom: RT-PCR analysis of RNA from HeLa cells transfected with the empty parent vector (pMRSR, lane 1), the chimeric construct containing E3b (pMRSR.SLC6A3.E3b, lane 2), or the E3b construct with a mutated RSS (5'ss component). The spliced RNA species (E3b- and E3b+) are identified at the right of the gel.

Effects of schizophrenia-associated SNPs on E3b splicing

(These experiments were conducted in collaboration with undergraduate researcher Kathleen McCann and are already publidhed in Talkowski et al 2011).

Chimeric construct pMRSR.SLC6A3 contained non-risk alleles at all the SNP positions that were included in the cloned fragment, but it did not include all of the potentially relevant SNP positions of intron 3. Furthermore, E3b in this construct was flanked by heterologous exons. To study the splicing of E3b in a more natural context and to evaluate the possible effects of risk and non-risk alleles, a second set of constructs was generated that placed the native region of *SLC6A3* containing the end of exon 3 through the beginning of exon 4 under the control of the CMV promoter and SV40 cleavage/polyadenylation site (construct series pCMV.SLC6A3.E3-E4; see Materials and Methods and Talkowski et al, 2011). Translation start codons upstream of E3b were avoided to prevent degradation of mRNA by NMD. The *SLC6A3* fragment in each construct of this series contained either the non-risk alleles at all the known risk-associated SNPs within intron 3 (the “non-risk haplotype”: 5 constructs) or the risk-associated alleles (the “risk haplotype”: 5 constructs) (see Materials and Methods). The fragments differed in sequence at other SNP positions not known to be associated with differential risk and at a VNTR (variable number tandem repeat) polymorphism within intron 3, but none of these variations distinguished between the risk and non-risk sets.

RT-PCR analysis after transfection of these constructs into the human neuroblastoma cell line SH-SY5Y or into HeLa cells revealed that inclusion of

exon E3b in the minigene mRNAs was increased by 3- to 4-fold when the RNA was derived from any of the the risk haplotype constructs compared to any of the non-risk haplotype constructs (Talkowski et al 2011). These results were confirmed with three to four independent replicates for each construct and the differences between risk and non-risk constructs were found to be highly significant. Wilcoxon's rank sum test indicated that the differences in average percent E3b inclusion between risk and non-risk constructs were significant at $P<0.005$ ($w_{non\text{-}risk}=15$, $w_{risk}=40$). Analysis of variance showed that the differences in average E3b inclusion across the entire set of constructs were significant ($F=4.25$, $P<0.01$), whereas those among risk or non-risk constructs were not ($F_{risk}=0.938$; $F_{non\text{-}risk}=0.39$; $P>>0.05$ in both cases). Application of Student's t-test to the pooled risk versus pooled non-risk data confirmed that inclusion of E3b was significantly higher for risk than for non-risk constructs ($t=5.85$, $P<0.001$). Thus, the risk haplotype is associated with increased inclusion of exon E3b, at least in the context of these transfected constructs. The difference in behavior between the heterologous (pMRSR.SLC6A3) and native (pCMV.SLC6A3.E3-E4) constructs derived from HeLa and non-risk haplotype brain DNA, respectively, is consistent with the conclusion that the efficiency of inclusion of E3b in mRNAs depends on the sequence context. Although all of these constructs contain the non-risk alleles immediately surrounding E3b, the pMRSR.SLC6A3 construct (which exhibits increased splicing of E3b; compare Figure 4.6 with Talkowski et al, 2011) lacks the more distal SNP positions, places E3b in a shorter intron (total 1.7 kb vs 8.5 kb) and pairs E3b with heterologous

exons.

Mapping Sequence Variations Reponsible for iIncreased Inclusion of E3b

(These experiments were conducted in collaboration with undergraduate researcher Kathleen McCann and are already publidhed in Talkowski et al 2011).

To map the approximate location of sequence variations responsible for increased inclusion of E3b we swapped the region extending from genomic coordinates 1487252 to 1488492 (-519 to +721 relative to the first 3'ss of E3b) between risk and non-risk constructs that were derived from CEPH samples 12155 and 12249, respectively. The results showed that exchanging this region was sufficient to confer elevated E3b inclusion on the otherwise unaltered 12155 non-risk construct (Talkowski et al, 2011). The swapped region contains the four previously identified schizophrenia-associated SNPs that immediately flank E3b (rs462523 through rs458609) plus two additional single-nucleotide differences that were identified between risk and non-risk haplotypes during sequencing of the CEPH panel samples. The new polymorphisms were at nt 1487928 (45 nt downstream of the first 3'ss for E3b; G in CEPH 12249, T in 12155) and at nt 1487382 (591 nt downstream of the first 3'ss for E3b; C in CEPH 12249, T in 12155).

DISCUSSION

With experiments described in this chapter I have obtained evidence for the existence of recursive splicing in human gene transcripts and for its potential importance as a mediator of normal or aberrant alternative splicing. I verified recursive splicing by testing the predictions of 3'ss and 5'ss component activity. Unfortunately the lariat analyses have not been feasible in transfected human cell lines, so currently the activity of the regenerated 5'ss component must be inferred from the mutagenesis experiments.

RSSs and NMD

For the *CDH4* and *MAPKAPK2* minigenes, my results indicate that the predicted RSSs contain a functional 3'ss, and that the 5'ss component is required for normal processing of the recursive intermediate, as predicted. In both cases the RSS was predicted to be non-exonic. However, in the 5'ss competition assay, the proximal 5'ss engineered in the minigene was not used. In both *CDH4* and *MAPKAPK2*, mutation of the regenerated 5' ss shifted the completion of splicing primarily to use of a downstream site within the *CDH4* or *MAKAPK2* intron, thus creating an exon. Use of these downstream sites and inclusion of these exons has not been annotated, or even revealed through analysis of EST tags, so they may represent cryptic 5' splice sites and pseudoexons rather than natural alternative splicing. However, in both cases inclusion of the novel exon is expected to trigger degradation of the mRNA through NMD, as in-frame stop

codons populate both exons. Thus, they might be bona fide alternatively spliced exons whose discovery could have been prevented by low steady-state levels of accumulation of the corresponding transcripts from the native gene. I was also able to test a predicted RSS in the third intron of the *SLC6A3* gene. In this case the RSS is associated with a newly predicted and validated cassette exon (E3b) that introduces an in-frame PTC into the mRNA and appears to trigger NMD. Thus, endogenous E3b(+) mRNAs accumulate to very low but detectable levels in brain tissue, whereas analysis of E3b inclusion in non-translatable minigene transcripts shows that the exon can be retained efficiently. As noted in the Introduction, Grellschied and Smith (2006) described an NMD-triggering cassette exon in the alpha-tropomyosin gene of rat whose inclusion appears to be controlled by recursive splicing. Inclusion of this cassette exon leads to mRNA degradation through the nonsense mediated decay (NMD) pathway, and is suppressed by hnRNP F and H (Coles et al 2009).

The above results thus suggest that recursive splicing in mammals may be commonly associated with alternative splicing of NMD-triggering cassette exons. This may be another reason why recursive splicing has been difficult to detect in mammals. It also suggests that recursive splicing of such NMD-triggering cassette exons could have an important role in quantitative regulation of gene expression. Conversely, mutations or genetic polymorphisms at recursive splice sites that lead to activation of cryptic NMD-triggering exons could have important health consequences by reducing the levels of functional gene product. This idea is consistent with the finding that common SNPs that may be

associated with schizophrenia risk can alter the efficiency of E3b retention in *SLC6A3* minigene RNAs.

In Drosophila, over 50% of non-exonic RSSs have a pseudo-5'ss at 50-100 bp downstream of the RSS (Papasaikas et al 2010). Although in the one tested case (*Ubx* RP3) the downstream 5'ss is never used for splicing under normal circumstances and is part of a U1-snRNA-dependent regulatory element, it can be activated to define a pseudo-exon by mutation of other cis-acting elements associated with the RSS (Chapter 2). At least two other Drosophila RSSs which were believed to be non-exonic, exhibited use of a cryptic 5'ss downstream of the RSS when the regenerated 5'ss was mutated, similar to the mammalian examples analyzed here (Papasaikas et al 2010). It is thus also possible that a subset of apparent non-exonic RSSs in Drosophila may actually be associated with NMD-triggering cassette exons.

Biological Significance of NMD

The nonsense mediated decay pathway was originally interpreted as a mechanism to deal with aberrant mRNAs that have a premature termination codon (PTC) (reviewed in Silva and Romao 2009, Maquat and Gong 2009, Brogna and Wen 2009). When a ribosome encounters a PTC without an accompanying poly-A binding protein-mediated termination signal, the ribosome stalls for a period of time, allowing UPF1 to bind to the SURF complex. UPF2 and UPF3 then bind and phosphorylate UPF1, licensing the mRNA for degradation, which can be enhanced by the presence of an exon junction

complex downstream of the PTC. The mRNA is then degraded by two pathways, one which removes the mRNA cap and poly-A tail followed by exonucleolytic degradation, and another by which the mRNA is cut by an endonuclease followed by degradation beginning from the cleavage site.

PTCs can arise in mRNAs in a number of ways. At the DNA level, point mutations can directly change a normal amino-acid codon into a PTC, while frameshift mutations can alter the frame, introducing a PTC. At the RNA level, transcription errors or aberrant processing can also introduce PTCs. The inclusion of a pseudoexon between exons 20 and 21 of the *ATM* gene truncates the open reading frame of the mRNA, causing the degradation of the mRNA by NMD (Pagani et al 2002). In another case, the inclusion of pseudoexons into the mRNA of the *Dystrophin* gene also affects the expression of normal protein, leading to Duchenne muscular dystrophy or Becker muscular dystrophy (Gurvich et al 2007). However, PTCs can also be spliced into mRNAs physiologically through normal alternative splicing for regulatory purposes. A classic example is sex determination in Drosophila, which hinges on alternative splicing to control the inclusion of a PTC-containing exon in mRNAs from the master switch gene *Sex lethal* (*Sxl*) (reviewed in: Lopez 1998; Salz 2011). ORF-truncating, NMD-triggering alternative exons with potential negative regulatory effects are predicted to be common in humans (~45% of alternative splicing events; reviewed in: Stalder and Muhleman, 2008). However, it remains unknown what proportion of these cases are actually exploited for regulation.

Alternative Splicing and NMD in the Human *SLC6A3* Gene

The dopamine reuptake transporter (DAT, encoded by *SLC6A3* in humans) sequesters synaptic dopamine (DA) into presynaptic nerve terminals (Amara and Kuhar 1993, Giros and Caron 1993, Gainetdinov et al 2002, Torres et al 2003, Cragg and Rice 2004). *SLC6A3* may contribute to voluntary movement, reward and cognitive function (Mozley et al 2001, Sotnikova et al 2006). Expression of the transporter varies across brain regions, so it is considered a critical spatio-temporal regulator of synaptic DA activity (Amara and Kuhar 1993, Giros and Caron 1993, Cragg and Rice 2004). In humans, the highest levels of *SLC6A3* are present in the striatum; much lower levels are found in the neocortex (Farde et al 1994, Hall and Strange 1999). *SLC6A3* spans ~60 kb (Giros et al 1992, Vandenbergh et al 1992) and contains 15 previously documented exons, with the protein-coding portion spanning exons 2–15 (Bannon et al 2001). Until our study of E3b (Talkowski et al, 2011), only one mRNA isoform had been described; alternative splicing had not been observed in *SLC6A3* transcripts.

It has been difficult to understand the control of *SLC6A3* expression because neuronal cell lines stably expressing *SLC6A3* have not been available (Bannon et al 2001). *SLC6A3* has been a target for conventional candidate gene association studies of various psychiatric disorders, including schizophrenia (SZ) and schizoaffective disorder (SZA). Most studies have considered only a variable number tandem repeat polymorphism (VNTR) in the 3' UTR, with inconsistent results (reviewed by Talkowski et al 2007). The results presented here and in Talkowski et al 2011 suggest that alternative splicing of the NMD-triggering exon

E3b may be a mechanism for quantitative regulation of *SLC6A3* expression and that differences in the inclusion of this exon determined by common flanking SNPs may influence risk for schizophrenia in some populations.

If the function of E3b in *SLC6A3* RNAs is to mediate negative regulation by truncating the ORF, its boundaries and exact sequence need not be strictly conserved as long as it brings stop codons into frame. This is the case for the species whose *SLC6A3* sequence has been analyzed (Talkowski et al 2011). However, in the tenrec (superorder Afrotheria, an outgroup to the remaining species) the 3'ss motif is very weak and there is no 5' splice site, whereas the entire E3b region is absent from the sequences of opposum, mouse, rat, guinea pig, and rabbit. This suggests that E3b splicing may have evolved in the Boroeutheria and been lost secondarily in the Glires. The secondary loss of E3b would suggest that alternative splicing of E3b does not play an essential role in regulation of *SLC6A3*, although the Glires could have evolved an alternative mechanism for the same task or not require it. It will be important to determine whether E3b splicing is regulated under normal circumstances and to elucidate its functional effects in species where it is used. Genotypic variation for E3b splicing could be functionally relevant to schizophrenia because increased inclusion of E3b should result in decreased expression of functional dopamine reuptake transporter. Indeed, quantitative real-time RT-PCR analysis of post-mortem brain samples from 9 control and 9 schizophrenia cases shows anticorrelated variation in E3b inclusion levels and total *SLC6A3* mRNA levels (Talkowski et al, 2011). The most pronounced outlier, with highest E3b inclusion

and lowest *SLC6A3* level, was a schizophrenia patient. However, as the genotypes of these samples were not known, these data do not distinguish whether the changes in E3b inclusion were cause or effect of the disease condition, its treatment or side effects. Analysis of samples for which there is both high quality RNA and genotype data will be necessary to gain greater insight into these questions.

Chapter 5: General Conclusions and Future Directions

Since the initial discovery of non-exonic recursive splicing (Burnette et al 2005) numerous questions have arisen regarding its mechanisms, its origins and biological role, and its existence in organisms outside of Diptera, particularly in mammals. In this thesis I have presented work addressing aspects of each of these questions. The results provide a better understanding of recursive splicing but also raise many new questions

Activation and Coordination of Non-Exonic RSSs

Our current knowledge shows that many factors can contribute to the recognition and activation of splice sites that define regular exons. While the inherent strength of the splice sites, based on resemblance to the consensus motif, does contribute to their recognition, exonic and intronic splicing enhancers and silencers also play an important role in multicellular eukaryotes, particularly iwith regard to regulation of alternative splicing (reviewed by Chen and Manley 2009, Wang and Burge 2008). Splice sites associated with large introns are usually recognized and activated through the exon definition mechanism, in which the activated 3'ss and 5'ss interact across the exon to stabilize the binding of early spliceosome assembly factors (Berget 1995). If the exon size is reduced below a limit of about 50 nt, the splice sites appear to interfere with one another. Thus, the architecture of RSSs, where the 3'ss and 5'ss actually overlap, would appear to rule out this mechanism. Furthermore, non-exonic RSSs are used efficiently

first as a 3'ss, then as a 5'ss during the normal splicing of their respective introns, suggesting that some mechanism coordinates their sequential action (Burnette et al 2005). My work in chapter 2 investigated their activation.

A common downstream architecture for non-exonic RSSs in Drosophila consists of a pseudo-5'ss at approximately +50 nt, surrounded by predicted splicing enhancers. An obvious hypothesis was that these modules would function to activate the 3'ss component of the RSS by defining a pseudo-exon with the downstream pseudo-5'ss and enhancers, and that some unknown mechanism would suppress use of the pseudo-5'ss. I tested this hypothesis with *Ubx* RSS RP3, which contains a highly conserved module of this type. Surprisingly, I found that this module plays only a weak role in activation of the 3'ss component of the RSS. Instead, the main function of the module is to activate the regenerated 5'ss during the second stage of recursive splicing. This still leaves us with the question of how the RSS is activated as a 3'ss without interference by the overlapping 5'ss component. The enhanced information content and special features of the 3'ss component and branch site (Panagiotis 2010) may allow it to override the 5'ss component without need for additional specialized *cis*-elements. However, there may also be enhancers and silencers upstream of RP3, a region that has not been investigated. Minigenes where RP3 is pre-spliced to the downstream exon could facilitate the discovery of a mechanism for the activation of RP3 as a 3'ss.

A general question raised by this work is the mechanism for the directional effect of the downstream module, which stimulates only upstream 5' splice sites.

The failure to splice at pseudo-5'ss a is relatively easy to understand as element A is immediately upstream (in what would be be the pseudo-exon), so bound factors could block spliceosome assembly or activation at this position. Failure to splice at pseudo-5'ss b, further downstream, is more difficult to explain. A similar directionality has been seen in the mechanism of hnRNP H in the modulation of 5'splice sites (Fisette et al 2009, reviewed in Wang and Burge 2008). Use of a 5'ss is enhanced by the binding of hnRNP H downstream but suppressed by the binding of hnRNP H upstream. When there are two hnRNP H binding sites associated with two 5'splice sites, the upstream one is favored. The proposed mechanism is similar to the model previously formulated for some examples of regulation by hnRNP A1, namely that self-interaction between hnRNPs loops out the RNA containing the suppressed 5'ss, leaving the upstream 5'ss to be used (Martinez-Contreras et al 2006, Blanchette and Chabot 1999). A related model is also proposed for the mechanism behind the silencing of alternative exons by the neuronal splicing factor Nova (Ule et al 2006). Such models fail to explain suppression or activation by only upstream or only downstream sites, however, unless looping can be achieved by heterotypic interactions with other downstream factors. In the case of *Ubx* RP3, interactions between factors bound at the *cis*-element module with factors bound at unknown sites downstream of the pseudo-5' splice sites could loop them out, leaving the regenerated 5'ss for genuine splicing. Affinity purification using the cis-element module as bait may identify relevant factors. Ultimately, however, it is unclear why looping out a splice site in RNA would suppress its activity, unless some

scanning process is impeded or the splice site is also directly blocked by factors associated with the loop.

Another possibility is that the directionality of the RP3 enhancer module results from the direction of transcription, either because the upstream sites are available first or because of a role for the 5' cap complex. Another is that the factors that interact with the downstream module function by interaction with factors that are deposited near the upstream 5' splice sites during processing of the upstream exons (e.g. the exon-junction complex). This last possibility appears to be ruled out by my results showing normal processing of minigene transcripts in which all upstream exons are already pre-spliced at the DNA level to each other and to RP3 as a 3'ss. However, these results do reveal an interesting effect of the upstream splicing events: in RNAs from these minigenes, pseudo-5' ss a and element B become dispensable, whereas element A becomes even more important for correct splicing at the regenerated 5'ss. Furthermore, all missplicing is directed upstream rather than to pseudo-5'ss a or b. This suggests that the mechanism for activation of the regenerated 5'ss is different in the two circumstances, possibly due to the presence or absence of upstream exon-junction complexes and whether or not the 3'ss component of RP3 is available to define a pseudoexon with the downstream pseudo-5' splice sites when element A is mutated. A comprehensive analysis of the *cis*-element module, its interaction with *trans*-acting factors, and the structure of the RNA-protein complex would be needed to clearly understand the mechanism of directionality.

Biological Role of Non-Exonic RSSs

Recent computational and experimental studies (Papasaikas et al 2010) have identified many RSSs in the telomere-associated non-LTR retroelements of Drosophila and Lepidoptera. This has suggested the hypothesis that the non-exonic RSSs of long introns in Diptera resulted from an ancient spread of these retroelements into euchromatic genes. Nevertheless, the conservation of most non-exonic RSSs throughout Drosophila species as well as Anopheles suggests that recursive splicing plays an important role in the function of large genes in these organsims.

Previous work on the role of recursive splicing in Drosophila has focused on their association with alternatively spliced cassette exons (Hatton et al 1998). In this context, recursive splicing has an obvious effect on mRNA and protein structure. The role of non-exonic RSSs, which do not alter mRNA structure, has not been clear. Deletion of *Ubx* RP3 from minigenes with an intron of up to 10 kb had no obvious effect on the splicing of the minigeneor its level of steady state expression (Burnette et al, 2005; Lopez, AJ personal communication). However, the normal function of RP3 may not be required in this context, which bypassses the complex developmental regulation of *Ubx* transcription and splicing in vivo. My work in Chapter 3 takes the next logical step by deleting RP3 from its native context in the endogenous *Ubx* gene of flies. Because of the technical complexity of the approach and unanticipated hurdles, I was only able to answer some simple questions about the role of RP3 in the available time. Its deletion

led to a partial loss of *Ubx* function, confirming that it plays a role in *Ubx* expression in vivo. Consistent with this, the molecular analysis of steady-state *Ubx* mRNAs in the ΔRP3 mutant and controls shows a change in alternative isoform ratios during larval development, with a greater proprtion of isoform IVa in the mutant. However, I am unable to distinguish between different scenarios: (1) the shift in isoform ratio could be due to a defect in accuracy of splicing caused directly by the deletion of RP3; (2) it could be due to a change in transcription of *Ubx* in epidermal or mesodermal tissues, leading indirectly to an increase in the proportion of isoform IVa, which is specific to the central nervous system (Lopez and Hogness, 1991; Bomze and Lopez, 1994; Subramaniam et al, 1994; Lopez et al 1996. It should be possible to distinguish the two scenarios by quantitative RT-PCR of RNA from individual larval tissues it should be possible to distinguish the two scenarios.

Trans-heterozygotes between ΔRP3 alleles and $Ubx^{Cbx\text{-}Hm}$ exhibited significantly decreased viability prior to the pupal stages. Adults that eclosed exhibited poor mobility and had to be rescued from the food. This hints at a perturbation of normal *Ubx* function in organs that control motor functions. Immunohistochemical staining with isoform-specific and general antibodies should should indicate whether there are changes in tissue-specific expression of the different isoforms.

ΔRP3 alleles exhibited an interaction specifically with the processivity-deficient C4 allele of RNA Polymerase II 215-kd subunit in that the haltere-to-wing transformation was enhanced. This may reflect a role for RP3 in stimulating

elongation, possibly through interactions between the associated splicing factors and the transcriptional machinery (Fong and Zhou 2001, Lin et al 2009). Alternatively, it could be due to an exacerbation of a splicing defect by the reduced Pol-II processivity. The latter would appear to be inconsistent with a previous *Ubx* splicing defect attributed to the C4 allele (de la Mata et al 2003); the reported defect was a reduction of isoform IVa rather than an increase, and isoform IVa is not normally expressed in the haltere. However, the effect of the Pol-II C4 allele could be different on splicing of ΔRP3 compared to wild-type mRNAs. To address this question, the mRNA isoform ratios in haltere imaginal discs of the different genotypes should be compared by RT-PCR. Chromatin immunoprecipitation assays should also be performed to investigate whether ΔRP3 exhibits changes in polymerase density along the *Ubx* transcription unit compared to wild type in the presence or absence of C4.

A possible processivity defect as a result of RP3 deletion could be related to the profound suppression of the *white* marker during integration of the gene replacement donor element into the *Ubx* RP3 target region, which was surprising, given that transposable element insertions into other regions of the *Ubx* transcription unit do express the *white* marker. This phenomenon suggests the presence of a local chromosome structure that suppresses expression, at least in the Drosophila eye. Deletions of the upstream duplicated target region that arise during aberrant reduction to single copy re-activate expression of the *white* marker, and further studies of these deletions can narrow down the region causing the suppression. Additionally, it would be interesting to analyze

transcription of *white* in different tissues by in-situ hybrization to determine whether it is anticorrelated with the expression of *Ubx*.

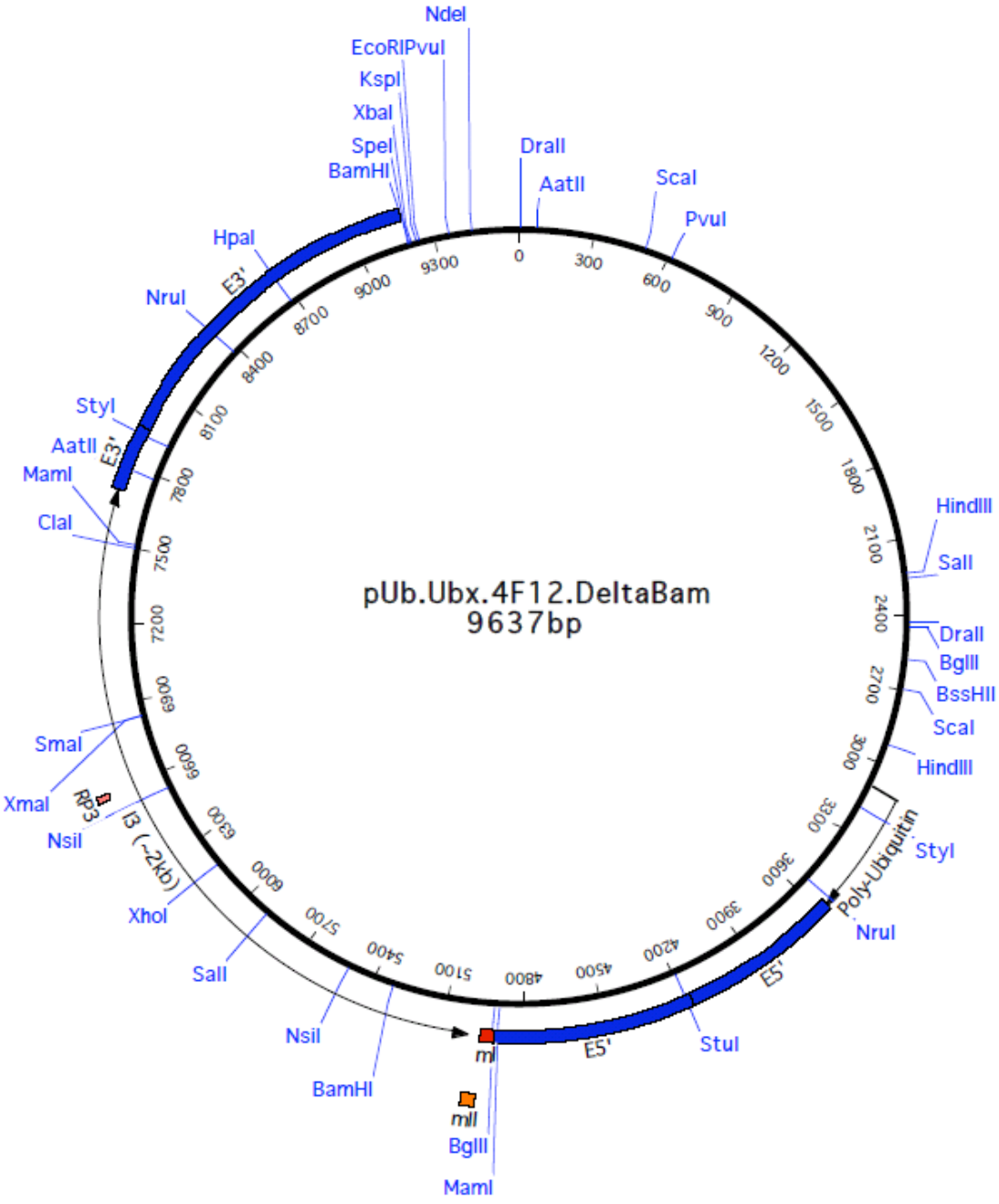
Recursive Splicing in Human Genes

Although recursive splicing has been well characterized in Drosophila, the existence of recursive splicing in humans has been an open question. I addressed this question in chapter 4, by replicating the methods used in Burnette et al 2005 for Drosophila RSSs. I was unable to stabilize recursive lariats for analysis, but I was able to prove in three cases that the 3'ss component of the predicted RSS is used, and that the 5'ss component is required to complete normal excision of the intron. Additionally, I found that each of the three cases of recursive splicing is associated with a cryptic exon that becomes activated upon mutation of the 5'ss component of the RSS. In the case of *SLC6A3*, I found that this presumed cryptic exon is actually a genuine cassette exon in vivo that may play a significant role in regulating dopamine reuptake transporter levels and dopaminergic signaling activity by truncating the open reading frame. The other case of mammalian recursive splicing with substantial experimental support (in rat alpha tropomyosin; Grellschied and Smith 2006) also defines an ORF-truncating cassette exon that triggers mRNA degrdation by nonsene-mediated decay. Although the sample of human RSSs analyzed is still small, it is possible that most are are associated with this type of alternative splicing, in contrast to the non-exonic character of most Drosophila RSSs. The verified CDH4 and MAPKAPK2 RSSs should be investigated more thoroughly to determine whether

the associated pseudo-exons are in fact *bona fide* cassette exons with regulatory implications, and a larger sample of predicted RSSs in humans and other mammals should also be tested to understand how the type, distributions and functions may differ from those in insects.

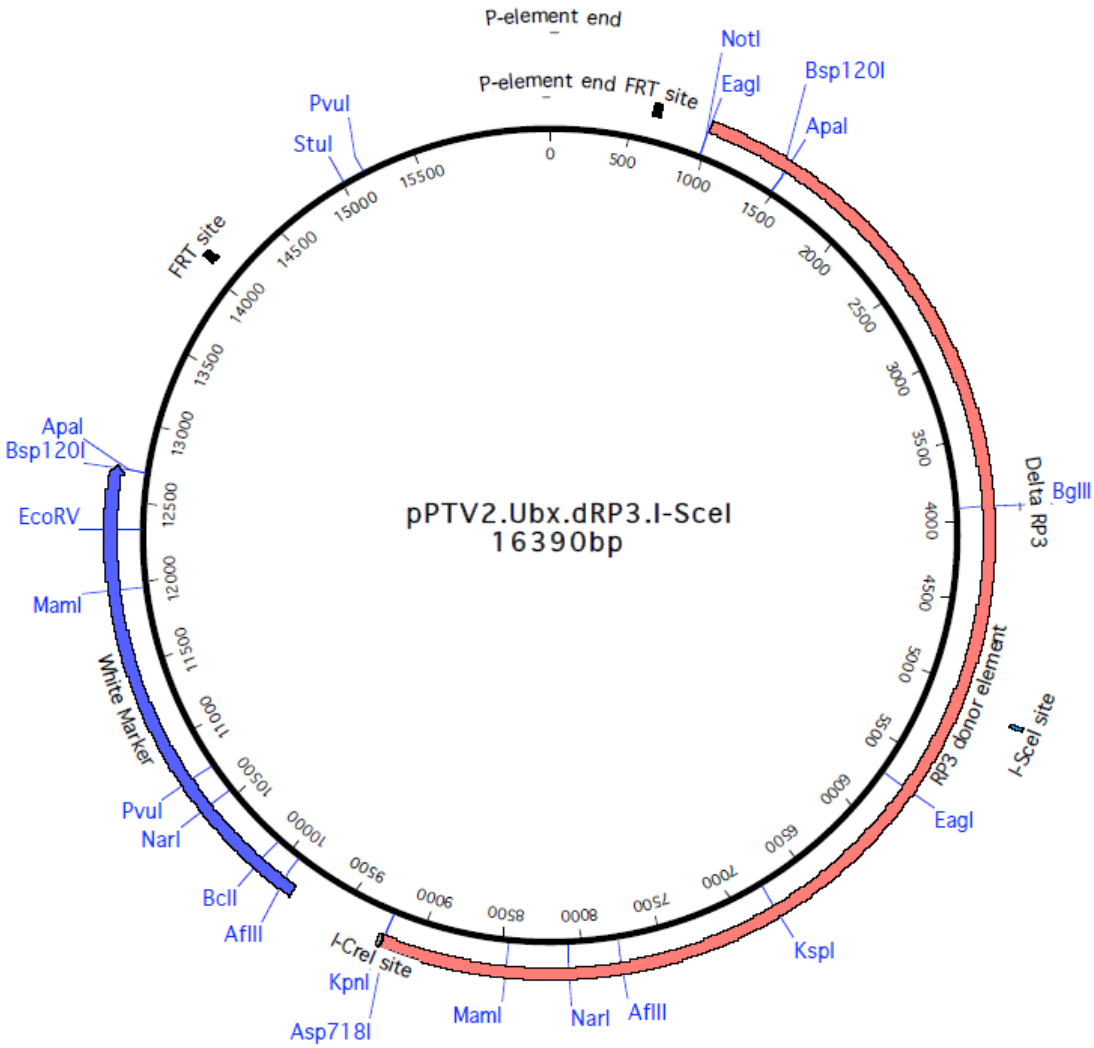
Appendix A

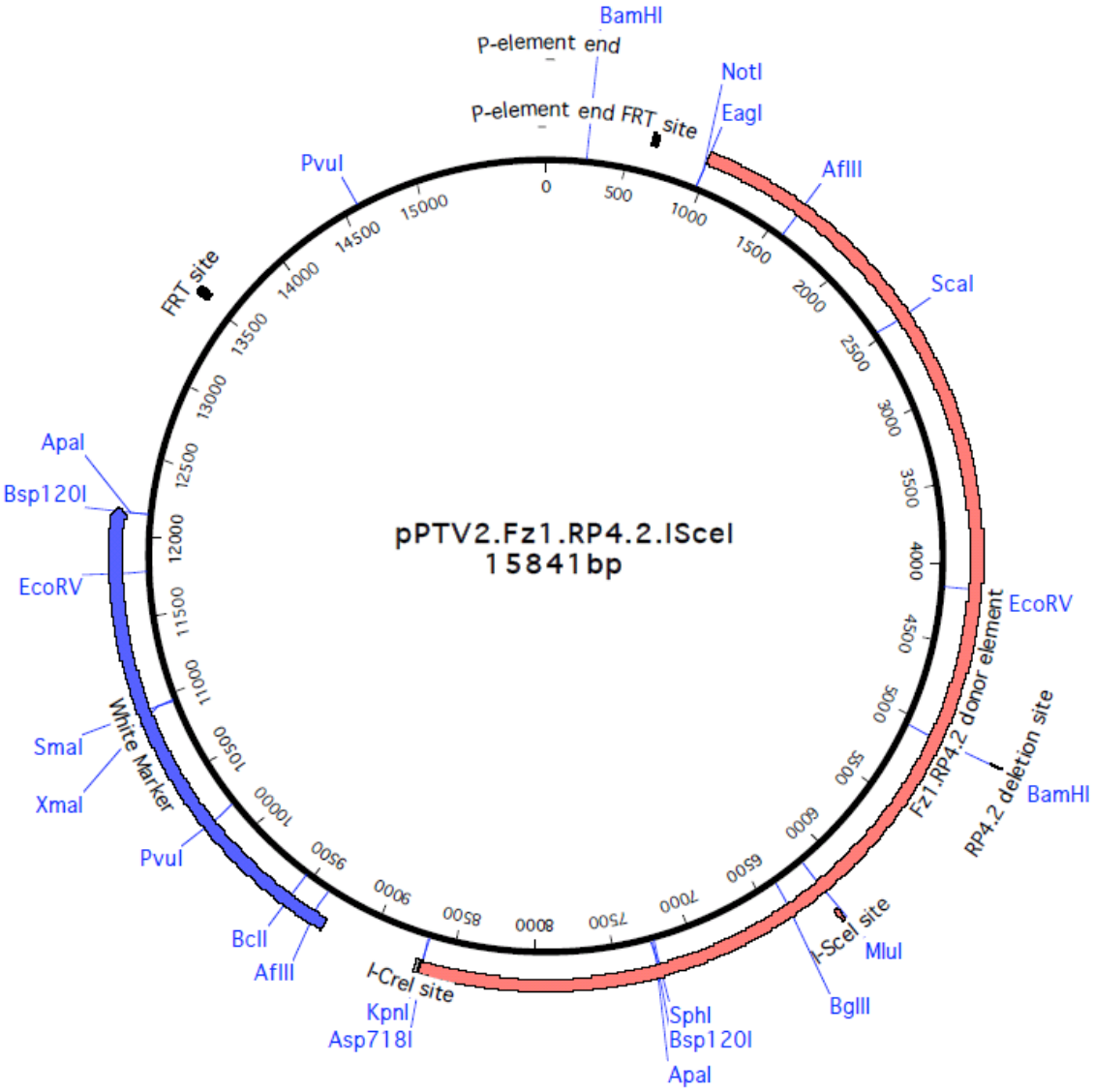
Plasmid Maps and Sequences

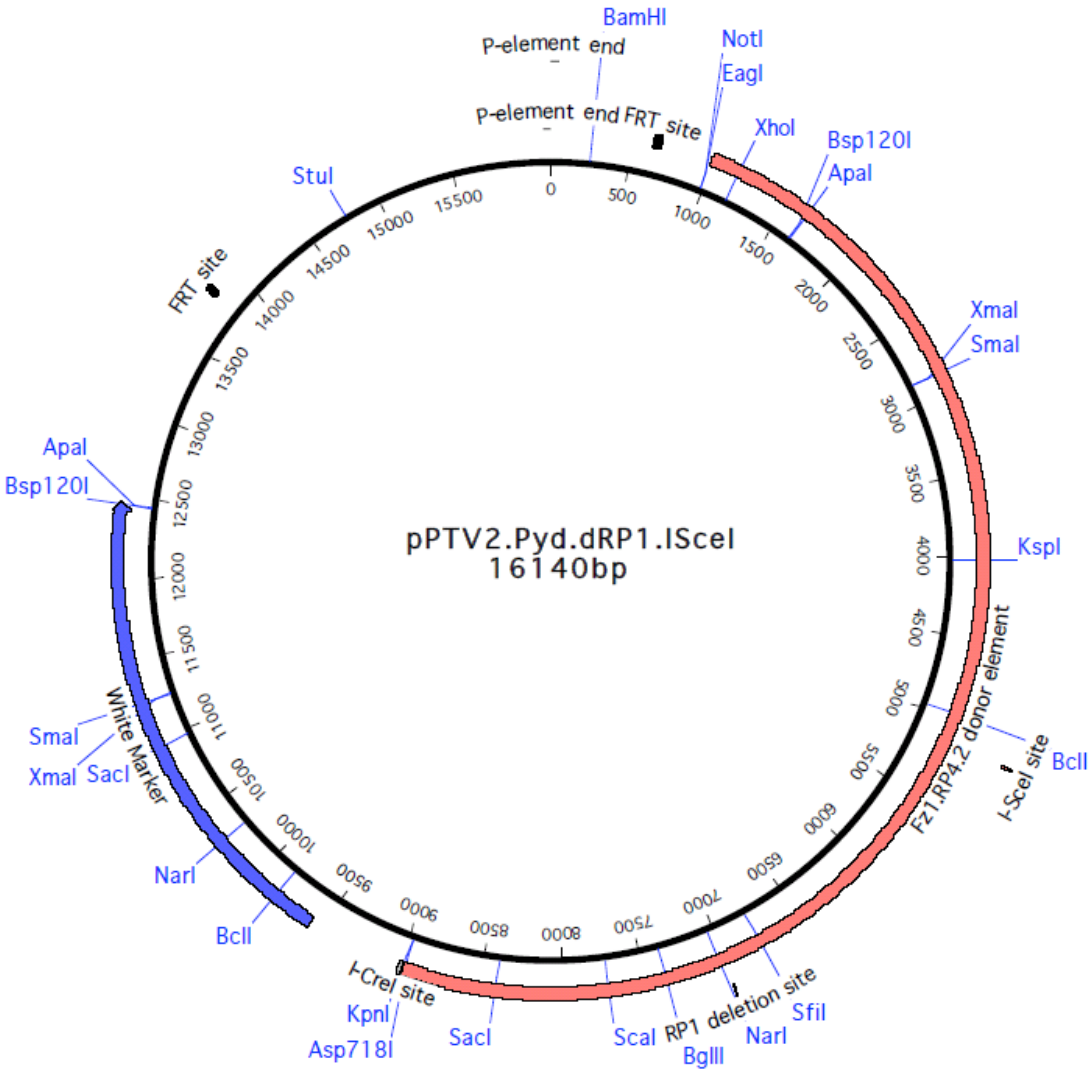

pUb.Ubx.4F12.DeltaBam
9637bp
NdeI
EcoRI
PvuI
KspI
XbaI
SpeI
BamHI
DraII
AatII
ScaI
PvuI
HpaI
E3'
NruI
StyI
AatII
E3'
MamI
ClaI
HindIII
SalI
DraII
BglII
BssHII
ScaI
HindIII
StyI
Poly-Ubiquitin
NruI
E5'
StuI
E5'
mI
mII
BglII
MamI
BamHI
NsiI
SalI
XhoI
I3 (~2kb)
RB3
NsiI
XmaI
SmaI
0
300
600
900
1200
1500
1800
2100
2400
2700
3000
3300
3600
3900
4200
4500
4800
5100
5400
5700
6000
6300
6600
6900
7200
7500
7800
8100
8400
8700
9000
9300

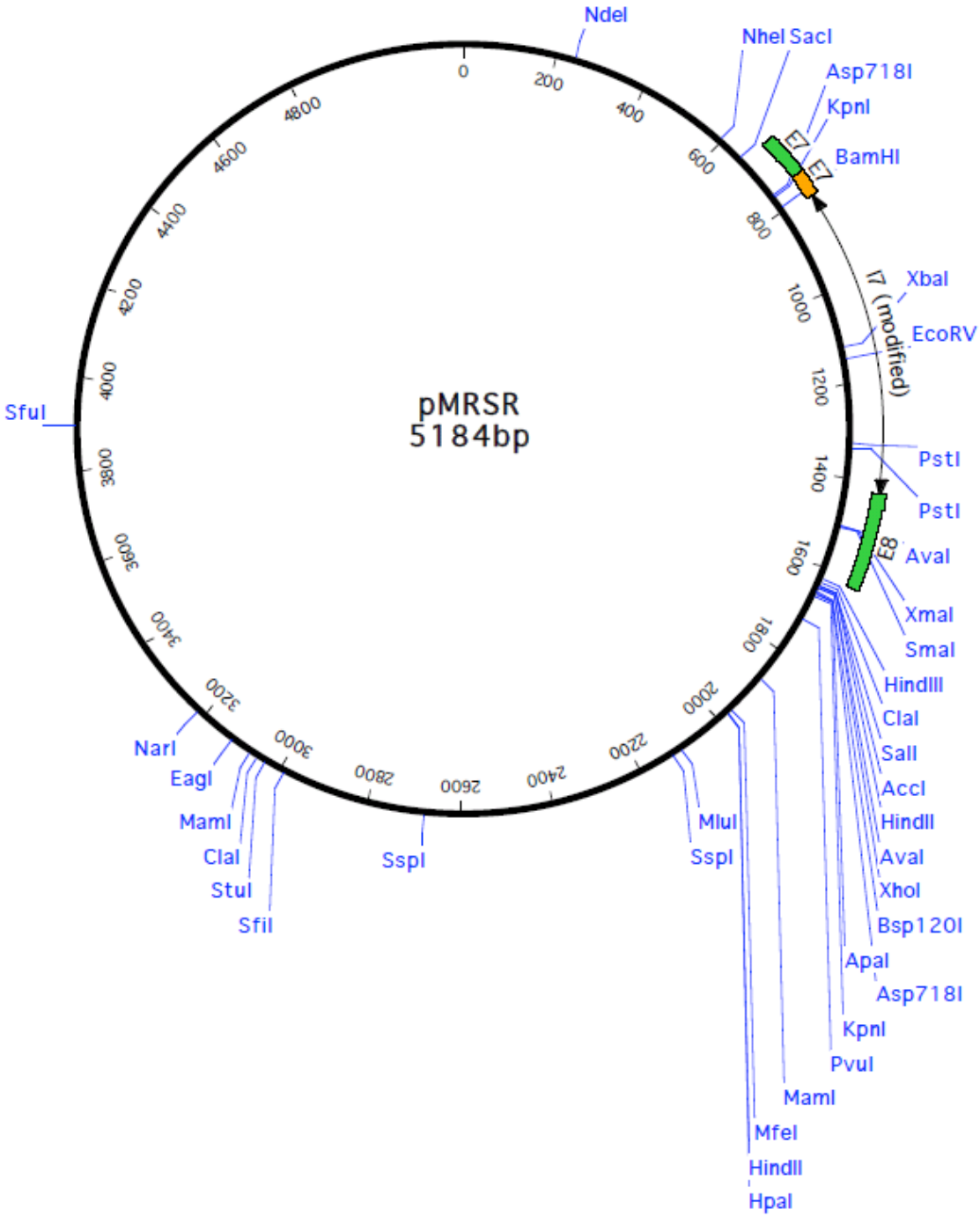
Name:	Sequence:	Ψ5'ssβ:
WT	AACCAAAACAAAAACATTGACAAAGTGAGTAAATAAGTATAATAATAAA	GTGAGT
DE1	Deletion decribed in text	GTGAGT
mut1	CGAAAATCGGATCCCGGTTTGGCCGTGAGTAAATAAGTATAATAATAAA	GTGAGT
mut2	AACCAAAACAAAAACATTGACAAAGGATCCAAATAAGTATAATAATAAA	GTGAGT
mut3	AACCAAAACAAAAACATTGACAAAGTGAGTATCGGATCCCGGTTTGGCC	GTGAGT
mutβ	AACCAAAACAAAAACATTGACAAAGTGAGTAAATAAGTATAATAATAAA	CCATGG
mut1mutβ	CGAAAATCGGATCCCGGTTTGGCCGTGAGTAAATAAGTATAATAATAAA	CCATGG
mut2mutβ	AACCAAAACAAAAACATTGACAAAGGATCCAAATAAGTATAATAATAAA	CCATGG
mut3mutβ	AACCAAAACAAAAACATTGACAAAGTGAGTATCGGATCCCGGTTTGGCC	CCATGG
m1m2	CGAAAATCGAATTCCGGTTTGGCCGGATCCAAATAAGTATAATAATAAA	GTGAGT
m1m3	CGAAAATCGAATTCCGGTTTGGCCGTGAGTATCGGATCCCGGTTTGGCC	GTGAGT
m2m3	AACCAAAACAAAAACATTGACAAAGAATTCATCGGATCCCGGTTTGGCC	GTGAGT
m1m2mβ	CGAAAATCGAATTCCGGTTTGGCCGGATCCAAATAAGTATAATAATAAA	CCATGG
m1m3mβ	CGAAAATCGAATTCCGGTTTGGCCGTGAGTATCGGATCCCGGTTTGGCC	CCATGG
m2m3mβ	AACCAAAACAAAAACATTGACAAAGAATTCATCGGATCCCGGTTTGGCC	CCATGG
m1m2m3	CGAAAATCGAATTCCGGTTTGGCCGGGCCCATCGGATCCCGGTTTGGCC	CCATGG
RP*	AACCAAAACAAAAACATTGACAAAGTGAGTAAATAAGTATAATAATAAA	GTGAGT
RP*m1	CGAAAATCGGATCCCGGTTTGGCCGTGAGTAAATAAGTATAATAATAAA	GTGAGT
RP*m2	AACCAAAACAAAAACATTGACAAAGGATCCAAATAAGTATAATAATAAA	GTGAGT
RP*m3	AACCAAAACAAAAACATTGACAAAGTGAGTATCGGATCCCGGTTTGGCC	GTGAGT
RP*mβ	AACCAAAACAAAAACATTGACAAAGTGAGTAAATAAGTATAATAATAAA	CCATGG
RP*m2mβ	AACCAAAACAAAAACATTGACAAAGGATCCAAATAAGTATAATAATAAA	CCATGG
RI-Long	AACCAAAACAAAAACATTGACAAAGTGAGTAAATAAGTATAATAATAAA	GTGAGT
RImut1	CGAAAATCGGATCCCGGTTTGGCCGTGAGTAAATAAGTATAATAATAAA	GTGAGT
RImut2	AACCAAAACAAAAACATTGACAAAGGATCCAAATAAGTATAATAATAAA	GTGAGT
RImut3	AACCAAAACAAAAACATTGACAAAGTGAGTATCGGATCCCGGTTTGGCC	GTGAGT

*Underlined sequences are restriction enzyme sites


pPTV2.Ubx.dRP3.I-SceI
16390bp
P-element end
P-element end
FRT site
NotI
EagI
Bsp120I
ApaI
PvuI
StuI
FRT site
ApaI
Bsp120I
EcoRV
MamI
White Marker
PvuI
NarI
BclI
AflII
I-CreI site
KpnI
Asp718I
MamI
NarI
AflII
KspI
EagI
RP3 donor element
I-SceI site
Delta RP3
BglII


pPTV2.Fz1.RP4.2.ISceI
15841bp
BamHI
P-element end
NotI
P-element end FRT site
EagI
PvuI
AflII
ScaI
FRT site
ApaI
Bsp120I
EcoRV
EcoRV
Fz1.RP4.2 donor element
RP4.2 deletion site
BamHI
White Marker
SmaI
XmaI
PvuI
I-SceI site
MluI
BglII
BclI
AflII
I-CreI site
KpnI
Asp718I
SphI
Bsp120I
ApaI


pPTV2.Pyd.dRP1.ISceI
16140bp
BamHI
P-element end
NotI
EagI
P-element end FRT site
XhoI
Bsp120I
ApaI
XmaI
SmaI
KspI
BclI
I-SceI site
F21.RP4.2 donor element
I-CreI site
KpnI
Asp718I
SacI
ScaI
RP1 deletion site
BglII
NarI
SfiI
BclI
NarI
White Marker
SmaI
XmaI
SacI
ApaI
Bsp120I
FRT site
StuI
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
7500
8000
8500
9000
9500
10000
10500
11000
11500
12000
12500
13000
13500
14000
14500
15000
15500


pMRSR
5184bp
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
3600
3800
4000
4200
4400
4600
4800
NdeI
NheI
SacI
Asp718I
KpnI
BamHI
E7
E7
I7 (modified)
XbaI
EcoRV
PstI
PstI
E8
AvaI
XmaI
SmaI
HindIII
ClaI
SalI
AccI
HindII
AvaI
XhoI
Bsp120I
ApaI
Asp718I
KpnI
PvuI
MamI
MfeI
HindII
HpaI
MluI
SspI
SspI
NarI
EagI
MamI
ClaI
StuI
SfiI
SfuI

CDH4

GTACTTGGCCACGGAGAGTCCAGGTGGGAAGAATTTATTAGGTGAGCGATGTAGCTGGA
ACGAGGCCTGGCTGAGATTCAGGAGACGGCTGCTGCCGTCTCTGCCCCCTCATCCTTGC
TGAATCTCAGAACCTTTGGGCCTGCATCTCTGCGGCCCGGAGACGCTGACTCCAGTTGC
CTAACAGAGCTGTGGTGAGACTCAAAGGCCACACAGAAGCAGCGTGTTCCGGAAAGCAT
GCCGTGCCCTGTGAACTAGAAGAGGGTACTGGAATACTAGCTATCTTCACAAGCACCAC
CCCGCCACTGCCTCCACCATTCGTTAGGGAGCCACTCATTATGGGACGGACCACCAGAC
GCGACTAATGAAACTTTTCTGTTCTCTCCTTTCTAGGTAAGTGATTCCGGTGCTAACGC
TGGACACGTGTGTCCTTGTATATACCTGTTGCGCCTGCGAGAACATGTTCTCCACTCCC
CCAGCTCCTCTGAGGAATGCCTGAGTGAGGACCCCGTGTAAAGTAAGATGAACCGAGTG
GGCACGTTAGCCCAGGCCGTTACCCTTTAGTGGGCGCGTGAGGGCTGTGGGTCACAGTT
TCCAAGACTGATGGCATTGCATGATCTTTCATGGTCAAACTTGTGCAAATGAAGAAGTG
GAGTGGGGGACAGGGGTGGGATGAGATAGCTAAGCAGTGAGCCAGAAAAAAAGCAGTTG
ACGTGGACGATCTAACCTGAGAAACCATCATAGGAAATGATGACTCTCTGCTCCAATGC
AGCATCTACAAGGCCCAAATGAGATAGAACACAGAGTTTCAATGTGAACTTCAAAACCT
GCTCCTGCTGTAAGGTGCTTTTCCCATCTTAGCAGCCTGGTGGTCTACTTTGTCGTTTT

*Recursive splice site is underlined

MAPKAPK2

GATGGGAAAGTGAAGCAATGGGTCGGGTCTGTTTGTTCCTGAGCATGTCCTGTGACAAA
GGAAGTTGCAACATGTCTGTCTGTCTTTCTCTCTCTTTTTCTTTTCTCCTTTCTTTTCG
TTTCTTTCTCTCTCTCTCTCTTCTTTTTTTCTCTTTCTTTCTTCTTCTTACTTTTTTTT
TTTTTTTTTTTAAATAAAGAAAGAAGCAGGTAAAACCATCTTCCAGAAAATGCCAGAAA
GAATATCTGGTCTGTGATGGGTGGTTCTTGCCCCTCCCGCTGCAATGCTTTTCCTGTGT
AGTTTTTCTTTTGTCCTTCCTTTCCCATCCACACTTTCCACTCTCCTTTCCTCTGAGAG
GTTGTGTCTCTGCAGCAGGCCAGCTATGTCTCCAAAGGGAAGTAGTCAGTGCTTGGACT
GAGGGAACCTTGAGAGAGAATCCTCTCCTTTGTCCTCCAGGTGAGAGAGCTGTTGCATC
CACTCACACCTGGAAGGTTGCTGGTCTCCTCTCGCTGGACACTGAGAGCTGTTTAAATT
TGACAGGTGGTTCCTGATTCCCACCTGGCATTTTTTGCTGCTGCACAGGTCTGTGTTGT
GTCTTAGAAGGCTGAGGGAATACTTACTTCTCAACTCAGATAACCACCTTGTTGGTAGG
TTCAGTGTGTGTCTGTGTCAAGGCAGATGCTGGGAAAAAAGGGCCCAGAATGTGAGTGG
GAGGTAGAAATCAAGGCTAGAAGACCACACATGTTAATCTGGAGTACTTGGTCAGGTTT
GATTTTGGCCT

*Recursive splice site is underlined

GRK5

GTGTCCAAACCCACACTGAGATCACACAGCGAGGGAAGGAAGCCGGGGTCCTGGCTGAC
ACTTGCGGGGCAGAGCACCTTCTCTTGGCCCCGAATATGGATTTTGGTGTCCCTAGTCA
AGATTTGTTCTGCTGACTTCTCCTGGGGAAAGAGGTGTCACGTCTGCCTGAGCCCCTGA
GCCTGGTGGGTGGGGTGTGAGGATCTGGCCGGGGCCGCTGCACCTGCAGGTGTTTCCTG
CAGACATTCTGCTGGGGAGCACTTCCCCTCTTCTCCCTCCAGGTGACTCACTGGGTGCA
GTGTCCTCTAGCGTCTACTGGGTGCAGGGCACTGGGGAGACATGGGGGATGCTTCTTTG
GGCTCGTCTGTACTAGGTATGGTTATAGGGTACTGGGGGTTGCCTAAAAAAATGCTGTG
TGACCTTGATGGCTCACTTTTTCTCCCTGGGCCTCTGTGTCTCCAGCCGTCAGTGAGGA
GCAGGCGTGGGTGATCAGCAAGGTTCCTCTCAGCT

*Recursive splice site is underlined

SLIT3

AATTGGATCAGGTGGACTGGTGAATTGCAGGCATGTACCTTATGCAAAGTGGGCAATTG
CCTCTTAGTACCAGCTTATGATGGCCACTAAAATCCATTGTTATTTGATGGACTGACCT
TTGGAGACAAGCCAGAATTGTAGATTTGTATTTGAAATGCTCCCAAATTGGCAATTAAT
TTGCATATGTGAAAATCACAGTGAGAACCAGACTACGCAGGTCTGCAGGCCAGAACTCA
CCGGTGGGCTACCAGTTTGCAAGTTCGTCTCTGCTGTTCCCAGTAAAGAGCTAGTTGTA
CGTTTTGTTTGTTTGTTTGTTTGTTTGTTTTTGAGCATGGAGCTGAACTGAGCAAAACA
GGGGCAGGGAGTCCTCCTGGTGCAAGGGATGAAGGAAAAGAGACCCCATGCCCAATACT
CCTCTTTACATGCCCTCACCATGACCACATCCAACCACCCGGGTTGCCCCTCTGTTTGG
CCCATTCTTCTCCCTGCAGGTAGCTGTCTGCACAGGGCTGGCTGCAAGCTTTTGCAGGT
GTAAAATTATATATGCGAGAGTGTTGCCAAGCCAAAAATTTAGCCCAGAGAGGCTTAGT
TCCCAAACAACTGAGGTAGTGTGAAGCCAGTTTTTAAAATCAAATTCAGCTCATTTCTT
TCCCAAATGCGTGTATGTTTTTAAACTTGCATGTGGTTATGTAACAGATGAGAACGAAG
TGCAAAAATCAGAATAGTTAATAATGCCTCCTGTGGTGTGAGTCTGGAGAGGTAAAGCA
TGAGATGGGAGCCCTATAAGGGTGGCCCAAAGGGAAACACTGCTCTCCTTGACCTGGTA
GGAAGGTGATTTGATGGTTTTCAAACACATGCTGTTTCTCTCTTTCATCAGGGTAGCTG
GGTGGGTGGCATTAGGTCCTTTGGACTGGGGAGAGGAGACAGAGGAAGACACCCATTAG
GACCGCCTGATGTGAGCAGTTTCAGAGGTCTGCCCTTGG

*Recursive splice site is underlined

RSU1

TGCTTGTCACTGCCTTTTTGGAAGGGACTCTGAGGAGTGTGAGCATTGAAAAGTACACA
GAAGCCCATGGAATCGTGGAGTTAAGACAGTGGAGGTCACTTAGGCCAAACTCTTGACC
TGGGTGGAGTCCTTTTCACGGCCTCCCTGAGCTGGTTGTCGTCTGGCCCCTGCTTGAAT
ACCCCCAGTGACCTGGGACTCACCACCTCAACATAGTCACTGTATAGGACTCAGGCATT
CCAGAACATTCCTTGAGGACTCAGTCTCACCACACCAGTGCTGAATAACCTTTCTTCCA
TGATAAGAAGAAAGACTAAAGCTAGCTTGGGGTAACAGCTCTGGATGTTTTTTTTTCCT
CTCCCCAGGTGGGGATGGGTGTTGGACACAATAAGCATTTTATTTCTGTTTTTTCTTGG
AGGCACAAATTAACATTTTTTCCTACCCAGGGGCCATCATACCCTAAGCTGTGCTGATT
TTTAAATTGGTCTATAGAGACAACTCGTTTCCCATGTTCGGCTGATGCTCTCCATAGCT
ACCCAAAAGTCCATTTATTTGTAACTCTAAAGCAATGACTCCCAGCCACAGCAATTTCA
CACACACCCCCACTCCCGCTGGGGACACTTGGCGCTCTCTGGAGACACTTTGGTTCCCA
CAGCTGGTGTGGGGGTGCTATTGATATCCAGTGGGTAACAGCCGGGGGTGCTGATAAAG
ACCCTACAGTTCACAGGACAGCCCCCCAAAACAAAGAATGATTTGACCCAGAACATCTC
AGTGCCAAGGCTGAGAAACTCTAGTCTAAAGGTACCATAAATAGCCAACCAGTAGGACC
ATCTGCATTTATATTTGTCTTCAGGGTTTTTTTTTAAATTTTGTTGTTTGAGATGGGAT
CTCACTCAGGCTGGAG

*Recursive splice site is underlined

ACTN4

TGCCCTGTCTGACCTTTTCTCCCCAAATGCCAAAGGCCTCTTGTCTGGATGAACAAATC
CCTGGGCGGATTTGTCCAGACTAAAACGTTAATTCTAAAGACTGGGCCCCTTAGGCACT
CGCTAAGATTCCAGGCCACACTGTCTCAAGGCCAGAATTGGCTTACTTGGCTATCTTTT
TGAAAGATCAAAGTTTAAACCAGATGATCTGTAGTCCCCCACCCCCACCCTGAAACCTG
AGCTTAGCTGTAAGCATTGAAAGTAAATGGGGTGTTTGTAGCTCACCCTCCCTGTTCTC
CAGGTGAAGGGCCCCGTGTGCCTGATCATTTCATAGCAAAATGCTAGATGGGGCCAGAG
GAGGCCCCAGCCTCTGCTGCTGCCCTAATTTTAAAACTGCCTTTTGGGAGTGTAAGTTT
CCTCTGTTAAAGGTAGTTATTTCAAGGTAGGCCTCACCATCTCCTCCTCCTGGTGAGAA
GCTCTGCCTGGAGGGCTGAGCACTGCCTCCCGCTCTGTGGGCCCCACCTGCCTTGGGTT
GAGACCTATCTCTTCCTGGACTCTGTGTGGGGAGTGCAGGCTCTTCCCCTTGGGGAGAA
CCCAGTTCTTTGACGTATAATCTGAGTGGTTTGGGTTGGTTGGTTGGTTGGTTGGTTGG
TTTCCCATGTGTGGGATGGCTCCGGAAGTCTGTTTGAGAACAGAGGCAGGC

*Recursive splice site is underlined

BACH2

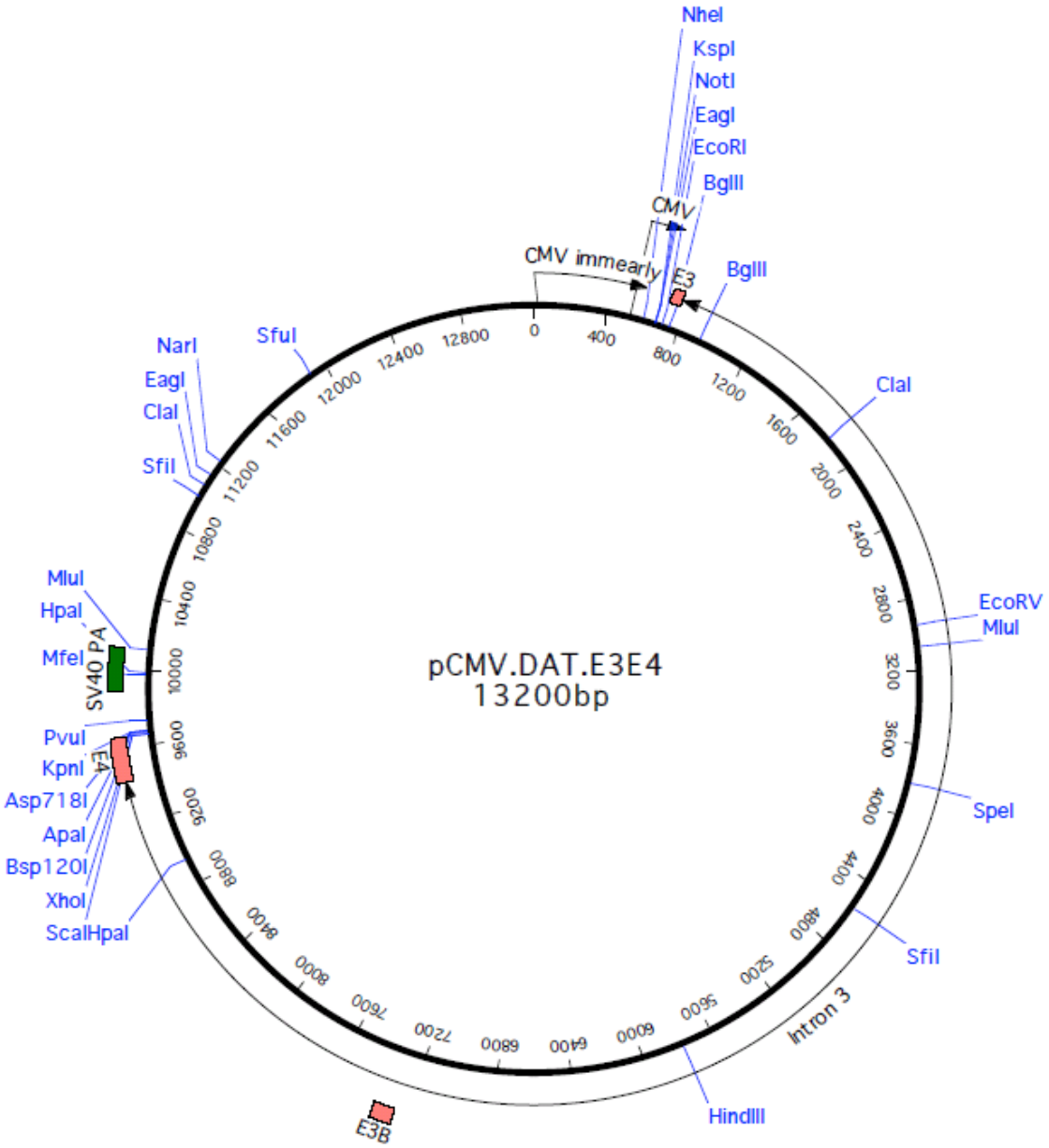
GTAACCCAGGGTGTAGCACAACTTTATAGACACAGTGTGGTCTAATCTTCTTTTCCAGT
TGTCTAAAATACAAATGGTTATCCTCCTCACAGACACACAGGATTTAAAATCTGAAAGG
ATTATGAACTTTTCCTTGACTGGCATTCTAACTAAGAATTCCTAGTGTCTCTTTTGGTA
TCATTTGCAACCTAGGTAACCAGTAAAAGCTCTGAATCCAAGCATGTCTTTTAGTCAGC
TTGAGTTCTATTGAAAAGAAACATTGTCGGGCATTTTATGTGAACTTCTGCTGTGCCAT
TTTGATTTTTTTTTAAGAAGGTGATGCTCCTTTCTGGTTTTCAAGTGTCTCCTTTAGAA
ACTTACGTTTTTTTTTAACAGGTGAGCATCTTTTTCTTTCAAAGAATTTTGAAAGCATT
GTAATAGGGAATTTAACACGCATATGCTGTTATTTAATCAATTCTTTGCTAAGCTGCTC
AGAAAACCATCCACAGAAGACTGTTCCTATATAAATGCACCTCCATTTTCCCAGGAAGA
TCATGAGTGGTTTGTTTTTACATTGGTTGTGTTGGGTCCAATGTCTTACAATTCCAAAT
TATAAGTTATTAAAGAGGACTTTGTCATCAGACCACTGTGAGTAGAACTAAAGCACATT
CCTGTGCGTGAGGAAAAGTTGCTCTGTGGCCTTCCTGCTGGTCTGATTTT

*Recursive splice site is underlined

SLC6A3.E3B

ACCCTCCACAGTGCTCTCTGGAAACAATGTGGCTCACCGACAGTGTGGCTCCCAACCTG
GCTGCCTGGGTGAGTTCACTGTGGATCACAACCCAGCCTCTCTCCTAAGGGACTCCGGA
CAGACGGTAATATAGAATTATTTAATATGGACCAGATCCACGTGGGAGAAGGCCTTCCA
AAGGCAATCCGTGACAGACTGCAATACAGAATTATTTAATATGGACCAGATCCATATGG
GAGAAGGCTTTTCAAAGGCAATCCATGACAGACTGCAATACAGAATTATTTAATATGGA
CCAGATCCACATAGGAGAAGACCTTCCAAAGGCAGCAGCTTGGCTTTCATCGTCACCAC
TACTGAGCATGCTTTCCAAGGGGGATTACCCGCACTCCTGATCTTAGATTTGTTTAAAA
CAAAGTTTTGAGTCTTCTTTTTGCTTTCAAG**GTAGGAAGAGAACTTTACTGAGGTGCCC**
TGAGCATGAGAACAGCTTCTCCTAAGGATTGAGACTATAAAAAGCAACCCAGGCCACCC
CCTGCAAAAGTCACCTTGAAGGTATGCTCCTACCCCGGCCATGAACAGGCAAGACGGCA
TGGTGCCTACTGGGTTTTAATAAAGTAAATCAAAGTTGTACCCAAACTAATCATGTCAG
TAAACTGAGAAGAAATGTGGAAATGAAAAAAATTCTTCCTGGAGCTTAGTAAAGTGAAC
CCCAGTAGCAAGAACGTGATGGTGCCCATCCAGCAGTGAACAAGGAGGAAGTCATCTGA
CCACCAGGCCCATCTGCCCACCAGTCAGGCTGACACCACTCCAAAGACTGCTGACCACT
GAGTTCTGTTCCAGTTTACCAGGAGACCCCATAAATGATGGATCCCAAATTCCAGGTCT
GTGATCCTGGAAAGGACACTCTAAAAGACCGTGGATGGCA

*Recursive splice site is underlined
*Cassette exon is bolded


pCMV.DAT.E3E4
13200bp
NheI
KspI
NotI
EagI
EcoRI
BglII
CMV
CMV immearly
E3
BglII
ClaI
EcoRV
MluI
SpeI
SfiI
Intron 3
HindIII
E3B
ScaIHpaI
Asp718I
ApaI
Bsp120I
XhoI
KpnI
PvuI
E4
SV40 PA
MfeI
HpaI
MluI
SfiI
ClaI
EagI
NarI
SfuI
0
400
800
1200
1600
2000
2400
2800
3200
3600
4000
4400
4800
5200
5600
6000
6400
6800
7200
7600
8000
8400
8800
9200
9600
10000
10400
10800
11200
11600
12000
12400
12800

Appendix B

Primer Sequences

Primers used in Chapter 2

Name:	Sequence:	Use:
5S1	GCCTGAATGCCAATTGCACCATC	mRNA/Recursive Intermediate
Hae3.1	CATCGTATGGGTAAAAGATGC	mRNA
Ubx.I3A2	GGATTTGATTTCGGGCTATGG	Recursive Intermediate
Rp49.F1	CAGTCGGATCGATATGCTAAGCTGT	Quantitative Control
Rp49.B1	CTCGACAATCTCCTTGCGCTTCTTGG	Quantitative Control
Ubx.I3A4	GTTTTTGTTTTGGTTTAAGGGTG	Construction of DE1
Ubx.RP3.F4	CATCAACCAGTTAAAGCAGCG	Construction of DE1
RP3.PB5.N.F	AGCGCCATGGGGAGAAAAACAATTTTCCGGA	Mutation of Ψ5'ssβ
RP3.PB5.N.R	CTCCCCATGGCGCTGCTTTAACTGGTTGATG	Mutation of Ψ5'ssβ
RP3.DE1mut1.F	AATCGGATCCCGGTTTGGCCGTGAGTAAATAAGTATAATAATAAAAAGAACGCAT	Mutation of Element A
RP3.DE1mut2.F	CAAAGGATCCAAATAAGTATAATAATAAAAAGAACGCATCAAC	Mutation of Ψ5'ssα
RP3.DE1mut3.F	TATCGGATCCCGGTTTGGCCAAGAACGCATCAACCAGTTAAAGC	Mutation of Element B
RP3.DE1mut1.R	ACCGGGATCCGATTTTCGTAAGGGTGTATTAAATATTTGACACTTACCTA	Mutation of Element A
RP3.DE1mut2.R	ATTTGGATCCTTTGTCAATGTTTTTGTTTTGGTTTA	Mutation of Ψ5'ssα
RP3.DE1mut3.R	ACCGGGATCCGATACTCACTTTGTCAATGTTTTTGTTTTG	Mutation of Element B
m2m3.EcoRI.F	CAAGAATTCATCGGATCCCGGTTTGCCAAGAACGCATCAAC	Construction of mut2mut3
m2m3.EcoRI.R	CGATGAATTCTTTGTCAATGTTTTTGTTTTGGTTTA	Construction of mut2mut3
RP3.mut1mut3.F1	ATTCGAATTCCGGTTTGGCCGTGAGTATCGGATCCCGGTTTGGCCAAGAACGCAT	Construction of mut1mut3
RP3.mut1mut2.F1	ATTCGAATTCCGGTTTGGCCGGATCCAAATAAGTATAATAATAAAAAGAACGCAT	Construction of mut1mut2
RP3.5ssAR2	TTATACTTATTTGCTCAGTTTGTCAATGTT	Construction of Ψ5'ssα compensatory mutatior
RP3.5ssAF2	AACATTGACAAACTGAGCAAATAAGTATAA	Construction of Ψ5'ssα compensatory mutatior
dU1.21D.R1	GCAGTTCTCCACCTTCGACT	Cloning of U1 snRNA/Compensatory mutation
dU1.21D.F1	CTCGTTGACCGCAAATTTCT	Cloning of U1 snRNA/Compensatory mutation
dU1.mut+1+6.F1	GAAAGCATGCTTAGCTGGCGTAG	Compensatory mutation
dU1.21D.R2	CTCAGCTCAGGGAATGGG	Compensatory mutation

Primers used in Chapter 3

Name:	Sequence:	Use:
RP3.TK.A.F2	TTTTGCGGCCGCTTTTTTGCTCGCACGGATTC	Construction of pPTV2.ΔRP3.ISceI
RP3.del.B.R	AAAAAGATCTTTAGGCAAACAGTGAGTACAAAAAGTAG	Construction of pPTV2.ΔRP3.ISceI
RP3.del.B.F	AAAAAGATCTGTCAAATATTTAATACACCCTTAAACC	Construction of pPTV2.ΔRP3.ISceI
RP3.TK.B.R	TTTTGGTACCTCAATAGACCAATGCGAGACCAG	Construction of pPTV2.ΔRP3.ISceI
RP3.TK.A.R	TAGGGATAACAGGGTAATTTAGCCTGATTTATTTGTCGGTCTG	Construction of pPTV2.ΔRP3.ISceI
RP3.TK.B.F	ATTCAATTTCCACGCAATATTCC	Construction of pPTV2.ΔRP3.ISceI
downst.RR4	ATTACCCCTTTCAGGCGTTT	Molecular verification
downst.RR3	TGTGAAGGGCTACGAAAGTACA	Molecular verification
pTV2.FRT.F4	ACCTCTACATCAACAGGCTTCC	Molecular verification
pTV2.FRT.F3	CTGAAGGAAGCATACGATACCC	Molecular verification
dRP3.upst.R	TCGGAGGATGTAGGATGGAG	Molecular verification
wtRP3downst.F	CAAAGTCCATCCCTTCCTGA	Molecular verification
pTV2.CreI.R3	AGGCGGACATTGACGCTATC	Molecular verification
pTV2.CreI.R2	CTGCCTCCGCGAATTAATAG	Molecular verification
upst.LF3	GACGTCGAGGCAAAAACTTC	Molecular verification
upst.LF2	AAAAACCATCCACGAACGAG	Molecular verification
downst.RR2	TCCAAGATGGATTGCTGTGA	Molecular verification
downst.RR	CGGACAGTATGGCAGCACTA	Molecular verification
deltaRP3.RF	TGTACTCACTGTTTGCCTAAAGATCT	Molecular verification
deltaRP3.R1	GGGTGTATTAAATATTTGACAGATCT	Molecular verification
pTV2.CreI.R	GCAAACTGCTCACGACGTTTTG	Molecular verification

pTV.FRT.F1	GTTACAGTCCGGTGCGTTTT	Molecular verification
pTV2.CreI.F	TGTACTCACTGTTTGCCTAATACTAAT	Molecular verification
wtRP3.LR	ATATTTGACACTTACCTAGAAAAGAG	Molecular verification
upst.LF	AGGAAGCAAATGGCAGCTAA	Molecular verification

Primers used in Chapter 4

Name:	Sequence:	Use:
HF7.F1	TCCTGTTGTTGGTGAATGGAGC	Construction of pMRSR
HF7.B1	CAGCGTCCTCTCAGAGAACGTC	Construction of pMRSR/mRNA analysis
MRSR.F1	ACAAAAGCTGGAGCTCAGTTGTGTG	Construction of pMRSR/mRNA analysis
MRSR.Xba.R1	AGTGGCCCTCTAGAGTGCTCGTC	Construction of pMRSR
MRSR.RV.F1	ACTGTGGAGATATCGGGGCAC	Construction of pMRSR
hF7.m6.F1	AAAAGGATCCGCGGTGCCAGGTGAGTACCACTCTCCCCTGTCTG	Construction of pMRSR
hF7.md25.R1	TTTTGGATCCAGCACCGCGGTCGGAC	Construction of pMRSR
Cdh4.RP246886.R1	AAAACGACAAAGTAGACCACCAG	Cloning of RSS into pMRSR
Cdh4.RP246886.F1.X	AAAATCTAGAGTACTTGGCCACGGAGAGTC	Cloning of RSS into pMRSR
BACH2.F1	AAAATCTAGACCATTTCCTTCAGCCTTTGA	Cloning of RSS into pMRSR
BACH2.R1	AAATCAGACCAGCAGGAAGG	Cloning of RSS into pMRSR
ACTN4.F1	AAAATCTAGATGCCCTGTCTGACCTTTTCT	Cloning of RSS into pMRSR
ACTN4.R1	GCCTGCCTCTGTTCTCAAAC	Cloning of RSS into pMRSR
RSU1.F1	AAAATCTAGATGCTTGTCACTGCCTTTTTG	Cloning of RSS into pMRSR
RSU1.R1	CTCCAGCCTGAGTGAGATCC	Cloning of RSS into pMRSR
MAPKAPK2.F1	AAAATCTAGAGATGGGAAAGTGAAGCAATG	Cloning of RSS into pMRSR
MAPKAPK2.R1	AGGCCAAAATCAAACCTGAC	Cloning of RSS into pMRSR
GRK5.F1	AAAATCTAGAGTGTCCAAACCCACACTGAG	Cloning of RSS into pMRSR
GRK5.R1	TTAGGGGACCATGATTCAGC	Cloning of RSS into pMRSR
SLIT3.F1	AAAATCTAGAAATTGGATCAGGTGGACTGG	Cloning of RSS into pMRSR

SLIT3.R1	CAAGGGCAGACCTCTGAAAC	Cloning of RSS into pMRSR
hCDH4.I2B1	TTTACACGGGGTCCTCACTC	Recursive intermediate analysis
BACH2.R2	TGTGGATGGTTTTCTGAGCA	Recursive intermediate analysis
RSU1.R2	GGGTGTGTGTGAAATTGCTG	Recursive intermediate analysis
ACTN4.R2	AAAATTAGGGCAGCAGCAGA	Recursive intermediate analysis
SLIT3.R2	AATTTTTGGCTTGGCAACAC	Recursive intermediate analysis
GRK5.R2	AGCTGAGAGGAACCTTGCTG	Recursive intermediate analysis
MAPKAPK2.R2	AGTGTCCAGCGAGAGGAGAC	Recursive intermediate analysis
MRSR.IR2	AGCCCCCAGTCTTTTATCGT	Recursive lariat analysis
MRSR.IF2	CCAGATTCACCCCAGTTCAC	Recursive lariat analysis
MRSR.IR1	GCTCGTCTCACCCATAAACC	Recursive lariat analysis
MRSR.IF1	AGGGCACAGCATCCCTTC	Recursive lariat analysis
CDH4.RP*.R	ACCGGAATCCGGGACCTAGAAAGGAG	Mutating the 5'ss component
CDH4.RP*.F	GCTAACGCTGGACACGTGTGTC	Mutating the 5'ss component
MAPKAPK2.RP*.R	CAACAGCTCCGGGACCTGGAGGA	Mutating the 5'ss component
MAPKAPK2.RP*.F	CCTCCAGGTCCCGGAGCTGTTGC	Mutating the 5'ss component
BACH2.RP*.R	GACCGGGACCTGTTAAAAAAAAACGTAAG	Mutating the 5'ss component
BACH2.RP*.F	TTTTTCTTTCAAAGAATTTTGAAAGCATTG	Mutating the 5'ss component
DAT.RSS2.R	TGCCATCCACGGTCTTTTAG	Cloning of *SLC6A3* RSS into pMRSR
DAT.RSS2.XF	AAAATCTAGAACCCTCCACAGTGCTCTCTG	Cloning of *SLC6A3* RSS into pMRSR
DAT.RSS2*R	AAGTTCTCTTGGGACCTTGAAAGC	Mutating the 5'ss component
DAT.RSS2*F	TACTGAGGTGCCCTGAGCAT	Mutating the 5'ss component
hDAT.E3.RI.F1	AAAAGAATTCGCCAGTTCAACAGGGAAG	Cloning of E3-I3-E4 portion of *SLC6A3* into pCMV
hDAT.E4.XI.B1	AAAACTCGAGAAGTACTCGGCAGCAGGT	Cloning of E3-I3-E4 portion of *SLC6A3* into pCMV
pCMV.F1	ACGCCAAGCTCGAAATTAAC	pCMV.SLC6A3.E3E4 and haplotype mRNA analysis
pCMV.R1	GAAGGGCGATCGAGTGAA	pCMV.SLC6A3.E3E4 and haplotype mRNA analysis
hDAT.E3F1	GCCAGTTCAACAGGGAAGG	*SLC6A3* mRNA analysis
hDAT.E4R2	GAAGGAGGAGAAGAGATAGTGCA	*SLC6A3* mRNA analysis

hDAT.E2F1	GAGACCTGGGGCAAGAAGAT	*SLC6A3* mRNA analysis
hDAT.E5R1	GTAGAGCAGCACGATGACCA	*SLC6A3* mRNA analysis
hDAT.I3R1	GGGTTCACTTTACTAAGCTCCAGG	*SLC6A3* mRNA analysis
hDAT.I3R2	GGCCTGGGTTGCTTTTTATAG	*SLC6A3* mRNA analysis
hDAT.sE3b.A	AGCTTGAAACCCTGGGAAGT	SNP swap experiments
hDAT.sE3b.B	AAGGACTTGATACAGAAAGTTTTAACC	SNP swap experiments
hDAT.sE3b.C	GAGCCATCCAAGGTCACACT	SNP swap experiments
hDAT.sE3b.D	CACATGGCTGTAAATGAGCTCA	SNP swap experiments
hDAT.sE4.E	TAGGGAGCCCATGCAAATAG	SNP swap experiments
hDAT.sE4.F	GGTAGCCCTGGGTGCTTCT	SNP swap experiments
hDAT.I3.H	GCTTCCTGGGAGTCAGACAG	SNP swap experiments

References

Aguilera, A, & Gomez-gonzalez, B. (2008). Genome instability: a mechanistic view of its causes and consequences. *Nat. Rev. Genet. 9*, 204-217.

Alexander, M. R., Wheatley, A. K., & Purcell, D. F. J. (2010). Efficient transcription through an intron requires the binding of an Sm-type U1 snRNP with intact stem loop II to the splice donor. *Nucleic Acids Research*. (16), 1-13.

Amara, S. G., & Sonders, M. S. (1993). Neurotransmitter transporters as molecular targets for addictive drugs. *Drug and alcohol dependence*. 51(1-2), 87-96.

Ashton-Beaucage, D., Udell, C. M., Lavoie, H., Baril, C., Lefrançois, M., Chagnon, P. (2010). The exon junction complex controls the splicing of MAPK and other long intron-containing transcripts in Drosophila. *Cell*, 143(2): 251-62.

Bannon MJ, Michelhaugh SK, Wang J, Sacchetti P. (2001). The human dopamine transporter gene: Gene organization, transcriptional regulation, and potential involvement in neuropsychiatric disorders. *Eur Neuropsychopharmacol.* 11(6):449–455.

Bender, W., & Hudson, A. (2000). P element homing to the Drosophila bithorax complex. *Development*. 127: 3981-3992.

Bender, W., Weiffenbach, B., Karch, F., Peifer, M. (1985). Domains of cis-interaction in the Bithorax complex. *Cold Spring Harbor Symp. Quant. Biol*. 50: 173-180.

Bentley, D. L. (2005). Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Current Opinion in Cell Biology*. *15*: 251-256

Berget SM (1995) Exon recognition in vertebrate splicing. *J. Biol. Chem*. 270:2411-2414.

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Review Literature And Arts Of The Americas*. 72:291-336.

Blanchette, M., & Chabot, B. (1999). Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization. *The EMBO journal*. 18(7): 1939-52.

Borensztajn, K., Sobrier, M.-L., Duquesnoy, P., Fischer, A.-M., Tapon-Bretaudière, J., & Amselem, S. (2006). Oriented scanning is the leading mechanism underlying 5' splice site selection in mammals. *PLoS genetics*. *2*(9): 1297-1306.

Botas, J., Cabrera, C.V., Garcia-Bellido, A. (1988). The reinforcement-extinction process of selector gene activity: a positive feed-back loop and cell-cell interactions in Ultrabithorax patterning. *Rouxs Arch. Dev. Biol*. 197: 424-434.

Brogna, S., & Wen, J. (2009). Nonsense-mediated mRNA decay (NMD) mechanisms. *Online*. 16(2): 107-113.

Burge CB, Tuschl T, Sharp PA. (1999). *Splicing signals. In The RNA World II* Edited by: Gesteland RF, Cech TR, Atkins JF. Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press; pp 525-560.

Burnette, J. M., Hatton, Allyson R, & Lopez, A Javier. (1999). Trans-acting Factors Required for Inclusion of Regulated Exons in the Ultrabithorax mRNAs of Drosophila melanogaster. *Genetics*. 151:1517-1529.

Burnette, J. M., Miyamoto-sato, E., Schaub, M. A., Conklin, J., & Lopez, A Javier. (2005). Subdivision of Large Introns in Drosophila by Recursive Splicing at Nonexonic Elements. *Genetics*. 674: 661-674.

Chasin LA (2007) Serching for splicing motifs. *Adv. Exp. Med. Biol*. 623:85-106.

Chavez, S., Traude, B., Rondon, A. G., Erdjument-Bromage, H., Tempst, P., Svejstrup, J. Q. (2000). A protein complex containing Tho2, Hpr1, Mft1 and a novel protein, Thp2, connects transcription elongation with mitotic recombination in Saccharomyces cerevisiae. *EMBO Journal*. 19(21): 5824-5834.

Chen, L. L., Sabripour, M., Wu, E. F., Prieto, V. G., Fuller, G. N., & Frazier, M. L. (2005). A mutation-created novel intra-exonic pre-mRNA splice site causes constitutive activation of KIT in human gastrointestinal stromal tumors. *Oncogene*. *24*(26): 4271-80

Chen, Y., Weeks, J., Mortin, M. a, & Greenleaf, a L. (1993). Mapping mutations in genes encoding the two large subunits of Drosophila RNA polymerase II defines domains essential for basic transcription functions and for proper expression of developmental genes. *Molecular and cellular biology*. *13*(7): 4214-22.

Coles, J. L., Hallegger, M., & Smith, C. W. J. (2009). A nonsense exon in the Tpm1 gene is silenced by hnRNP H and F A nonsense exon in the Tpm1 gene is silenced by hnRNP H and F. *Spring*. 15: 33-43.

Comeron, J. M., & Kreitman, M. (2000). The correlation between intron length and recombination in Drosophila: Dynamic equilibrium between mutational and selective forces. *Genetics*. 156: 1175-1190.

Comeron, J. M., & Kreitman, M. (2002). Population, evolutionary and genomic consequences of interference selection. *Genetics*. 161: 389-410.

Conklin, J. F., Goldman, A., & Lopez, A Javier. (2005). Stabilization and analysis of intron lariats in vivo. *Design*. 37: 368-375.

Cooper, Thomas A, Wan, L., & Dreyfuss, G. (2010). RNA and Disease. *Cell*. 136(4): 777-793.

Coulter, D. E., & Greenleaf, a L. (1985). A mutation in the largest subunit of RNA polymerase II alters RNA chain elongation in vitro. *The Journal of biological chemistry*. 260(24): 13190-13198.

Cragg, S. J., & Rice, M. E. (2004). DAncing past the DAT at a DA synapse. *Trends in neurosciences*. 27(5): 270-7.

Das, R., Dufu, K., Romney, B., Feldt, M., Elenko, M., & Reed, R. (2006). Functional coupling of RNAP II transcription to spliceosome assembly. *Genes & Development*. 20: 1100-1109.

Das, R., Yu, J., Zhang, Zuo, Gygi, M. P., Krainer, A. R., Gygi, S. P., et al. (2007). SR proteins function in coupling RNAP II transcription to pre-mRNA splicing. *Molecular cell*. 26(6), 867-81.

Deutsch, M., & Long, M. (1999). Intron-exon structures of eukaryotic model organisms. *Nucleic acids research*. 27(15): 3219-28.

Dewey, C. N., Rogozin, I. B., & Koonin, E. V. (2006). Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC genomics*. 7: 311.

Dhir, A., Buratti, Emanuele, Santen, M. A. V., Lu, R., & Baralle, Francisco E. (2010). The intronic splicing code: multiple factors involved in ATM pseudoexon definition. *EMBO Journal*. 29(4): 749-760.

Dolezal, T., Gazi, M., Zurovec, M., & Bryant, P. J. (2003). Genetic analysis of the ADGF multigene family by homologous recombination and gene conversion in Drosophila. *Genetics*. 165(2): 653-666.

Dominguez, C., & Allain, F. H.-T. (2006). NMR structure of the three quasi RNA recognition motifs (qRRMs) of human hnRNP F and interaction studies with Bcl-x G-tract RNA: a novel mode of RNA recognition. *Nucleic acids research*. 34(13): 3634-3645.

Duncan, I. W. (2002). Transvection effects in Drosophila. *Annual review of genetics*. 36: 521-56.

Dye, M. J., Gromak, N., & Proudfoot, N. J. (2006). Exon tethering in transcription by RNA polymerase II. *Molecular cell*. *21*(6): 849-59.

Elmore, T., Ignell, R., Carlson, J. R., & Smith, D. P. (2003). Targeted mutation of a Drosophila odor receptor defines receptor requirement in a novel class of sensillum. *The Journal of neuroscience: the official journal of the Society for Neuroscience*. *23*(30): 9906-12.

Fairbrother, W G, & Chasin, L. A. (2000). Human genomic sequences that inhibit splicing. *Mol. Cell. Biol*. *20*: 6816-6825.

Fairbrother, William G, Yeh, R.-F., Sharp, P. a, & Burge, C. B. (2002). Predictive identification of exonic splicing enhancers in human genes. *Science*. *297*(5583): 1007-13.

Farde L, Halldin C, Muller L, Suhara T, Karlsson P, Hall H. (1994). PET study of [11C]beta-CIT binding tomonoamine transporters in the monkey and human brain. *Synapse*. 16(2):93–103.

Faustino, N. A., & Cooper, T A. (2003). Pre-mRNA splicing and human disease. *Genes Dev*. *17*: 419-437.

Fedorova, L., & Fedorov, A. (2003). Introns in gene evolution. *Genetica*. 118: 123-131.

Fedorova, O., & Zingler, N. (2007). Group II introns: structure, folding and splicing mechanism. *Biological chemistry*. *388*(7): 665-78.

Fisette, J.-F., Toutant, J., Dugré-Brisson, S., Desgroseillers, L., & Chabot, B. (2010). hnRNP A1 and hnRNP H can collaborate to modulate 5' splice site selection. *RNA*, 16(1): 228-38.

Fong, Y. W., & Zhou, Q. (2001). Stimulatory effect of splicing factors on transcriptional elongation. *Nature*. *414*(6866): 929-33.

Gainetdinov RR, Sotnikova TD, Caron MG. (2002). Monoamine transporter pharmacology and mutant mice. *Trends Pharmacol Sci.* 23(8):367–373.

Giles, K. E., & Beemon, K. L. (2005). Retroviral Splicing Suppressor Sequesters a 3J Splice Site in a 50S Aberrant Splicing Complex. *Society*. *25*(11): 4397-4405.

Giros B, el Mestikawy S, Godinot N, Zheng K, Han H, Yang-Feng T, Caron MG. (1992). Cloning, pharmacological characterization, and chromosome assignment of the human dopamine transporter. *Mol Pharmacol*. 42(3):383–390.

Giros B, Caron MG. (1993). Molecular characterization of the dopamine transporter. *Trends Pharmacol Sci*. 14(2):43–49.

Gonzalez-Gaitan, M. A., Micol, J.-L., & Garcia-Bellido, A. (1990). Developmental genetic analysis Contrabithorax Mutations in Drosophila melanogaster. *Genetics*. *126*: 139-155.

Greenleaf AL, Weeks JR, Voelker RA, Ohnishi S, Dickson B. (1980). Genetic and biocheemical characterization of mutants at an RNA polymerase II locus in D. melanogaster. *Cell*. 21: 785-792.

Grellscheid, S.-nagaraja, & Smith, C. W. J. (2006). An Apparent Pseudo-Exon Acts both as an Alternative Exon That Leads to Nonsense-Mediated Decay and as a Zero-Length Exon. *Society*. *26*(6): 2237-2246.

Grimaldi D, Engel MS. (2005). *Evolution of the insects*. Cambridge Univ. Press, Cambridge.

Gurvich, O. L., Tuohy, T. M., Howard, M. T., Finkel, R. S., Medne, L., Anderson, C. B., et al. (2008). DMD pseudoexon mutations: splicing efficiency, phenotype, and potential therapy. *Annals of neurology*. *63*(1): 81-89.

Hall DA, Strange PG. (1999). Comparison of the ability of dopamine receptor agonists to inhibit forskolin-stimulated adenosine 3’5’-cyclic monophosphate (cAMP) accumulation via D2L (long isoform) and D3 receptors expressed in Chinese hamster ovary (CHO) cells. *Biochem Pharmacol*. 58(2):285–289.

Hatton, A R, Subramaniam, V, & Lopez, A J. (1998). Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon exon junctions. *Mol. Cell*. *2*: 787-796.

Haugen, P., Simon, D. M., & Bhattacharya, D. (2005). The natural history of group I introns. *Trends in genetics. TIG*. *21*(2): 111-9.

Hibbert, C. S., Gontarek, R. R., & Beemon, K. L. (1999). The role of overlapping U1 and U11 5’ splice site sequences in a negative regulator of splicing. *Spring*. 5: 333-343.

Huertas, P., & Gene, D. D. (2003). Cotranscriptionally Formed DNA: RNA Hybrids Mediate Transcription Elongation Impairment and Transcription-Associated Recombination. *Science And Technology*. *12*: 711-721.

Jurica, M. S., & Moore, M. J. (2003). Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell*. *12*: 5-14.

Kandul, N. P., & Noor, M. a F. (2009). Large introns in relation to alternative splicing and gene evolution: a case study of Drosophila bruno-3. *BMC genetics*. *10*: 67.

Kornblihtt, A R, Mata, M. D. L., Fededa, J P, Oz, M. J. M. ˜, & Nogues, G. (2004). Multiple links between transcription and splicing. *RNA*. *10*: 1489-1498.

Kornfeld, K., Saint, R. B., Beachy, P. A., Harte, P. J., Peattie, D. A., & Hogness, David S. (1989). Structure and expression of a family of Ultrabithorax mRNAs generated by alternative splicing and polyadenylation in Drosophila. *Genes & Development*. 3: 243-258.

Lankenau, S., Barnickel, T., Marhold, J., Lyko, F., Mechler, B. M., & Lankenau, D.-H. (2003). Knockout targeting of the Drosophila nap1 gene and examination of DNA repair tracts in the recombination products. *Genetics*. *163*(2): 611-23.

Lenasi, T., & Barboric, M. (2010). P-TEFb stimulates transcription elongation and pre-mRNA splicing through multilateral mechanisms. *Rna Biology*. 7(2): 145-150.

Lewis EB. (1954). The theory and application of a new method of detecting chromosomal rearrangements in Drosophila melanogaster. *Am. Nat*. 88:225–39.

Li, X., & Manley, J. L. (2005). Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell*. *122*(3): 365-78.

Lin, S., Coutinho-Mansfield, G., Wang, D., Pandit, S., & Fu, X.-D. (2009). The splicing factor SC35 has an active role in transcriptional elongation. *Molecular Pathology*. *15*(8): 819-826.

Long, J. C., & Caceres, J. F. (2009). The SR protein family of splicing factors: master regulators of gene expression. *Society*. *27*: 15-27.

Lopez, a J., & Hogness, D S. (1991). Immunochemical dissection of the Ultrabithorax homeoprotein family in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*. *88*(22): 9924-8.

Lopez, A.J., Artero, R., and Pérez-Alonso, M. (1996). Stage, tissue and cell-specific expression of alternative *Ultrabithorax* mRNAs and protein isoforms in the Drosophila embryo. *Roux's Arch. Dev. Biol.* 205: 450-459.

Lopez, A Javier. (1998). Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annual review of genetics*. 32: 279-305.

Lou, H., Gagel, R. F., & Berget, S. M. (1996). An intron enhancer recognized by splicing factors activates polyadenylation. *Genes & Development*. 10(2): 208-219.

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*. 21:1-7.

Maeda, R. K., & Karch, F. (2006). The ABC of the BX-C: the bithorax complex explained. *Development*. 133(8): 1413-22.

Martinez-Contreras, R., Fisette, J.-F., Nasim, F.-ul H., Madden, R., Cordeau, M., & Chabot, B. (2006). Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. *PLoS biology*. 4(2): e21.

Maniatis, T., & Reed, R. (2002). An extensive network of coupling among gene expression machines. *Nature*. 416: 499-506.

Maquat, L. E., & Gong, C. (2009). Gene expression networks: competing mRNA decay pathways in mammalian cells. *Biochemical Society transactions*. 37: 1287-92.

Mata, M. D., Lafaille, C., Kornblihtt, Alberto R. (2010). First come, first served revisited: Factors affecting the same alternative splicing event have different effects on the relative rates of intron removal. *RNA*. 16: 904-912.

Mata, M. D., Alonso, C. R., Fededa, Juan P, Pelisch, F., Cramer, P., Bentley, D., et al. (2003). A slow RNA polymerase II affects alternative splicing in vivo. *12*: 525-532.

Matlin, A. J., Clark, F., & Smith, C. W. (2005). Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol. 6*: 386-398.

McCall, K., O′Connor, M. B., & Bender, W. (1994). Enhancer traps in the Drosophila bithorax complex mark parasegmental domains. *Genetics*. *138*(2): 387-99.

McCullough, a J., & Berget, S. M. (2000). An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5' splice sites. *Molecular and cellular biology*. *20*(24): 9225-35.

Mcglincy, N. J., & Smith, C. W. (2008). Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense. *Trends Biochem. Sci. 33*: 385-393.

Micol, J. L., & García-Bellido, a. (1988). Genetic analysis of "transvection" effects involving contrabithorax mutations in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*. *85*(4): 1146-50.

Mineo, I., Clarke, P. R., Sabina, R. L., & Holmes, E. W. (1990). A novel pathway for alternative splicing: identification of an RNA intermediate that generates an alternative 5' splice donor site not present in the primary transcript of AMPD1. *Molecular and cellular biology*. *10*(10): 5271-8.

Mineo, I., & Holmes, E. W. (1991). Exon recognition and nucleocytoplasmic partitioning determine AMPD1 alternative transcript production. *Molecular and cellular biology*. *11*(10): 5356-63.

Mortin MA, Lefevre G Jr. (1981). An RNA polymerase II mutation in Drosophila melanogaster that mimics ultrabithorax. *Chromosoma*. 82:237-247.

Mozley, L. H., Gur, R. C., Mozley, P. D., & Gur, R. E. (2001). Striatal dopamine transporters and cognitive functioning in healthy men and women. *The American journal of psychiatry*. *158*(9): 1492-9.

Muñoz, M. J., Mata, M. de la, & Kornblihtt, Alberto R. (2010). The carboxy terminal domain of RNA polymerase II and alternative splicing. *Trends in biochemical sciences*. *35*(9): 497-504.

Ooi, S. L., Dann, C., Nam, K., Leahy, D. J., Damha, M. J., & Boeke, J. E. F. D. (2001). RNA Lariat Debranching Enzyme. *Methods*. *342*(1984): 233-248.

O'Connor, M. B., Binari, R., Perkins, L. a, & Bender, W. (1988). Alternative RNA products from the Ultrabithorax domain of the bithorax complex. *The EMBO journal*. *7*(2): 435-45.

Pagani, F., Buratti, E, Stuani, C., Bendix, R., Dork, T., & Baralle, F E. (2002). A new type of mutation causes a splicing defect in ATM. *Nat. Genet*. *30*: 426-429.

Pandya-Jones, A., & Black, D. L. (2009). Co-transcriptional splicing of constitutive and alternative exons. *RNA. 15*(10): 1896-908.

Peifer, M., & Bender, W. (1986). The anterobithorax and bithorax mutations of the bithorax complex. *The EMBO journal*. *5*(9): 2293-303.

Pessa, H. K. J., & Frilander, M. J. (2011). Minor Splicing, disrupted. *Science*. *332*(6026): 184-185.

Pozzoli, U., & Sironi, M. (2005). Cellular and Molecular Life Sciences Silencers regulate both constitutive and alternative splicing events in mammals. *Cellular and Molecular Life Sciences*. *62*: 1579-1604.

Reed, H. C., Hoare, T., Thomsen, S., Weaver, T. a, White, R. a H., Akam, M., et al. (2010). Alternative splicing modulates Ubx protein function in Drosophila melanogaster. *Genetics*. *184*(3): 745-58.

Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita P, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. (2010). The UCSC Genome Browser database: update 2010. *Nucleic Acids Res*. 38: D613-619.

Rong, Y. S. (2002). Gene targeting by homologous recombination: a powerful addition to the genetic arsenal for Drosophila geneticists. *Biochemical and Biophysical Research Communications*. *297*: 1-5.

Rong, Y. S., & Golic, Kent G. (2001). A Targeted Gene Knockout in Drosophila. *Genetics*. *1312*: 1307-1312.

Ruden, D. M., & Jäckle, H. (1995). Mitotic delay dependent survival identifies components of cell cycle control in the Drosophila blastoderm. *Development. 121*(1): 63-73.

Russo, C. a, Takezaki, N., & Nei, M. (1995). Molecular phylogeny and divergence times of drosophilid species. *Molecular biology and evolution*. *12*(3): 391-404.

Salz, H. K. (2011). Sex determination in insects: a binary decision based on alternative splicing. *Current opinion in genetics & development*. 21: 1-6.

Savitsky, M., Kwon, D., Georgiev, P., Kalmykova, A., & Gvozdev, V. (2006). Telomere elongation is under the control of the RNAi-based mechanism in the Drosophila germline. *Genes & development*. *20*(3): 345-54.

Schellenberg, M., Ritchie, D., & MacMillan, A. (2008). Pre-mRNA splicing: a complex picture in higher definition. *Trends in Biochemical Sciences*, *33*(6): 242-243.

Schmucker, D., & Chen, B. (2009). Dscam and DSCAM: complex genes in simple animals, complex animals yet simple genes. *Genes & development*. *23*(2): 147-56.

Siebel CW, Fresco LD, Rio DC. (1992) The mechanism of somatic inhibition of Drosophila P-element pre-mRNA splicing: multiprotein complexes at an exon pseudo-5' splice site control U1 snRNP binding. *Genes Dev*. 6: 1386–1401.

Selth, L. a, Sigurdsson, S., & Svejstrup, J. Q. (2010). Transcript elongation by RNA Polymerase II. *Annual review of biochemistry*. *79*: 271-93.

Shepard, S., Mccreary, M., & Fedorov, A. (2009). The Peculiarities of Large Intron Splicing in Animals. *Science*, *4*(11). doi: 10.1371/journal.pone.0007853.

Shpiz, S., Kwon, D., Uneva, A., Kim, M., Klenov, M., Rozovsky, Y., et al. (2007). Characterization of Drosophila telomeric retroelement TAHRE: transcription, transpositions, and RNAi-based regulation of expression. *Molecular biology and evolution*. *24*(11): 2535-45.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*. *15*(8): 1034-50.

Silva, A. L., & Romão, L. (2009). The mammalian nonsense-mediated mRNA decay pathway: to decay or not to decay! Which players make the decision? *FEBS letters*. *583*(3): 499-505.

Sims, R. J., Belotserkovskaya, R., & Reinberg, D. (2004). Elongation by RNA polymerase II: the short and long of it. *Genes & development*. *18*(20): 2437-68.

Singh, J., & Padgett, R. A. (2009). Rates of in situ transcription and splicing in large human genes. *Nat. Struct. Mol. Biol*. *16*(11): 1128-1133.

Sironi, M., Menozzi, G., Riva, L., Cagliani, R., & Comi, G. P. (2004). Silencer elements as possible inhibitors of pseudoexon splicing. *Nucleic Acids Res*. *32*: 1738-1791.

Smith, C. W., & Valcarcel, J. (2000). Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci. 25*: 381-388.

Sotnikova TD, Beaulieu JM, Gainetdinov RR, Caron MG. (2006). Molecular biology, pharmacology and functional role of the plasma membrane dopamine transporter. *CNS Neurol Disord Drug Targets.* 5(1):45–56.

Stalder, L., & Mühlemann, O. (2008). The meaning of nonsense. *Trends in cell biology*. *18*(7): 315-21.

Stamm, S., Ben-ari, S., Rafalska, I., Tang, Y., Zhang, Zhaiyi, Toiber, D., et al. (2005). Function of alternative splicing. *Gene. 344*: 1 - 20.

Subramaniam, Vaidyanathan, Bomze, H. M., & Lopez, A Javier. (1994). Functional Differences Between Ultrabithorax Protein Isoforms. *Genetics.* 136: 979-*991*.

Sugnet, C. W., Srinivasan, K., Clark, T. a, O'Brien, G., Cline, M. S., Wang, H., et al. (2006). Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS computational biology*. *2*(1): e4.

Talkowski, M. E., Kirov, G., Bamne, M., Georgieva, L., Torres, G., Mansour, H., et al. (2008). A network of dopaminergic gene variations implicated as risk factors for schizophrenia. *Human molecular genetics*. *17*(5): 747-58.

Tamura, K., Subramanian, S., & Kumar, S. (2004). Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Molecular biology and evolution*. *21*(1): 36-44.

Tazi, J., Bakkour, N., & Stamm, S. (2009). Alternative splicing and disease. *Biochimica et Biophysica Acta*. *1792*(1): 14-26.

Thummel C. (1992). Mechanisms of transcriptional timing in Drosophila. *Science*. 3: 39-40.

Torres, G. E., Gainetdinov, R. R., & Caron, M. G. (2003). Plasma membrane monoamine transporters: structure, regulation and function. *Nature reviews. Neuroscience*. 4(1): 13-25.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., et al. (2006). An RNA map predicting Nova-dependent splicing regulation. *Nature*. 444(7119): 580-6.

Vagin, V. V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., & Zamore, P. D. (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. *Science.* 313(5785): 320-4.

Vagner, S., Ruegsegger, U., Gunderson, S. I., Keller, W., & Mattaj, I. W. (2000). Position-dependent inhibition of the cleavage step of pre-mRNA 3J-end processing by U1 snRNP. *RNA*. *6*: 178-188.

Vandenbergh DJ, Persico AM, Hawkins AL, Griffin CA, Li X, Jabs EW, Uhl GR. (1992). Human dopamine transporter gene (DAT1) maps to chromosome 5p15.3 and displays a VNTR. *Genomics*. 14(4):1104–1106.

Venables, J. P. (2004). Aberrant and Alternative Splicing in Cancer. *International Journal of Cancer*. 64: 7647-7654.

Vinogradov, A. E. (2006). "Genome design" model and multicellular complexity: golden middle. *Nucleic Acids Research*, *34*(20): 5906-5914.

Voynov, V., Verstrepen, K. J., Jansen, A., Runner, V. M., Buratowski, S., & Fink, G. R. (2006). Genes with internal repeats require the THO complex for transcription. *Proceedings of the National Academy of Sciences of the United States of America*. *103*(39): 14423-8.

Wahl, M. C., Will, C. L., & Lu, R. (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell*. 136: 701-718.

Wang, Z., & Burge, C. B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA*. 617: 802-813.

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): 831-45.

Weir, M., Eaton, M., & Rice, M. (2006). Challenging the spliceosome machine. *Genome biology*. *7*(1):R3.

Weir, M., & Rice, M. (2004). Ordered partitioning reveals extended splice-site consensus information. *Genome research*. *14*(1): 67-78.

Wesolowska, N., & Rong, Y. S. (2010). The past, present and future of gene targeting in Drosophila. *Fly*. *4*(1): 53-9.

Will, C. L., & Lührmann, R. (2005). Splicing of a rare class of introns by the U12-dependent spliceosome. *Biological chemistry*. *386*(8): 713-24.

Xie, H. B., & Golic, K G. (2004). Gene deletions by ends-in targeting in Drosophila melanogaster. *Genetics*. *168*: 1477-1489.

Xu, Q., Modrek, B., & Lee, C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic acids research*. *30*(17): 3754-66.

Yeo, G. W., Van Nostrand, E. L., Nostrand, E. L. V., & Liang, T. Y. (2007). Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS genetics*. *3*(5): e85.

Zhang, C., Hastings, M. L., Krainer, A. R., & Zhang, M. Q. (2007). Dual-specificity splice sites function alternatively as 5’ and 3' splice sites. *PNAS*. 104(38): 15028-15033.

Zhang, X. H., & Chasin, L. A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*. *18*: 1241-1250.

Zhang XH, Leslie CS, Chasin LA. (2005). Computational searches for splicing signals. *Methods*. 37: 292-305.