Title
RNA Splicing Regulation in Drosophila melanogaster

Author
Brooks, Angela Norie

Publication Date
2011

Peer reviewed|Thesis/dissertation

RNA Splicing Regulation in *Drosophila melanogaster*

by

Angela Norie Brooks

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Molecular and Cell Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor Steven E. Brenner, Co-Chair
Professor Donald C. Rio, Co-Chair
Professor Michael B. Eisen
Professor Sandrine Dudoit

Spring 2011

RNA Splicing Regulation in *Drosophila melanogaster*

Angela Norie Brooks

Abstract

RNA Splicing Regulation in *Drosophila melanogaster*

by

Angela Norie Brooks

Doctor of Philosophy in Molecular and Cell Biology

Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Steven E. Brenner, Co-Chair

Professor Donald C. Rio, Co-Chair

A majority of metazoan genes contain introns in their primary transcripts (pre-mRNA) that require removal by the spliceosome—a cellular complex composed of protein and RNA. Upon removal of introns from the primary transcript, the remaining exonic portion of the transcript is spliced together. It is essential to remove the correct intronic portion of a primary transcript in order to produce the desired product, typically a protein-coding mRNA. Pre-mRNAs are alternatively spliced when different intron boundaries are used by the spliceosome, thus creating different mRNA products. Alternative splicing allows for an additional step of gene regulation by producing transcript isoforms that can be differentially processed in a particular tissue or developmental time point. Alternative splicing is primarily regulated by RNA binding proteins that bind to pre-mRNA and act to recruit or inhibit the spliceosome at specific splice sites. A central aim of this work is to gain a better understanding of splicing regulation by the identification and characterization of protein regulators of splicing and *cis*-acting splicing regulatory sequences in the model organism, *Drosophila melanogaster*.

To identify splicing regulatory elements, many previous studies in vertebrate genomes have used computational methods. In collaboration with Anna I. Podgornaia, I applied such an approach to predict splicing regulatory elements in *Drosophila melanogaster* and compared them with elements found in vertebrates. I identified 330 putative splicing enhancer sequences enriched near weak 5' and 3' splice sites of constitutively spliced introns. I found that a significant proportion (58%) of *D. melanogaster* enhancers were previously reported

as splicing enhancers in vertebrates. To provide additional evidence for the function of the intronic splicing enhancers (ISEs), I identified intronic hexamers significantly enriched within sequences phylogenetically conserved among 15 insect species. This analysis uncovered 73 putative ISEs that are also enriched in conserved regions of the *D. melanogaster* genome. The functions of nine enhancer sequences were verified in a heterologous splicing reporter by Julie L. Aspden, demonstrating that these sequences are sufficient to enhance splicing *in vivo*. Taken together, these data identify a set of predicted positive-acting splicing regulatory motifs in the *Drosophila* genome and highlight those regulatory sequences that are present in distant metazoan genomes[1].

To identify and characterize splicing regulators, collaborators and I have combined RNAi and RNA-Seq to identify exons that are regulated by 58 known or putative splicing regulators. To identify and quantify alternative splicing events from RNA-Seq data, I developed the JuncBASE (Junction Based Analysis of Splicing Events) software package. For a pilot study, I identified 404 splicing events significantly affected upon depletion of *pasilla*. Preliminary analysis showed 879 splicing events affected by at least one of the 57 other proteins. The sequence regions upstream and within Pasilla-repressed exons and downstream of Pasilla-activated exons are enriched for YCAY repeats, which is consistent with the location of these motifs near regulated exons of the mammalian ortholog, Nova. Thus, the RNA regulatory map of Pasilla and Nova is highly conserved between insects and mammals despite the fact that the pre-mRNAs that are regulated by Pasilla and Nova are almost entirely non-overlapping. This observation strongly suggests that the regulatory codes of individual RNA binding proteins are nearly immutable, yet the regulatory modules controlled by these proteins are highly evolvable. I also present RNA regulatory maps for the four hnRNP proteins: hrp36, hrp38, hrp40, and hrp48[2].

Lastly, I examine splicing regulation throughout the life cycle of *D. melanogaster*. Using transcriptome data from 30 developmental time points produced by collaborators from the modENCODE Consortium, I identified a total of 23,859 alternative splicing events in *Drosophila*, taking into account all transcript information from *D. melanogaster* annotations, short sequenced reads (Illumina RNA-Seq), sequenced cDNA, long RNA-Seq reads (454 RNA-Seq) from adult flies, and short read sequences of rRNA-depleted RNA from embryonic time points. I observed that 60.7% of intron-containing genes in *D. melanogaster* are alternatively spliced. Using only the Illumina RNA-Seq reads throughout development, 21,216 splicing events were expressed and 13,951 events were differentially spliced in at least one time point. I also observed exons with similar patterns of splicing changes throughout development as well as sex-biased alternative splicing. Integrating information from our *pasilla* study, I also observed correlations of *pasilla* gene expression with alternative splicing changes of its target exons throughout development.

[1]Paragraph was modified from a manuscript co-written by Julie L. Aspden, Anna I. Podgornaia, Donald C. Rio, and Steven E. Brenner.

[2]Paragraph contains excerpts from previously published work, Brooks et al. 2011, and co-written by Brenton R. Graveley, Li Yang, Michael O. Duff, Kasper D. Hansen, Sandrine Dudoit, and Steven E. Brenner.

Contents

List of Figures

List of Tables

Acknowledgments

I have been fortunate to have wonderful scientific mentors who have given support and encouragement throughout my graduate studies. First, I would like to thank my primary mentor and advisor Steven Brenner for teaching me how to think as a scientist, encouraging me to pursue my own scientific interests, and for providing opportunities to work on interesting and exciting research projects. I am grateful for the rest of my dissertation committee. Donald Rio has been a wonderful mentor and collaborator and I cannot thank him enough for his encouragement. Michael Eisen has given me great advice and mentorship since I was an undergrad. I also thank Sandrine Dudoit for her statistical advice and for teaching me exploratory data analysis. Although, Jennifer Doudna was not on my committee, I would like to thank her for her continued mentorship. It has been a pleasure to collaborate with Brenton Graveley at the University of Connecticut, Health Center, who has taken the time to help and advise me as if I were part of his own lab.

All of the material in my dissertation would not be possible without the help and hard work from many amazing collaborators. I would like to thank Kasper Hansen who was a great collaborator and friend, as well as a great teacher of statistics. Julie Aspden has also been a great friend and collaborator and has also taught me so much about biochemistry, RNA biology, and writing. I would like to thank members of the Graveley lab who were crucial in providing their expertise and ideas to our work: Michael Duff, Li Yang, Gemma May, and Jung Park. I would also like to thank the rest of the modENCODE Consortium, specifically Sue Celniker and the rest of the Fly Transcriptome Group, for their work on the fly development study and for input on the RNAi work. My interest in *Drosophila* splicing regulation started with work from Marco Blanchette while he was at Berkeley and I am glad that he continued to be a wonderful colleague and collaborator.

One of the reasons I chose UC Berkeley for graduate school was for the opportunity to work with talented undergraduates. I was fortunate to work with Anna Podgornaia who made a substantial contribution to the project on splicing enhancers. She is extremely bright and creative and I know that she will continue to be a successful scientist. I also had the pleasure of working with Isha Jain and Angadhjot Hundal who also made important contributions to studies on RNA splicing.

I am very appreciative of all the past and present Brenner lab members for their scientific input and for a fun and friendly work environment—particularly Liana Lareau, David Soergel, Max Shatsky, Neel Dey, Susanna Repo, Gaurav Pandey, Orna Mann, Maki Inada, and the rest of the RUST subgroup. A special thanks to Robin Peters, Eric O'Dell, Dean Ekkaia, and Connie Wang for their administrative and technical support.

I am grateful for financial support from the National Institutes of Health, the National Science Foundation, and UC Berkeley's Chancellor's Fellowship. A majority of this dissertation was written in Fertile Grounds Cafe on Shattuck Avenue. I thank them for their excellent espresso, cozy environment, unlimited WiFi, and electrical outlets.

To all of my amazing friends and classmates of 2005, particularly Emily Crane, Jane Klimenko, Andrea Manzo, Shirali Pandya, Nadine Jahchan Rabbat, and Jody Westbrook: all of the study groups, venting sessions, coffee breaks, lunch breaks, dinner breaks, BBQs, and evenings out have been the most memorable parts of graduate school. I was also fortunate to find fun and talented scientists to play music with—thanks to McClintock and Royal Jam band members for the good times and musical therapy. My dearest friend and roommate who witnessed all my ups and downs in graduate school, and life, is Laura Lombardi. I thank her for her support both scientifically and personally. I also thank Laura for her infectious energy and for making our place feel like a home.

I cannot express how greatly I appreciate my parents for their love and encouragement and for always giving me the support to explore my interests. I would also like to thank my loving grandmother for never missing a commencement ceremony, even though she has to travel thousands of miles.

My final acknowledgement goes to my best friend and partner Thomas Schiavone. I thank him for lifting my spirits during the bad times and for being there to celebrate the good ones.

Chapter 1

Introduction

1.1 pre-mRNA splicing of eukaryotic genes[1]

Based on Ensembl annotations in 2010 (Flicek et al. 2010), *Drosophila melanogaster* and *Canaerhabditis elegans*, have 14,869 and 38,120 genes, respectively. The number of annotated genes in the human genome is 51,715, suggesting that the number of genes in an organism alone cannot account for differences in organismal complexity (Hahn and Wray 2002) . Other genomic features can partly account for differences in complexity, including the number of transcripts derived from these genes.

The primary process by which multiple transcripts, or messenger RNAs (mRNA), are produced from a given gene is through alternative splicing. Recent deep sequencing surveys of 10 human tissues found that nearly all (95-98%) multi-exon human genes are alternatively spliced (Pan et al. 2008; Wang et al. 2008), while a survey of *Drosophila* transcripts throughout the entire life cycle of the animal found about 60% of multi-exon genes to be alternatively spliced (Graveley et al. 2011). Although there are many theories for the origin and ancestral role of introns and alternative splicing (Roy 2006), they are present in genes of all eukaryotic genomes examined to date.

The initial product of transcription is a pre-mRNA that may contain introns. The precise excision of introns and the joining of flanking exons is essential in maintaining an accurate mRNA product, typically a coding sequence for protein synthesis. Introns contain several sequence elements required for pre-mRNA splicing: 5' and 3' splice sites (5'ss, 3'ss) that define the start and end of an intron, a branch point, and a polypyrimidine tract (Figure 1.1). Splice sites can be classified as "weak" or "strong" according to their similarity to consensus motifs. The strength of the splice site is correlated with its identification

[1]Portions of this section were co-written by Brenton R. Graveley, Julie L. Aspden, Anna I. Podgornaia, Donald C. Rio, and Steven E. Brenner and modified from (Brooks et al. 2011), and Brooks et al., in preparation.

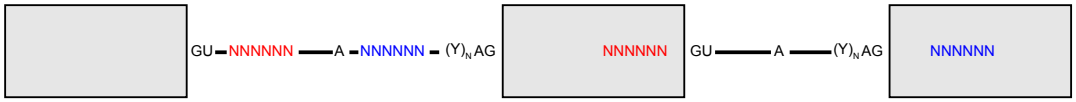


Figure 1.1: **Components of a typical pre-mRNA.** Exons are represented as gray boxes. Introns boundaries are indicated by GU-AG dinucleotides. Branch point, A. Polypyrimidine tract, $(Y)_N$. Splicing regultory sequences are shown as generic hexamers; however, regulatory sequences can range in length and vary in position. Red hexamers, enhancers. Blue hexamers, silencers.

as an intron boundary (Roca et al. 2005; Lim and Burge 2001), exemplified by the fact that constitutively spliced introns have stronger splice sites than alternatively spliced introns (Koren et al. 2007). There are also splicing regulatory elements (SREs) within the pre-mRNA, which influence splicing efficiency (Lim and Burge 2001). SREs are named according to their function and location: exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs), or intronic splicing silencers (ISSs) (Figure 1.1). Enhancer sequences are thought to most often be recognized by serine-rich proteins (SRs) and silencers most often recognized by heterogeneous nuclear ribonucleoproteins (hnRNPs) and are thought to either recruit or inhibit assembly or activity of spliceosomal components at nearby splice sites (Chen and Manley 2009). The specific combination of SREs and their distances from splice junctions contribute to ensuring the proper splicing outcome (Zhang et al. 2009). In addition, current models for splice site selection suggest that splice sites are recognized through interactions of the spliceosome across exons, "exon definition," when exons are flanked by long introns and across introns, "intron definition," when introns are short (Romfo et al. 2000; Lim and Burge 2001; Yeo et al. 2004). Therefore, the relative size of introns is important for splice site selection.

Alternative splicing can generate multiple mRNA products by joining exons together in different combinations. This process is used to both increase protein diversity and to regulate gene expression (Nilsen and Graveley 2010). Alternative splicing is most commonly controlled by RNA binding proteins which bind to enhancers and silencers, such as SR proteins and hnRNPs (Nilsen and Graveley 2010). In addition to SR and hnRNPs proteins, several other splicing regulators have been identified that function in a tissue specific manner (Chen and Manley 2009).

Through examining many instances of alternative splicing, common patterns were observed and were used to classify alternative splicing into different types (Figure 1.2). The most commonly examined form of alternative splicing is a cassette exon where an entire exon can be included or skipped in the resulting mRNA. The most frequent form of alternative splicing observed in *D. melanogaster* are alternative 5' splice sites and alternative 3' splice sites (See Table 5.1 on page 84) where only a portion of an exon is included or skipped.

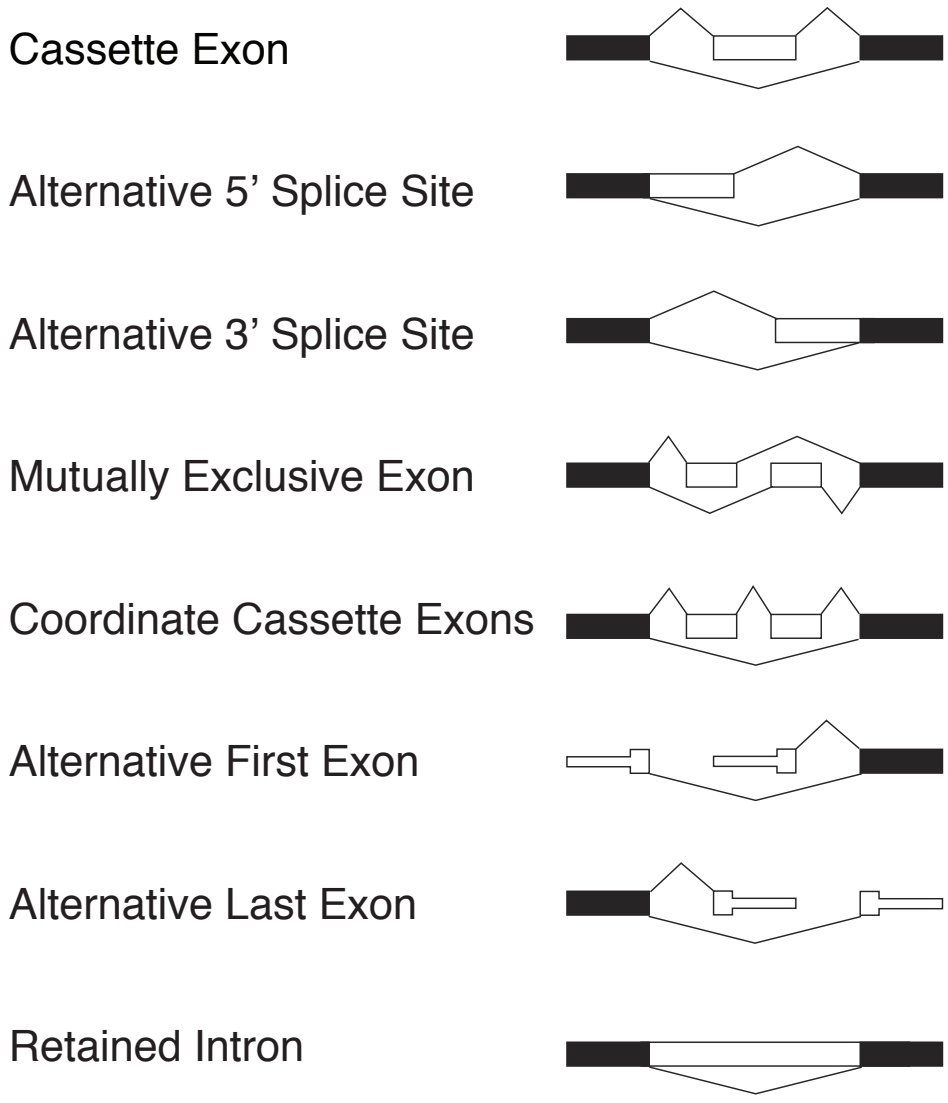


Figure 1.2: **Types of alternative mRNA processing.** Boxes, exons. Lines, introns. White boxes, alternative exons. Black boxes, constitutive exons. Thinner portions of alternative exons indicate UTR portions.

There are more complex forms of alternative splicing such as mutually exclusive exons or coordinate cassette exons. Alternative first exons and alternative last exons should more correctly be referred to as alternative RNA processing events. These events depend on alternative transcriptional promotor usage or alternative transcriptional termination and not solely on changes in splice site usage. Lastly, there are retained introns where the entire intron is either removed or not removed. Retained introns are more prevalent in *D. melanogaster* than humans, which may be due to the short length of most *D. melanogaster* introns. If an intron is short, it is less likely to disrupt the coding sequence if it is retained.

Given the ubiquity of alternative splicing and the key roles it plays in the control of gene function and expression, it is important to develop a complete understanding of the mechanisms by which alternative splicing is regulated.

1.2 Methods for genome-wide detection of alternative splicing

Molecular techniques used to identify the exact intron(s) removed from a pre-mRNA require a first step of selecting polyadenylated (poly(A)+) mRNA—an indicator that the mRNA has been fully transcribed and processed. Selecting poly(A)+ mRNA can start with reverse-transcription polymerase chain reaction (RT-PCR) on the mRNA using a poly(T) primer, converting the mRNA to complementary DNA (cDNA), or by affinity purification using poly(T) oligos. The exon-exon junction ("splice junction") sequence formed upon removal of an intron is most informative in determining what splicing events have occurred in a given sample. Genome-wide methods to detect alternative splicing typically focus on assaying these exon-exon junction sequences.

1.2.1 Splice junction microarrays

One high-throughput method to detect alternative splicing events in a given transcriptome is the use of microarrays. A microarray consists of a glass slide or chip with oligonucleotide probes attached and positioned in an ordered array. Each probe complementary base pairs to a region of interest in the genome. mRNA or cDNA is labeled with a dye, hybridized against the microarray, and "lights up" positions on the array where the sample complements with the probe sequence. The intensity of the signal at each spot in the microarray is proportional to the amount of mRNA that was in the sample; therefore, microarrays are used to not only detect the presence of an mRNA, but also the amount. Due to inherent differences in probe affinities that depend on the sequence, it is difficult to extrapolate exact abundances in the sample or relative abundances between different probes; therefore, relative fold-changes between two or more sets of samples for a given probe is used to determine changes in expression.

To detect alternative splicing the best design of a microarray should include splice junctions. Another microarray design that has been used to detect alternative splicing is an "exon array" where multiple probe sequences are created for each annotated exon. With exon arrays, it is difficult to quantify changes in exon-skipping isoforms, one of the most common forms of alternative splicing, as they do not include probes for the exon-exon junction sequence that is specific to the skipping isoform. When designing probes to detect splice junctions, it is important to have a balanced melting temperature (T_m) on each side of the junction to reduce signal from hybridization of only one side of the junction (Castle et al. 2003). Having more than one sequence spanning the junction can also improve detection.

The number of probes that can be put onto an array depends on the length of the probe and the manufacturer. The splice junction array used for the study presented in chapter and section 4.2 on page 51 contained approximately 44,000 features, including splice junction probes and probes to exons. A limitation of microarrays for detecting alternative splicing is that one needs to know *a priori* what splice junctions to detect; therefore, it is not feasible to detect all possible novel alternative splicing events. Current Agilent array technologies allow up to one million probes and would allow for detection of all our 88,045 observed splice junction sequences[2] in *Drosophila melanogaster*; however, one could not include the over 10 million potential splice junction sequence that could be formed if an unannoated or cryptic splice site is used.

1.2.2 Ultra-high-throughput sequencing of transcriptomes (RNA-Seq)

The newest high-throughput method to detect alternative splicing involves directly sequencing millions of cDNAs (RNA-Seq) in a relatively short amount of time at a relatively reasonable cost. There are many ultra-high-throughput sequencing platforms used for RNA-Seq or other genomic applications; however, given cost and throughput of reads, the most widely adopted platform for RNA-Seq is Illumina (formerly known as Solexa) and will be discussed here. Illumina sequencing data used in the studies described in this dissertation range from 37-75bp and include single and paired-end reads (Tables 1.1, 1.2, and 1.3). To sequence from the population of mRNAs in a biological sample, mRNAs are selected by affinity purification using poly(T) oligos. Next, RNA is fragmented in order to sample sequences evenly across the RNA molecule and then converted into cDNA using random hexamer priming (Mortazavi et al. 2008). After second strand synthesis, sequencing adapters are ligated onto the ends of the cDNA molecules and the library is size selected to approximately 300bp. The sample of cDNA is sequenced starting from the ends of the molecule and the number of base pairs sequenced is determined by the number

[2] "Observed" junction sequences include all annotated junction in FlyBase r5.12 and any additional splice junction detetected from RNA-Seq data presented in this dissertation.

of sequencing cycles. If paired-end sequencing is performed, the sequences at both the forward and reverse ends of the cDNA fragment are determined and the pairing maintained in reporting of the sequences. The paired-end mode of sequencing is useful for identifying and quantifying alternative splicing events because you are more likely to sequence a cDNA fragment that spans two exons than a cDNA fragment whose ends cross a splice junction. An advantage of RNA-Seq is that one can detect splice junctions that were previously unknown to exist. This is not only beneficial in identifying new alternative splicing events, but it also allows us to assay aberrant splicing that may occur when perturbing splicing regulation. In addition to detecting novel alternative splicing events, RNA-Seq has been shown to have a greater dynamic range of expression detection than microarrays and is less susceptible to misquantification due to cross-hybridization issues (Agarwal et al. 2010). RNA-Seq is also beneficial for quantifying gene expression since all exons are surveyed instead of only those that are annotated.

Illumina sequencing technology has advanced at a rapid pace during a very short time, making it challenging for method development or method assessment. The rapid change in sequencing technology can be exemplified from the the data presented here (Tables 1.1, 1.2, 1.3, and 1.4). Our initial RNA-Seq experiments were run in mid-late 2008 and had a combination of single and paired-end 37-45bp reads. The most recent experiments presented sequences up to 80bp reads. Today, it is common to obtain 100bp paired-end sequences and, with the right protocol, infer which strand of the genome the RNA came from. As the data change, the methods to analyze the data must evolve accordingly.

One of the first steps of analyzing RNA-Seq data is to align the reads to a reference genome, keeping in mind that splice junction reads will have a large gap in the alignment relative to the genome. New software was developed to first solve the problem of aligning millions of short 24-50bp sequences to a large reference sequence and perform the alignments in a relatively short amount of time, given a standard computer (Wang et al. 2008; Mortazavi et al. 2008; Trapnell et al. 2009; Langmead et al. 2009). One of these alignment programs, Bowtie (Langmead et al. 2009), uses an indexing method in order to increase the speed of the search. Bowtie allows mismatches in the alignment but not gaps, making it the wrong tool for direct analysis of RNA-Seq data. One solution is to create a set of all possible known or putative splice junctions that could exist in a given sample and align reads against these *in silico* junction sequences. There are many other approaches to spliced-alignments of short reads including directly aligning against the genome or by using exon expression to first determine the set of exons that can then be spliced together (e.g.,Wu and Nacu 2010; Au et al. 2010; Wang et al. 2010; Jean et al. 2010; Dimon et al. 2010; Bryant et al. 2010; Kent 2002; Trapnell et al. 2009). There are current efforts such as RGASP (RNAseq Genome Annotation Assessment Project, http://www.gencodegenes.org/rgasp) to assess these alignment methods, but no published results of these assessments are available. As mentioned previously, with changing sequencing technologies it is most likely that some of the spliced aligner methods will work better depending on the sequence input, for example as reads get longer they are more likely to span more than one splice junction and the

aligner must take this into consideration.

Once reads have been aligned, the next step is assigning these reads to specific transcript isoforms and then quantifying their expression. This step in the analysis pipeline also has a wide range of approaches. To examine alternative splicing there are two major approaches, an exon based approach where the presence an abundance of a particular splicing event or exon is examined (e.g. cassette exon events, alternative 5' splice site events, etc.) (Katz et al. 2010; Brooks et al. 2011) or entire transcript isoforms are predicted and quantified (Martin et al. 2010; Trapnell et al. 2010; Guttman et al. 2010; Jiang and Wong 2009).

RNA-Seq has greatly changed the field of splicing as we are able to examine the extent of alternative splicing at an unprecedented level. Also, there is now a wealth of data about alternative splicing from any researcher who performs an RNA-Seq experiment, even if they only care about gene expression. With microarrays, most experiments for gene expression were done on platforms designed specifically for gene expression and therefore are very difficult to use to examine splicing.

1.3 Genome-wide methods to identify splicing regulatory elements[3]

To understand how splicing is regulated on a specific pre-mRNA, it is important to know the full set of splicing regulatory elements and understand their function. Putative SREs have previously been identified with *in vitro* SELEX experiments where binding sites of known splicing regulators are determined (Shi et al. 1997; Liu et al. 1998; Amarasinghe et al. 2001; Wang et al. 2004b; Smith et al. 2006; Blanchette et al. 2009). Additional experimental determination of SREs involve *in vivo* functional selection for splicing effects of random sequences inserted into mini-gene reporter libraries (Wang et al. 2004b).

Another method to identify splicing regulatory elements identifies enriched motifs near co-regulated exons that are differentially spliced in particular tissues (e.g. Castle et al. 2008; Sugnet et al. 2006; Kalsotra et al. 2008; Das et al. 2007). Enriched motifs give clues to which proteins might be regulating the differential splicing.

Purely computational approaches have also been successful at identifying SREs. Since RNA-binding proteins typically bind to 6-8 nt, computational searches focus on finding enriched RNA elements of this size in functionally relevant locations (Fedorov et al. 2001). One such approach is the Relative Enhancer and Silencer Classification by Unanimous Enrichment (RESCUE) method (Fairbrother et al. 2002), which has been applied to numerous genomes including mammals, fish, and plants (Fairbrother et al. 2002; Yeo et al. 2004; Zhang and Chasin 2004; Pertea et al. 2007). The RESCUE method detects motifs

[3]Portions of this section were co-written by Julie L. Aspden, Anna I. Podgornaia, Donald C. Rio, and Steven E. Brenner and modified from Brooks et al., in preparation.

Biological Sample Description	Single/Paired/Mixed	Trimmed Read Length	Number of single reads	Number of paired reads	Uniquely aligned single reads	Uniquely aligned mate pairs	Number of sequenced base pairs	Number of aligned base pairs
Untreated	Mixed	37	32,989,325	22,805,923	28,232,489	14,858,720	2,908,243,327	2,144,147,373
CG8144-RNAi	Mixed	37	37,271,305	24,730,628	27,124,008	15,356,992	3,209,104,757	2,140,005,704

Table 1.1: Details of RNA-Seq data from Pasilla study, chapter and section 4.3 on page 53. Sequences associated with this study have the following Gene Expression Omnibus (GEO) accession numbers: GSM461176-GSM461181.

Biological Sample Description	Single/Paired/Mixed	Trimmed Read Length	Number of single reads	Uniquely aligned single reads	Number of sequenced base pairs	Number of aligned base pairs
CG10128-RNAi	Single	75	56,692,047	40,122,191	4,251,903,525	3,009,164,325
CG10203-RNAi	Single	75	36,443,346	27,059,564	2,733,250,950	2,029,467,300
CG10279-RNAi	Single	75	44,015,949	35,649,044	3,301,196,175	2,673,678,300
CG10328-RNAi	Single	75	28,021,729	22,752,024	2,101,629,675	1,706,401,800
CG10377-RNAi	Single	75	22,431,612	18,352,508	1,682,370,900	1,376,438,100
CG10851-RNAi	Single	75	59,854,018	40,726,336	4,489,051,350	3,054,475,200
CG1101-RNAi	Single	75	33,381,089	26,267,072	2,503,581,675	1,970,030,400
CG11266-RNAi	Single	75	31,517,659	25,766,498	2,363,824,425	1,932,487,350
CG12749-RNAi	Single	75	28,254,964	21,184,672	2,119,122,300	1,588,850,400
CG13425-RNAi	Single	75	69,802,318	48,385,744	5,235,173,850	3,628,930,800
CG1559-RNAi	Single	75	44,029,413	27,604,682	3,302,205,975	2,070,351,150
CG1646-RNAi	Single	75	47,282,143	39,850,917	3,546,160,725	2,988,818,775
CG16788-RNAi	Single	75	27,473,098	22,371,356	2,060,482,350	1,677,851,700
CG16901-RNAi	Single	75	31,114,651	22,104,329	2,333,598,825	1,657,824,675
CG17136-RNAi	Single	75	29,501,612	21,199,162	2,212,620,900	1,589,937,150
CG17838-RNAi	Single	75	31,194,944	22,507,820	2,339,620,800	1,688,086,500
CG18350-RNAi	Single	75	48,680,046	42,139,744	3,651,003,450	3,160,480,800
CG18426-RNAi	Single	75	41,528,441	29,467,916	3,114,633,075	2,210,093,700
CG1987-RNAi	Single	75	32,738,873	23,424,420	2,455,415,475	1,756,831,500
CG30122-RNAi	Single	75	53,195,919	33,269,298	3,989,693,925	2,495,197,350
CG31000-RNAi	Single	75	32,844,299	21,233,917	2,463,322,425	1,592,543,775
CG31716-RNAi	Single	75	37,831,927	31,577,806	2,837,394,525	2,368,335,450
CG32423-RNAi	Single	75	40,681,395	31,632,015	3,051,104,625	2,372,401,125
CG3249-RNAi	Single	75	72,808,064	46,118,222	5,460,604,800	3,458,866,650
CG33106-RNAi	Single	75	37,516,841	30,318,389	2,813,763,075	2,273,879,175
CG3312-RNAi	Single	75	41,394,281	35,175,798	3,104,571,075	2,638,184,850
CG33584-RNAi	Single	75	33,629,975	27,874,478	2,522,248,125	2,090,585,850
CG4262-RNAi	Single	75	47,564,827	41,447,570	3,567,362,025	3,108,567,750
CG4602-RNAi	Single	75	52,974,722	34,641,388	3,973,104,150	2,598,104,100
CG4760-RNAi	Single	75	43,995,680	35,380,355	3,299,676,000	2,653,526,625
CG4816-RNAi	Single	75	25,790,094	20,176,326	1,934,257,050	1,513,224,450
CG4878-RNAi	Single	75	33,700,037	28,001,239	2,527,502,775	2,100,092,925
CG5099-RNAi	Single	75	29,381,753	20,776,134	2,203,631,475	1,558,210,050
CG5170-RNAi	Single	75	32,743,356	27,102,731	2,455,751,700	2,032,704,825
CG5422-RNAi	Single	75	64,687,843	42,685,633	4,851,588,225	3,201,422,475
CG5442-RNAi	Single	75	59,651,643	40,686,826	4,473,873,225	3,051,511,950
CG5655-RNAi	Single	75	99,158,225	57,170,053	7,436,866,875	4,287,753,975
CG5821-RNAi	Single	75	44,123,002	29,973,777	3,309,225,150	2,248,033,275
CG5836-RNAi	Single	75	41,000,542	23,983,907	3,075,040,650	1,798,793,025
CG6049-RNAi	Single	75	31,389,362	24,836,500	2,354,202,150	1,862,737,500
CG6203-RNAi	Single	75	37,051,699	30,216,799	2,778,877,425	2,266,259,925
CG6227-RNAi	Single	75	27,028,661	22,443,206	2,027,149,575	1,683,240,450
CG6779-RNAi	Single	75	44,995,261	37,568,498	3,374,644,575	2,817,637,350
CG6841-RNAi	Single	75	40,981,591	33,743,278	3,073,619,325	2,530,745,850
CG6946-RNAi	Single	75	39,307,466	24,044,390	2,948,059,950	1,803,329,250
CG7437-RNAi	Single	75	38,598,955	25,561,241	2,894,921,625	1,917,093,075
CG7878-RNAi	Single	75	70,319,555	52,621,232	5,273,966,625	3,946,592,400
CG7971-RNAi	Single	75	29,711,143	20,510,698	2,228,335,725	1,538,302,350
CG8019-RNAi	Single	75	39,828,198	30,554,615	2,987,114,850	2,291,596,125
CG8241-RNAi	Single	75	32,993,939	25,713,056	2,474,545,425	1,928,479,200
CG8636-RNAi	Single	75	46,724,113	37,325,780	3,504,308,475	2,799,433,500
CG8749-RNAi	Single	75	33,638,190	27,180,736	2,522,864,250	2,038,555,200
CG8781-RNAi	Single	75	44,015,571	34,856,806	3,301,167,825	2,614,260,450
CG8912-RNAi	Single	75	38,415,708	30,996,531	2,881,178,100	2,324,739,825
CG9373-RNAi	Single	75	72,997,844	58,735,925	5,474,838,300	4,405,194,375
CG9412-RNAi	Single	75	43,784,294	35,309,209	3,283,822,050	2,648,190,675
CG9983-RNAi	Single	75	30,426,332	21,736,670	2,281,974,900	1,630,250,250
Untreated	Single	75	37,165,906	29,553,014	2,787,442,950	2,216,476,050

Table 1.2: Details of RNA-Seq data from 57 protein study, chapter and section 4.4 on page 65. Sequences associated with this study have the following Gene Expression Omnibus (GEO) accession number: GSE18508.

Biological Sample Description	Single/Paired/Mixed	Trimmed Read Length	Number of single reads	Number of paired reads	Uniquely aligned single reads	Uniquely aligned mate pairs	Number of sequenced base pairs	Number of aligned base pairs
embryos, 0-2 hr after egg laying	Single	75	12,184,337		9,779,436		913,825,275	733,457,700
embryos, 0-2 hr after egg laying	Mixed	76	15,119,607	68,027,196	29,515,881	12,552,779	11,489,223,924	4,151,229,364
embryos, 2-4 hr after egg laying	Single	75	15,316,730		11,296,779		1,148,754,750	847,258,425
embryos, 2-4 hr after egg laying	Mixed	76	18,585,798	45,701,954	18,780,953	16,621,672	8,359,217,656	3,953,846,572
embryos, 4-6 hr after egg laying	Single	75	14,747,695		11,532,426		1,106,077,125	864,931,950
embryos, 4-6 hr after egg laying	Mixed	76	51,517,516	127,144,564	68,818,824	16,037,908	23,241,304,944	7,667,992,640
embryos, 6-8 hr after egg laying	Single	75	12,044,718		9,905,011		903,353,850	742,875,825
embryos, 6-8 hr after egg laying	Mixed	76	35,070,528	71,558,306	42,812,229	10,031,616	13,542,222,640	4,778,535,036
embryos, 8-10 hr after egg laying	Single	75	11,805,956		8,796,547		885,446,700	659,741,025
embryos, 8-10 hr after egg laying	Mixed	76	40,737,291	55,027,572	31,802,022	11,473,499	11,460,225,060	4,160,925,520
embryos, 10-12 hr after egg laying	Single	75	11,727,233		9,080,347		879,542,475	681,026,025
embryos, 10-12 hr after egg laying	Mixed	76	46,883,299	80,257,600	46,375,122	12,771,059	15,762,285,924	5,465,710,240
embryos, 12-14 hr after egg laying	Single	75	14,262,678		11,162,001		1,069,700,850	837,150,075
embryos, 12-14 hr after egg laying	Mixed	76	58,012,865	107,835,850	61,173,626	12,590,478	20,800,026,940	6,562,948,232
embryos, 14-16 hr after egg laying	Single	75	12,758,841		10,133,798		956,913,075	760,034,850
embryos, 14-16 hr after egg laying	Mixed	76	41,118,958	83,377,862	44,243,743	8,934,220	15,798,475,832	4,720,525,908
embryos, 16-18 hr after egg laying	Single	75	12,569,683		9,797,018		942,726,225	734,776,350
embryos, 16-18 hr after egg laying	Mixed	76	48,504,770	56,159,990	45,876,787	11,694,407	12,222,681,000	5,264,185,676
embryos, 18-20 hr after egg laying	Single	75	14,097,356		9,735,065		1,057,301,700	730,129,875
embryos, 18-20 hr after egg laying	Mixed	76	57,550,871	54,938,270	53,845,668	10,137,664	12,724,483,236	5,633,195,696
embryos, 20-22 hr after egg laying	Single	75	15,259,290		10,262,697		1,144,446,750	769,702,275
embryos, 20-22 hr after egg laying	Mixed	76	19,131,922	39,376,286	74,550,875	14,255,491	7,439,221,544	7,832,701,132
embryos, 22-24 hr after egg laying	Single	75	15,498,353		11,018,634		1,162,376,475	826,397,550
embryos, 22-24 hr after egg laying	Mixed	76	43,331,733	96,654,222	47,547,096	10,440,089	17,984,653,452	5,200,472,824
L1 stage larvae	Single	75	13,227,708		10,535,903		992,078,100	790,192,725
L1 stage larvae	Mixed	76	52,053,872	60,378,448	45,699,850	13,783,666	13,133,618,368	5,568,305,832
L2 stage larvae	Single	75	13,502,577		9,675,417		1,012,693,275	725,656,275
L2 stage larvae	Mixed	76	52,872,463	122,435,590	62,899,392	15,433,635	22,628,516,868	7,126,266,312
L3 stage larvae, 12 hr post-molt	Single	75	7,317,378		5,645,100		548,803,350	423,382,500
L3 stage larvae, 12 hr post-molt	Mixed	76	18,301,843	39,872,200	17,196,843	14,021,099	7,451,514,468	3,438,167,116
L3 stage larvae, dark blue gut PS(1-2)	Single	75	15,961,255		10,813,552		1,197,094,125	811,016,400
L3 stage larvae, dark blue gut PS(1-2)	Mixed	76	20,715,955	42,783,456	19,709,060	10,356,308	8,077,497,892	3,072,047,376
L3 stage larvae, light blue gut PS(3-6)	Single	75	13,770,853		10,022,394		1,032,813,975	751,679,550
L3 stage larvae, light blue gut PS(3-6)	Mixed	76	29,166,259	29,519,130	24,819,924	10,405,462	6,703,543,444	3,467,944,448
L3 stage larvae, clear gut PS (7-9)	Mixed	76	67,593,286	35,813,048	48,251,671	9,275,325	10,580,673,032	5,076,976,396
white prepupae (WPP)	Single	75	14,944,452		11,949,054		1,120,833,900	896,179,050
white prepupae (WPP)	Mixed	76	66,392,025	32,417,412	50,626,945	10,120,781	9,973,240,524	5,386,006,532
WPP + 12 hr	Single	75	14,680,143		10,848,882		1,101,010,725	813,666,150
WPP + 12 hr	Mixed	76	44,416,994	70,401,648	37,209,117	14,133,508	14,076,742,040	4,976,186,108
WPP + 24 hr	Single	75	15,588,264		11,743,395		1,169,119,800	880,754,625
WPP + 24 hr	Mixed	76	38,605,004	51,093,524	34,230,244	11,069,011	10,700,195,952	4,283,988,216
pupae, WPP + 2 days	Single	75	14,558,555		11,420,257		1,091,891,625	856,319,275
pupae, WPP + 2 days	Mixed	76	62,430,206	40,409,592	53,261,218	10,278,259	10,886,953,640	5,610,147,936
pupae, WPP + 3 days	Mixed	76	85,632,302	68,644,172	73,161,987	7,890,329	16,941,969,096	6,759,641,020
pupae, WPP + 4 days	Single	75	14,579,573		11,731,036		1,093,467,975	879,827,700
pupae, WPP + 4 days	Mixed	76	48,909,581	42,898,328	41,401,923	11,993,655	10,237,674,012	4,969,581,708
adult female, eclosion + 1 day	Single	75	12,983,189		8,303,409		973,739,175	622,755,675
adult female, eclosion + 1 day	Mixed	76	67,159,001	41,553,846	56,257,333	11,839,068	11,420,268,668	6,075,095,644
adult male, eclosion + 1 day	Single	75	14,022,135		10,552,591		1,051,660,125	2,567,304,525
adult male, eclosion + 1 day	Mixed	76	48,738,944	48,874,294	42,940,370	11,922,169	11,133,052,432	5,075,637,808
adult female, eclosion + 5 days	Single	75	16,124,384		11,680,262		1,209,328,800	876,019,650
adult female, eclosion + 5 days	Mixed	76	21,815,527	52,803,556	25,565,983	17,178,858	9,684,120,564	4,554,201,124
adult male, eclosion + 5 days	Single	75	13,149,510		9,849,814		986,213,250	738,736,050
adult male, eclosion + 5 days	Mixed	76	20,888,331	88,870,060	23,343,444	31,021,484	15,095,762,276	6,489,367,312
adult female, eclosion + 30 days	Single	75	17,060,006		12,066,834		1,279,500,450	905,012,550
adult female, eclosion + 30 days	Mixed	76	13,165,747	59,185,604	15,514,166	19,430,399	9,996,808,580	4,132,497,264
adult male, eclosion + 30 days	Single	75	11,339,769		6,878,090		850,482,675	515,886,750
adult male, eclosion + 30 days	Mixed	76	64,635,262	28,042,810	50,915,097	4,606,588	9,174,787,032	4,569,748,748

Table 1.3: Details of RNA-Seq data from *Drosophila* development study, chapter 5 on page 82. Sequences associated with this study have the following Short Read Archive (SRA) accession number: SRP001065.

Number of single reads	Number of paired reads	Uniquely aligned single reads	Uniquely aligned mate pairs	Number of sequenced base pairs	Number of aligned base pairs
4,202,407,176	1,889,592,941	3,449,589,684	412,516,198	597,319,968,074	320,025,523,567

Table 1.4: Total amount of analyzed RNA-Seq data

enriched near weak splice sites of constitutively spliced introns with the assumption that neighboring sequences act as enhancers to compensate for poor splice site recognition. Other approaches use the premise that functional splicing regulatory elements are under stringent evolutionary constraint and have looked specifically at conserved sequences enriched in exons or introns (Goren et al. 2006; Yeo et al. 2007; Voelker and Berglund 2007; Churbanov et al. 2009; Kabat et al. 2006).

1.4 Reverse genetics approach to identify target splicing events of protein regulators

To understand splicing regulation, it is also important to know the set of *trans*-acting protein regulators and their specific target exons. Multiple studies have taken a reverse genetics approach to examine protein regulators by creating mutants with non-existing or non-functional versions of the proteins or by performing RNA interference (RNAi), depleting the expression of the genes coding for these regulators (e.g., Blanchette et al. 2005, 2009; Taliaferro et al. 2011; Ule et al. 2005; Llorian et al. 2010; Kress et al. 2008; Pleiss et al. 2007; Barberan-Soler et al. 2011; Rösel et al. 2011; Kawashima et al. 2009). A study in *D. melanogaster* identified splicing regulators by depleting putative genes using RNAi and then assayed splicing changes of five events (Park et al. 2004). Although targets of many regulators have been identified, the full set of splicing regulators and their targets are still unknown. It is also important to determine these targets in multiple species to examine the evolution of splicing regulation.

1.5 RNA-maps and the splicing code [4]

An interesting pattern has emerged from studying targets of the mammalian splicing regulators Nova1 and Nova2 (collectively named here as Nova). Nova1/2 encode RNA binding proteins with three KH-domains that recognize clusters of YCAY repeats. Over the past decade, several hundred splicing events have been shown to be regulated by Nova1/2 (Ule et al. 2005, 2006; Licatalosi et al. 2008). A comparison of the locations of the Nova1/2 binding sites with Nova-regulated splicing events has revealed a stereotypical "RNA map" for Nova1/2. Specifically, regions upstream of exons where Nova inhibits splicing and regions downstream of exons where Nova activates splicing were enriched with Nova binding sites (Ule et al. 2006; Licatalosi et al. 2008). Similar "RNA maps" that link the position of binding sites to typical activities of the regulatory proteins have also been developed for mammalian Fox1/2 (Zhang et al. 2008; Yeo et al. 2009), PTB (Xue et al.

[4]Portions of this section were co-written by Brenton R. Graveley and Steven E. Brenner and modified from (Brooks et al. 2011)

2009; Llorian et al. 2010), hnRNP C (König et al. 2010), ESRPs (Warzecha et al. 2010) and *Drosophila* LS2 (Taliaferro et al. 2011). Such maps, splicing expression data, and RNA sequence motifs have recently been used to predict regulated tissue-specific splicing changes in mouse, strongly supporting the existence of a splicing code (Wang and Burge 2008; Barash et al. 2010; Zhang et al. 2010) – a decipherable sequence-based information system that dictates the splicing pattern of a given pre-mRNA under a specific condition. Though considerable progress has been made, interpreting this code remains a formidable task in the field. In particular, it is unclear how the mouse splicing code can be applied to different species, especially distantly related organisms such as *Drosophila*. Moreover, the extent to which the RNA maps of individual splicing regulators are static or plastic throughout evolution has been unknown.

1.6 Summary of contents

The aim of this work is to gain a better understanding of splicing regulation in the model organism *Drosophila melanogaster* by expanding on the set of known alternatively spliced genes as well as identifying and characterizing the *cis*- and *trans*-elements involved in their regulation. *Drosophila* can be used as a model system to study splicing as all major components of the spliceosome are conserved and also many auxillary splicing regulators (Barbosa-Morais et al. 2006; Schwartz et al. 2008). More importantly, by examining and comparing splicing regulation in an animal at a long evolutionary distance to mammals, we can better understand splicing regulation in the common ancestor to all metazoans.

- In chapter 2, I present a set of SREs, including ESEs and ISEs as well as additional putative intronic SREs. These regulatory sequences were identified in *Drosophila* by implementing the RESCUE method, which has successfully identified regulatory elements in vertebrates. In addition, I used sequence information from 14 other insect species to identify those ISEs that are phylogenetically conserved and to identify additional intronic splicing regulatory sequences. Julie L. Aspden performed experimental validation on 10 SREs to show their function *in vivo*. Anna I. Podgornaia also contributed to the computational analysis.

- During the time of my dissertation, RNA-Seq became a feasible experiment to perform in order to identify splicing events. In chapter 3, I present our group's approach to use RNA-Seq data to identify splice junctions, identify all possible alternative splicing events based on those splice junctions, and then identify changes in alternative splicing between two samples.

- In chapter 4, I present work on the target splicing events of 58 known or putative splicing regulators. The first section describes RNA maps of target exons for hrp36, hrp38, hrp40, and hrp48 that were identified through splice junction microarrays by

Marco Blanchette. The second section includes work on the target splicing events of *pasilla*, the ortholog of mammalian Nova-1/2. The *pasilla* work was our group's pilot study on the use of RNA-Seq to identify target splicing events upon knockdown of a splicing regulator and from which most of the computational methods were developed. From the identification of targets of *pasilla*, I was able to show the conservation of an RNA map between *Drosophila* and mammals. Li Yang and Jung W. Park performed the experimental work in the *pasilla* study. Michael O. Duff, Kasper D. Hansen, and Brenton R. Graveley performed portions of the computational analysis. The last section gives a set of splicing events affected upon knockdown of 57 proteins. Here, I look at the differential and shared effects of each regulator on splicing events. The RNAi depletion of the 57 proteins and sequencing libraries were done by Li Yang and Gemma E. May. Sequence alignments of the sequenced RNA fragments from the RNAi-depleted cells was done by Brenton R. Graveley and Michael O. Duff.

- Chapter 5 includes an analysis of RNA-Seq data from 30 developmental time points of the *Drosophila* life cycle. This work was done in collaboration with the modENCODE consortium. From this analysis, I was able to identify 23,859 alternative splicing events, 13,951 of which are significantly changing during development. This study was able to survey the amount of alternative splicing in *Drosophila* at an unprecedented level. From the RNAi study of *pasilla*, described in chapter 4, I was able to identify exons regulated by the protein and, here, I show that the alternative splicing of those exons throughout development correlates with the expression of *pasilla*, further supporting that they are directly regulated by *pasilla*.

The work described in this dissertation has contributed to a collaboration with the modENCODE (model organism Encyclopedia of DNA Elements) Consortium (Celniker et al. 2009).

Chapter 2

Computational prediction of splicing regulatory elements in *D. melanogaster*

2.1 Introduction

Here I employed a combination of two methods, the RESCUE approach and a statistical model to define genomic regions under evolutionary sequence constraint, to predict SREs in *D. melanogaster*.[1] To define constrained sequences, I used 15 insect species, over large evolutionary distances, to identify phylogenetically conserved intronic elements. By comparing the set of fly splicing regulatory sequences with those found in vertebrates, I have identified sequence elements that have been functionally conserved across distant animal species. Interestingly, 58% of the enhancer elements identified here in *Drosophila* have also been identified in vertebrates. Several of the motifs are predicted binding sites of both *Drosophila* and mammalian RNA-binding proteins. Compared to vertebrate genomes with characterized splicing regulatory elements, the *D. melanogaster* genome has the unique feature of a large proportion of short introns. I have taken advantage of this feature to ask if there are different regulatory sequences present near short and long introns, indicating a specific involvement in exon or intron definition. A selection of putative SREs was tested for functionality *in vivo* in a mini-gene reporter. The majority of sequences examined had significant effects on the level of splicing, indicating the robustness of the computational approach employed in this study.

[1] This work was performed and co-written in collaboration with Julie L. Aspden, Anna I. Podgornaia, Donald C. Rio, and Steven E. Brenner for a manuscript that is in preparation.

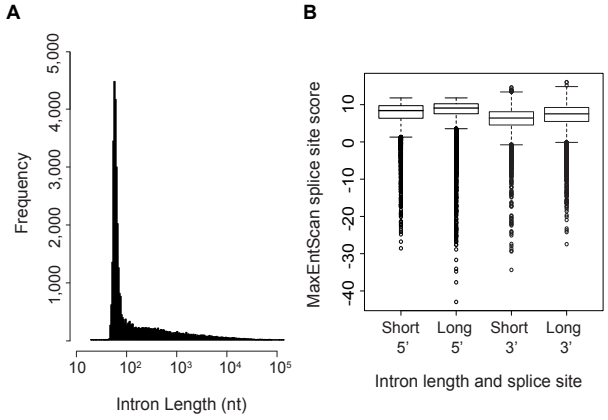


Figure 2.1: **Splice sites of short constitutively spliced introns are weaker than long constitutively spliced introns in *Drosophila*.** (A) Length distribution of constitutively spliced introns in *Drosophila*. (B) A significant difference in the distribution of MaxEntScan splice site scores (Yeo and Burge 2004) of short and long introns (p-val < 2.26e-16, Wilcoxon rank sum test). Higher MaxEntScan scores correspond to a stronger splice site sequence.

2.2 Results

2.2.1 Long and short introns have different distributions of splice site strengths

The number and type of regulatory elements near an intron is dependent upon intron length and splice site strength (Xiao et al. 2007; Yeo et al. 2004; Lim and Burge 2001); therefore, I looked for potential biases in SREs arising from intron length. The length distribution of constitutively spliced introns in *D. melanogaster* consists of a peak with a mode at 69 nt and a long tail (Figure 2.1). This length distribution is different from that of human introns, with a mode of 1500 nt (Fedorova and Fedorov 2005; Yeo et al. 2004; Lim and Burge 2001). Given the intron length distribution, I divided constitutively spliced introns into two categories: short ($\leq$ 80 nt; 22,329 introns) and long (> 80 nt; 15,474 introns). Using MaxEntScan (Yeo and Burge 2004) to score splice site strengths, I found that longer introns have significantly stronger 5' and 3' splice site strengths than shorter introns (p-val < 2.26e-16 for 5' and 3' splice sites, Wilcoxon rank sum test) (Figure 2.1). MaxEntScan runs were done by Anna I. Podgornaia. Although MaxEntScan's scores are derived from human splice sites, *Drosophila* splice site motifs are highly similar to human and many spliceosomal components involved in splice site recognition are highly

	Short introns, 5' splice site	Short introns, 3' splice site	Long introns, 5' splice site	Long introns, 3' splice site
1st Quartile	6.35	4.54	7.56	5.49
Median	8.40	6.41	9.09	7.54
3rd Quartile	9.73	8.10	10.24	9.24

Table 2.1: **Quartiles of splice site strengths, calculated using MaxEntScan, in constitutively spliced introns (FlyBase r5.4)**

conserved (Barbosa-Morais et al. 2006; Schwartz et al. 2008). Dividing the data into two length groups accounts for effects of intron length and permits the identification of motifs that may be specific to each length class.

2.2.2 Identification of ESEs and ISEs in *D. melanogaster*

To identify ESEs and ISEs in *Drosophila*, I implemented the RESCUE method (Fairbrother et al. 2002). Code used to implement the RESCUE method was contributed by Anna I. Podgornaia. Employing this method allows direct comparisons between the *Drosophila* motifs and those that were previously identified in other species by RESCUE. I applied the RESCUE method to constitutive splicing events to identify potential ESEs, characterized as hexamers significantly enriched in exons compared to introns and significantly enriched near either a weak 5' or weak 3' splice site, compared to strong splice sites (see Methods, Figure 2.2). ESEs were identified near short and long introns, separately. Putative ESE sequences were identified up to 100bp upstream of 5' splice sites and up to 100bp downstream of 3' splice sites, excluding the splice site sequences. If an exon was shorter than 100bp, sequence from the entire exon was used and was not extended into the next intron. I used the distribution of splice site strengths in both data sets to define cutoffs for weak and strong splice sites (Figure 2.1, Table 2.1): weak splice sites were defined as the 1st quartile of scores and strong splice sites in the 4th quartile.

This analysis identified 22 hexamers near 5'ss and 34 hexamers near 3'ss of short introns as putative ESEs (Figure 2.3). Five sequences were enriched near both splice sites (CTGGAG, CTGGAT, CTGGAA, CCTGGA, GGAAAC). Although enhancers are predicted to act a the RNA level, I include thymines when reporting their hexamers since they were discovered using the genomic DNA sequence. Near long introns, 19 hexamers were found to be enriched in exons at the 5'ss and 33 hexamers at the 3'ss (Figure 2.3). Two sequences were found in exons near long introns at both splice sites (AGAGGA, AATGGA). Interestingly, just two hexamers (AAGGAA, AGAGGA) were shared between the ESEs identified near short introns and near long introns for either splice site. This suggests that largely distinct regulatory sequences are present in exons proximal to introns of different sizes. Closely related hexamers were clustered according to their edit distances (Böckenhauer and Bongartz 2007; Fairbrother et al. 2002) and sequence logos were created to identify

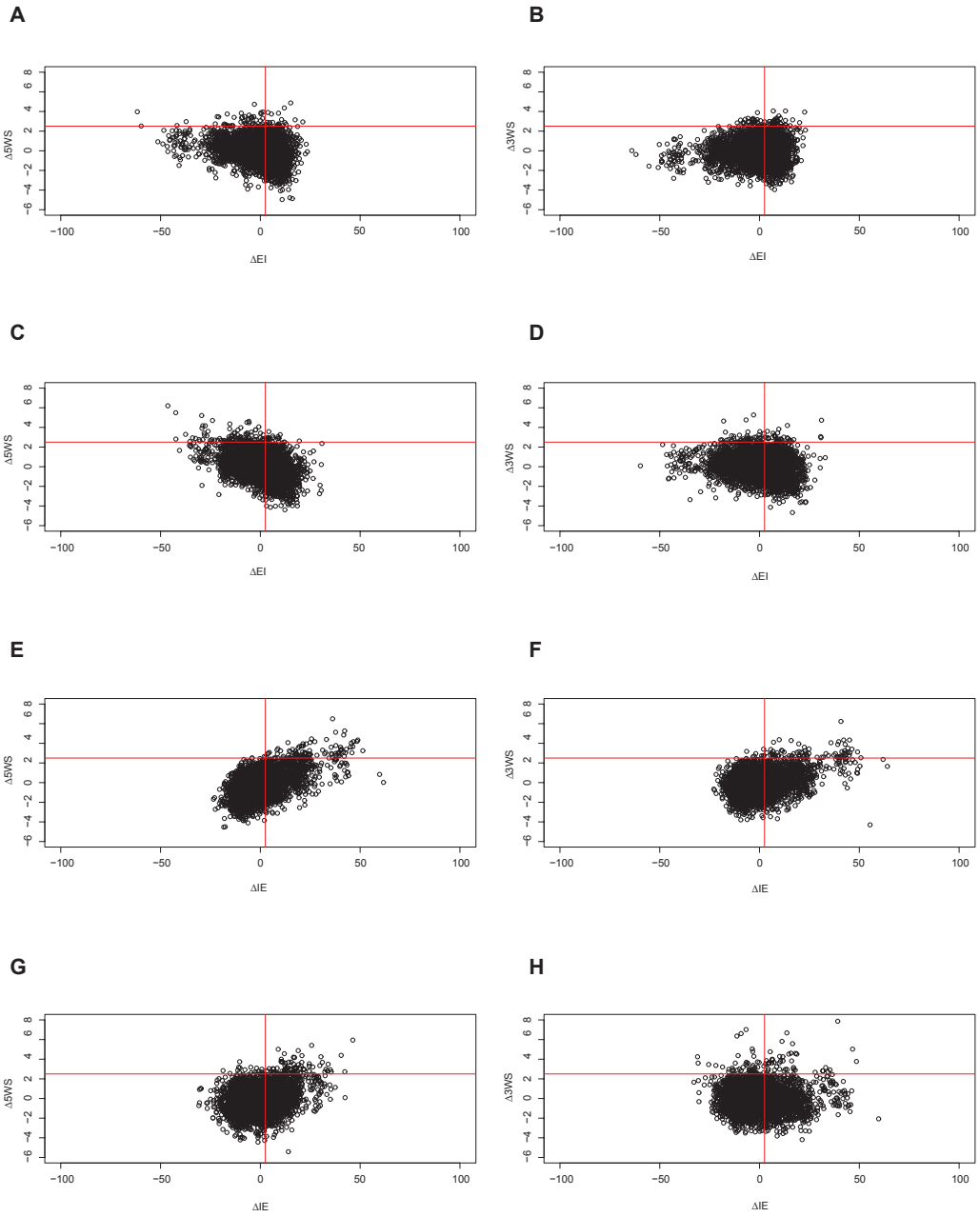


Figure 2.2: **Scatterplots of hexamer scores.** (A) ΔEI vs Δ5WS near short introns, (B) ΔEI vs Δ3WS near short introns, (C) ΔEI vs Δ5WS near long introns, (D) ΔEI vs Δ3WS near long introns, (E) ΔIE vs Δ5WS in short introns, (F) ΔIE vs Δ3WS in short introns, (G) ΔIE vs Δ5WS in long introns, and (H) ΔIE vs Δ3WS in long introns.

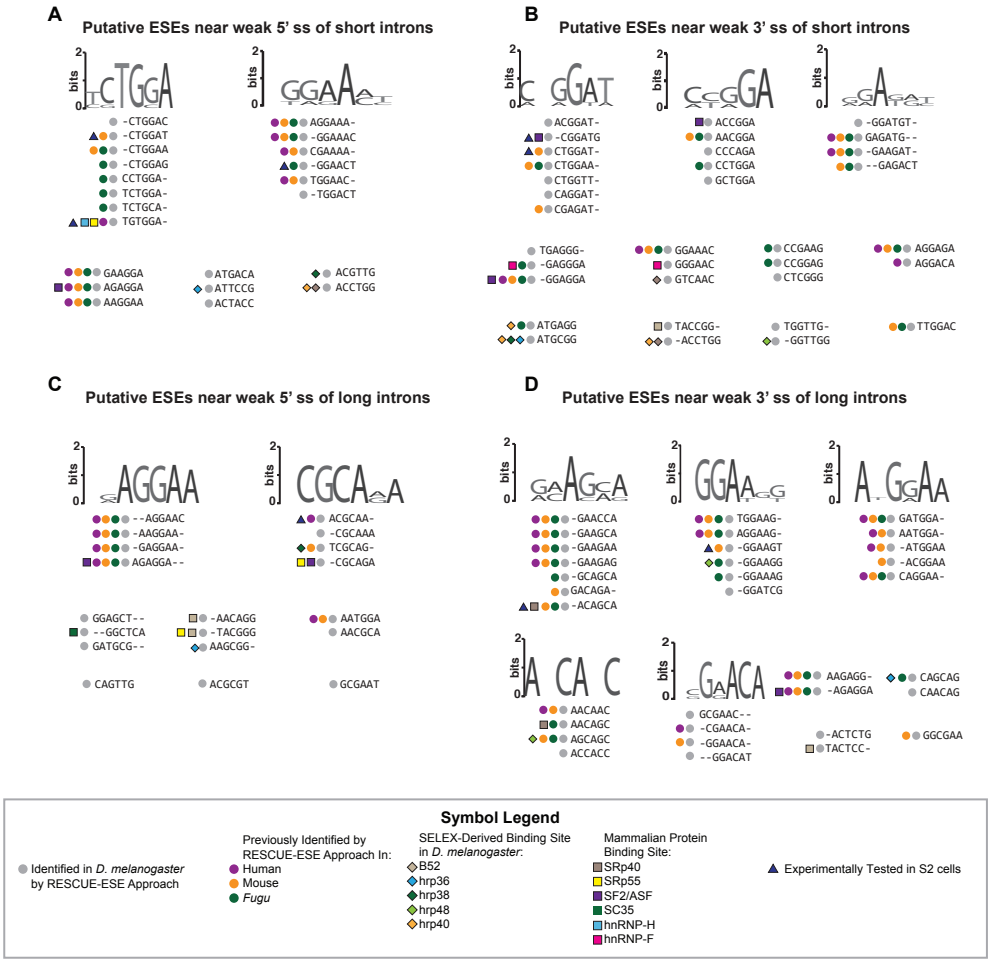


Figure 2.3: **Hexamers and motifs enriched in exons and introns near weak splice sites of constitutive introns.** Putative exonic splicing enhancers (ESEs) enriched near short introns of (A) weak 5' splice sites and (B) weak 3' splice sites, and long introns near (C) weak 5' splice sites and near (D) weak 3' splice sites. Hexamers identifed using the RESCUE-ESE method in this study and in human, mouse, or *Fugu* (Yeo et al. 2004) are indicated by colored circles. Hexamers containing high affinity binding sites of *D. melanogaster* splicing regulators (Shi et al. 1997; Amarasinghe et al. 2001; Blanchette et al. 2009) are indicated by colored diamonds, while binding sites for mammalian proteins are indicated by colored squares (Cartegni et al. 2002; Smith et al. 2006; Goren et al. 2006). The seven hexamers tested by Julie L. Aspden for enhancer activity in an in vitro splicing reporter are indicated by blue triangles. Julie L. Aspden contributed to this figure.

general motifs (Crooks et al. 2004; Schneider and Stephens 1990). Binding sites for many splicing regulators are known to be highly degenerate (Chen and Manley 2009); therefore, such motifs might correspond to sequence elements bound by proteins. Upon inspecting hexamer clusters, I found GGAA-containing ESE motifs present near both 5'ss and 3'ss, regardless of intron length (Figure 2.3). This 4-mer is part of the binding site of multiple hnRNP and SR proteins, including hnRNP-H, hnRNP-F, and SRp55 (Goren et al. 2006). I observed a motif unique to short introns, CTGGA, as well as motifs unique to long introns, CGCA and A[A/G/C]CA[A/G/C]C (Figure 2.3). By clustering similar hexamers, I identified motifs that are shared and distinct between short and long introns.

Potential ISEs were identified using the RESCUE method, by seeking hexamers overrepresented in introns relative to exons and enriched near weak splice sites relative to strong splice sites (Figure 2.2). As with ESEs, short and long introns were analyzed separately and sequences were identified within 100 bp of each splice site. For introns shorter than 100 bp, the entire intron sequence was used. 96 hexamers were identified in short introns at 5'ss and 76 hexamers at 3'ss (Figure 2.4).

Fifteen sequences found in short introns were located both near 5'ss and 3'ss. 78 hexamers were identified in long introns at 5'ss and 43 hexamers at 3'ss, seven of which were enriched at both splice sites (Figure 2.5).

Twenty-four putative ISEs were found near 5'ss in both long and short introns and 10 ISEs were found near 3'ss of both intron length classes. There is a greater overlap of ISEs found in short and long introns than ESEs found near both intron lengths. Similar ISEs were clustered into motifs. Most ISE clusters were found to be AT-rich and many were present near both short and long introns. CAA motifs were preferentially found near weak 3' splice site of long introns. Although some clustered hexamers revealed motifs preferentially found near short or long introns (e.g., TAAT and T[T/C]TC, respectively), sequences containing these motifs could be identified near both length classes (Figures 2.4 and 2.5).

I next looked to see if these enhancer sequences were enriched in positions closer to the splice sites, farther away from the splice sites, or evenly distributed across the search space—up to 100 bp from a splice site. Given that the search space length varied by the length of the exon or intron, I divided the search space into four equally sized bins to indicate a proximal, distal, or intermediate distance from each splice site. As a control, I also compared the distribution of SREs observed near weak splice sites to those observed near strong splice sites. As expected, there was a greater enrichment of enhancer sequences near weak splice sites than strong splice sites (Figure 2.6). ESEs appear to be evenly distributed across the search space, while ISEs show some positional biases (Figure 2.6). ISEs identified near weak 5' splice sites of short introns tend to be more distal to the 5' splice site. This trend is also observed in the shuffled control sequences, indicating that the trend may be due to a general occurrence of A and T nucleotides, as the ISEs are AT-rich (Figure 2.6.A, black lines). I also observe a trend for ISEs identified near weak 3' splice

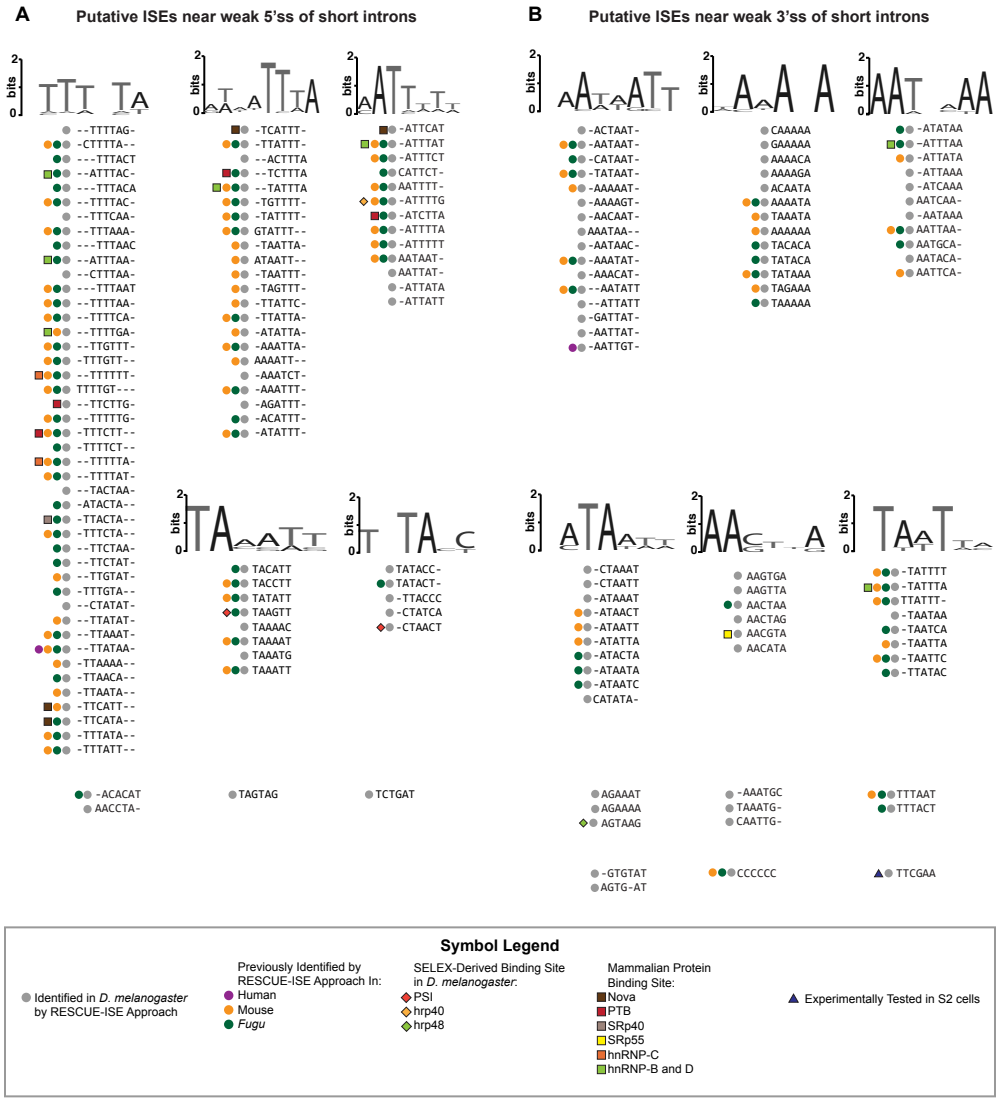


Figure 2.4: **Hexamers and motifs enriched in introns and near weak splice sites of short constitutive introns.** Putative ISEs near (A) weak 5' splice sites and (B) weak 3' splice sites of short introns. Hexamers were clustered by sequence similarity and sequence logos are given when a cluster contains at least 4 hexamers. Description of symbols can be found in the legend of Figure 2.3. Julie L. Aspden contributed to this figure.

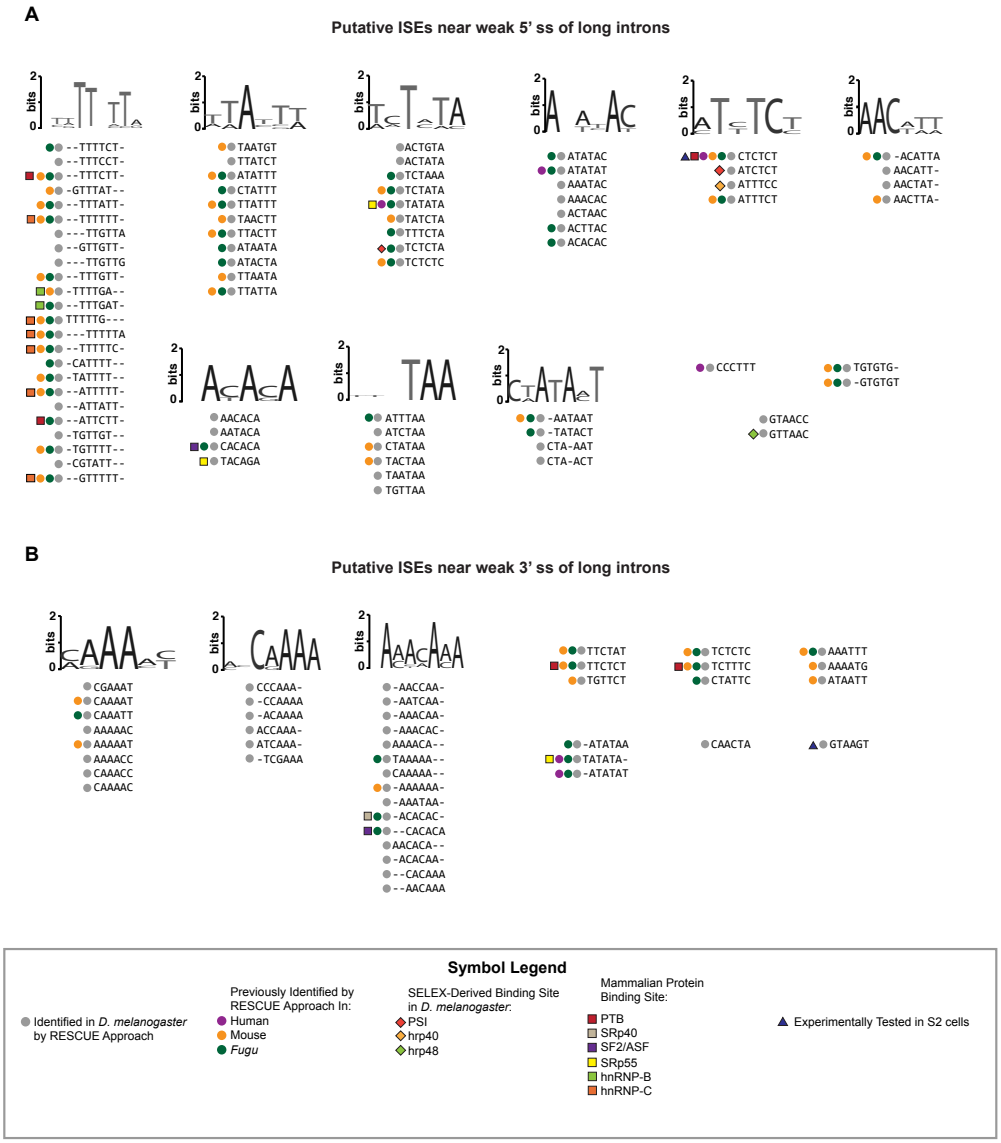


Figure 2.5: **Hexamers and motifs enriched in introns and near weak splice sites of long constitutive introns.** Putative ISEs near (A) weak 5' splice sites and (B) weak 3' splice sites of long introns. Hexamers were clustered by sequence similarity and sequence logos are given when a cluster contains at least 4 hexamers. Description of symbols can be found in the legend of Figure 2.3. Julie L. Aspden contributed to this figure.

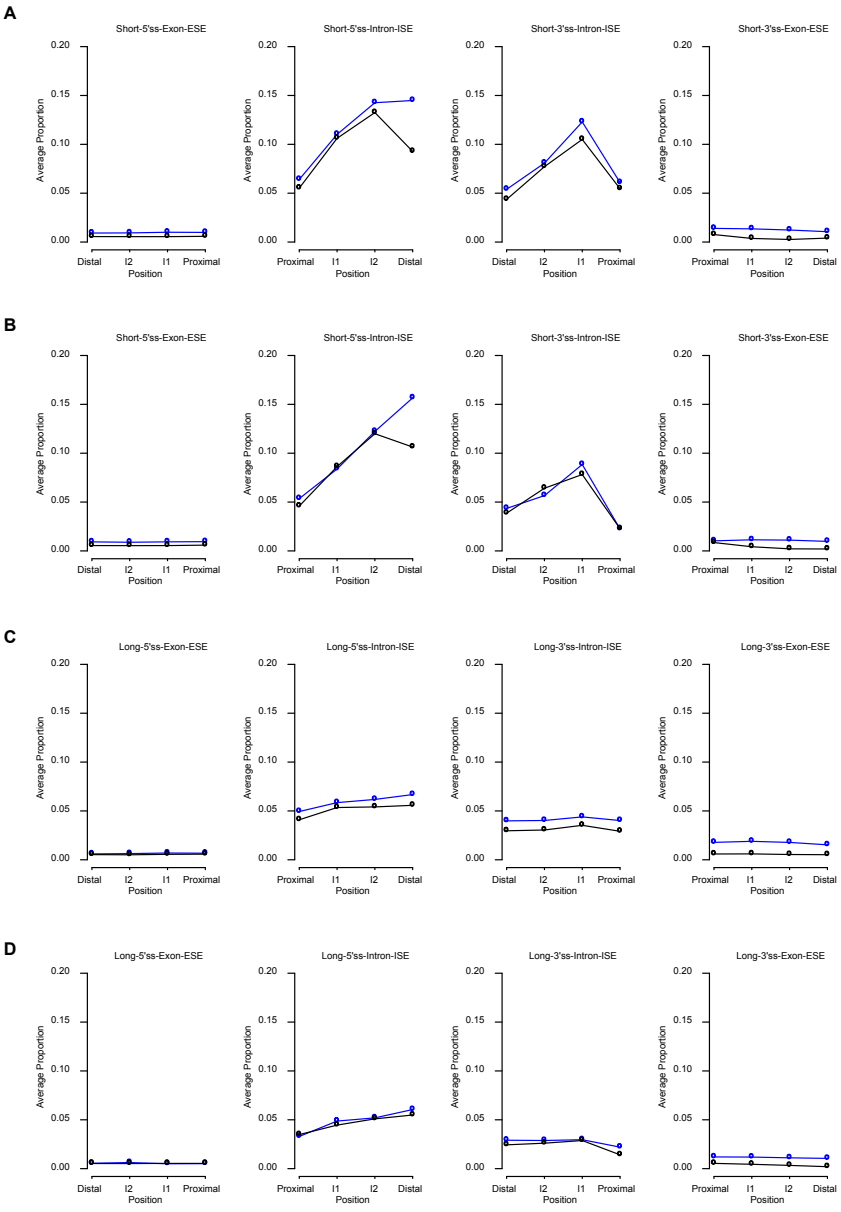


Figure 2.6: **Positional biases of enhancers.** Proportion of enhancers identified in sequence windows near or within short introns, where the splice sites are (A) weak or (B) strong. Proportions of enhancers identified in sequence windows near or within long introns, where the splice sites are (A) weak or (B) strong. Blue lines indicate the proportion of enhancers in each sequence window. Black lines are the proportion of enhancers in a random shuffle of the sequence window.

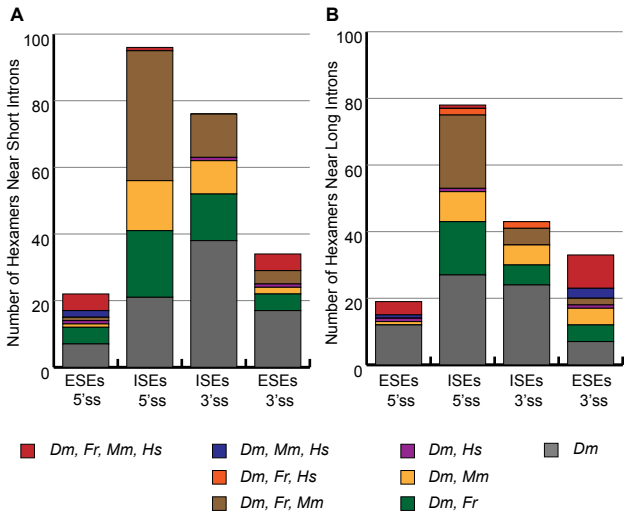


Figure 2.7: **A majority of *D. melanogaster* RESCUE-identified ESEs and ISEs are identical to those found in vertebrates.** The number of hexamers found near (A) short introns or (B) long introns that are shared with pufferfish (*Fugu rubripes*, *Fr*) , mouse (*Mus musculus*, *Mm*), and/or human (*Homo sapiens*, *Hs*) is shown along with the number of hexamers that were uniquely identified in fly (*Drosophila melanogaster*, *Dm*).

sites of short introns to occur more proximal to the 3' splice site (Figure 2.6.A).

2.2.3 58% of RESCUE-identified *D. melanogaster* hexamers are identical to those found in vertebrates

I compared the putative ESEs and ISEs I found in *Drosophila* with those identified in other vertebrate species using the same method, and found a significant overlap with human (*Homo sapiens*), mouse (*Mus musculus*), and pufferfish (*Fugu rubripes*) sequences (Fisher's exact test, p-val < 2.2e-16) (Fairbrother et al. 2002; Yeo et al. 2004). 57 of 99 putative *Drosophila* ESEs and 136 of 231 putative *Drosophila* ISEs were previously identified in one or more of the vertebrates (Figures 2.3, 2.4, and 2.5). I found that of the three species, *Drosophila* had the most overlap with *Fugu* (Figure 2.7), which may arise from the similar intron length distributions between the two species (Figure 2.1) (Yeo et al. 2004; Lim and Burge 2001). Among *Drosophila* ISEs, 45% (104 of 231) are identical to *Fugu* ISEs. A large proportion of ISEs are also identical to mouse; however, very few are shared with human (Figures 2.4, 2.5, and 2.7).

I observed that ESEs were more conserved across all four species than ISEs, perhaps due to added evolutionary constraint from the protein coding sequence of exons (Figure 2.7). 20% of predicted fly ESE hexamers (20 of 99) are shared with human, mouse, and

pufferfish predicted enhancers, with a greater proportion of shared ESEs near 3' splice sites of long *Drosophila* introns (30%, 10 of 33) (Figure 2.7). In contrast, only two putative ISEs are conserved between all four species: CTCTCT and TTATAA. CTCTCT is predicted to be a binding site for the splicing regulator PTB (Chen and Manley 2009; Robida et al. 2010), while no known cognate binding protein for TTATAA was found through literature searches. In HeLa cells, PTB has been shown to activate splicing of exons containing PTB binding sites in the downstream intron (Llorian et al. 2010). The CTCTCT ISE identified in our study was found in long introns near weak 5' splice sites, consistent with the location of PTB-bound enhancer sequences. In addition to the shared ESEs and ISEs between fly and vertebrates, there were 5 fly ESEs that were identified as vertebrate ISEs and 14 fly ISEs identified as vertebrate ESEs. These may be bound by proteins that can act both from exonic and intronic locations but are preferentially enriched in exons or introns due to inherent differences in genome composition of different organisms. All but one of these 19 enhancers were shared specifically with human and/or mouse. These sequences are identified as splicing enhancers in both vertebrates and fly, despite the difference in their enrichment in exonic sequences versus intronic sequence. Differences in exonic sequences may be the result of different codon usages in different organisms.

2.2.4 Overlap with known RNA protein binding sites

To identify potential cognate proteins for the putative ESEs and ISEs, I compared the hexamers against SELEX-derived binding sites of six *D. melanogaster* proteins (Shi et al. 1997; Amarasinghe et al. 2001; Blanchette et al. 2009) and against multiple mammalian RNA binding proteins defined in ESRSearch (Goren et al. 2006) and ESEfinder (Cartegni et al. 2002; Smith et al. 2006) (Figures 2.3, 2.4, and 2.5). Though there may be differences in the RNA binding specificities of mammalian and *Drosophila* RNA binding proteins, there are examples of proteins whose binding motif is well conserved such as PTB/Hephaestus and Nova/Pasilla (Pérez et al. 1997; Robida et al. 2010; Jensen et al. 2000; Brooks et al. 2011). I found one putative ESE, ATGCGG, to be a high affinity binding site for *Drosophila* hrp36, hrp38, and hrp40, despite the fact that their SELEX-derived consensus motifs are distinct. There are regions of the transcriptome that are known to be bound by these three hnRNPs, though not necessarily simultaneously (Blanchette et al. 2009). I identified several SR and hnRNP recognition motifs in both ESEs and ISEs (Figures 2.3, 2.4, and 2.5), further supporting the observation that SR proteins do not exclusively bind to exonic sequences and hnRNPs do not exclusively bind to intronic sequences or act as splicing silencers (Sanford et al. 2009, 2008; Blanchette et al. 2009; Llorian et al. 2010). In addition, the putative ISE TCTATC, found near weak 3' splice sites of long introns, was recently identified in C. elegans as a binding site for HRP-2, a homolog to hnRNPs Q and R (Kabat et al. 2009).

2.2.5 Hexamers enriched in conserved regions of constitutively spliced introns

Because stringent evolutionary constraints on intronic sequences have been used to indicate function (Yeo et al. 2007; Voelker and Berglund 2007; Kabat et al. 2006), I identified hexamers overrepresented in conserved regions within the set of constitutive introns. Code used in this analysis was contributed by Anna I. Podgornaia. Introns are more amenable to the identification of conserved sequence elements than exons, because they are free from protein coding constraints. This search identified additional SREs in introns and indicated which RESCUE-identified ISEs tend to be conserved. The PhastCons model was used to define genomic regions under evolutionary constraint between *D. melanogaster*, 11 other *Drosophila* species, and 3 additional divergent insects (Siepel et al. 2005; Rhead et al. 2010) (http://genome.ucsc.edu). This statistical model allowed us to identify conserved sequences across evolutionary distances greater than those used to identify conserved intronic sequences in vertebrates or in worm (Siepel et al. 2005; Kabat et al. 2006; Yeo et al. 2007; Voelker and Berglund 2007). As before, I separated constitutively spliced introns into short and long introns. I identified hexamers that were overrepresented in conserved regions of introns relative to non-conserved regions (Bonferroni-corrected p-value < 0.001 for 4,096 tested hexamers).

One hexamer was significantly enriched in conserved regions of short introns, CTAATT. In general, short introns did not overlap with PhastCons regions. Only 1% of all possible 4,096 hexamers overlapped with at least on PhastCons region in short introns compared to 12% of hexamers in long introns.

I identified 298 hexamers in long introns that were significantly enriched in conserved regions (Figures 2.8 and 2.9).

CTAATT was also found in conserved regions of long introns. Similar hexamers were clustered to identify common motifs shared by multiple sequences and most clusters were found to be markedly AT-rich in nature (Figure 2.8). Within the conserved elements in long introns, I identified previously reported high affinity binding sites for multiple *D. melanogaster* and mammalian proteins. 73 conserved hexamers overlapped with the 231 hexamers identified via RESCUE-ISE analysis, indicating which RESCUE-identified ISEs are also more likely to be phylogenetically conserved among insects. The hexamers CTCTCT and TTATAA identified as ISEs in fly, pufferfish, mouse, and human are also enriched in conserved regions of the fly genome, further supporting their functional role in splicing. I compared *Drosophila* conserved intronic hexamers to those identified in conserved regions of human introns (Yeo et al. 2007) and to 40 hexamers reported as conserved in regions near alternative introns in C. elegans (Kabat et al. 2006) (Figures 2.8 and 2.9). 35% (105 of 298) of conserved intronic hexamers in fly are also conserved intronic hexamers in human and 10 fly hexamers are also conserved intronic hexamers in worm (Figures 2.8 and 2.9). Most conserved intronic sequences shared between fly and human are AT-rich (Figure 2.8). However, the significance of this is uncertain as it is

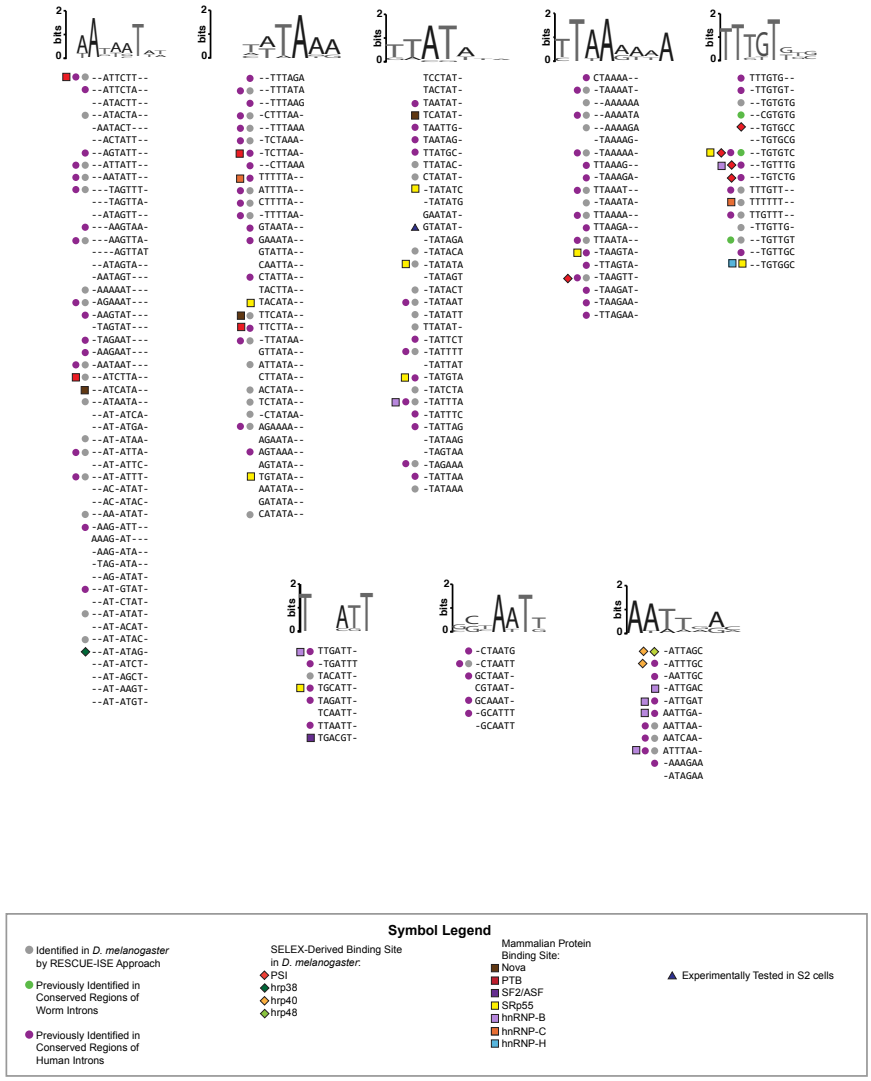


Figure 2.8: **AT-rich conserved hexamers and motifs identified in long constitutively spliced introns.** Hexamers with at least four A or T nucleotides were considered AT-rich. Hexamers identified using the RESCUE-ISE method is shown in gray circles. Hexamers identical to those found in conserved regions in worm (Kabat et al. 2006) or human (Yeo et al. 2007) are in green or purple circles, respectively. Hexamers containing high affinity binding sites for *D. melanogaster* or mammalian proteins are indicated, similar to Figure 2.3. A blue triangle indicates RESCUE-identified ISE hexamers that were tested in an in vitro reporter construct. Julie L. Aspden contributed to this figure.

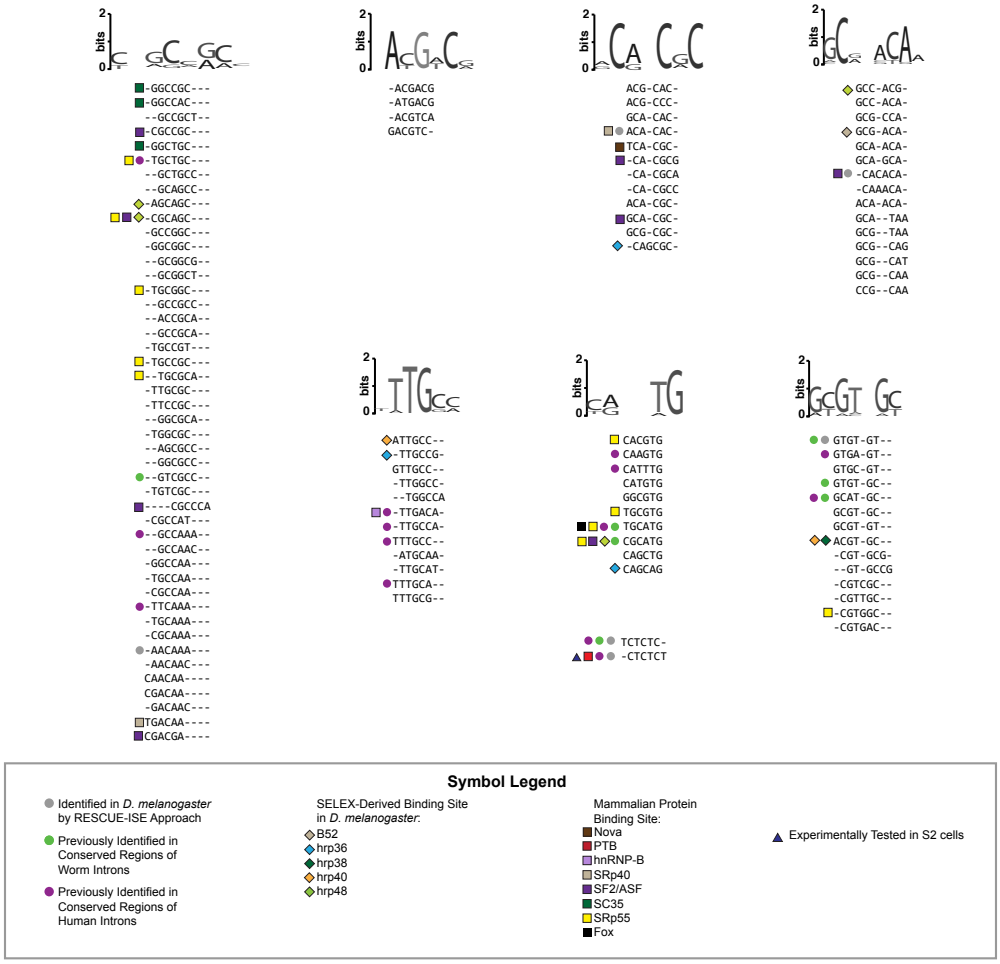


Figure 2.9: **Non-AT-rich conserved hexamers and motifs identified in long constitutively spliced introns.** Hexamer clusters with a minority of AT-rich sequences identified in conserved sequences of long introns. Hexamers identified using the RESCUE-ISE method is shown in gray circles. Hexamers identical to those found in conserved regions in worm (Kabat et al. 2006) or human (Yeo et al. 2007) are in green or purple circles, respectively. Hexamers containing high affinity binding sites for *D. melanogaster* or mammalian proteins are indicated, similar to Figure 2.3. A blue triangle indicates RESCUE-identified ISE hexamers that were tested in an in vitro reporter construct. Julie L. Aspden contributed to this figure.

possible that the sequence overlap between fly and human is an artifact resulting from the AT-rich nature of introns in general.

2.2.6 Computationally predicted ESEs and ISEs stimulate cassette exon inclusion *in vivo*[2]

To test whether putative ESEs and ISEs are sufficient for splicing enhancer activity *in vivo*, their ability to stimulate splicing in a mini-gene reporter was examined. These experimental validations were done by Julie L. Aspden. Anna I. Podgornaia also assisted in the design of the experiments. The mini-gene consisted of an alternatively spliced cassette exon event from an *Drosophila* endogenous gene (*pep*); therefore, the activity of putative ESEs and ISEs was monitored in a different context from the constitutive splicing events where they were identified. ESEs were inserted into the 101 nt cassette exon, which is within 100 nt of both 3' and 5' splice sites (Figure 2.10). ISEs were tested in the long (811 nt) upstream intron within 100 nt of the 3' splice site. The downstream intron of the cassette exon was also long (252 nt). Representative putative ESE and ISEs from long and short introns, and from both 5' and 3' splice site locations were tested.

The activities of seven ESEs were examined alongside the mini-gene reporter without any inserted sequence (-) (Figure 2.10.B, lane 1). To assess whether the effect of inserting hexamers was specific, and not just the result of inserting additional sequence within the cassette exon, a neutral control hexamer sequence was also tested (ATAGTA, N). This hexamer was selected based on its distinct sequence composition from my predicted ESEs. The neutral hexamer showed exon inclusion levels similar to the empty vector (Figure 2.10.B, lane 2).

Among hexamers selected to test for enhancer activity were sequences previously identified in other organisms, for example CTGGAT (ESE-A), which stimulated cassette exon inclusion from 12% with no inserted sequence (-) to 66% (Figure 2.10B, lane 3 and Figure 2.10.C). A single point mutation (PM) to ESE-A resulted in a hexamer that was underrepresented near weak splice sites and not predicted to possess enhancer activity. Julie L. Aspden found that the point mutation exhibited exon inclusion levels similar to background (Figure 2.10.B, lane 4). She also tested several hexamers predicted to be bound by splicing factors. TGTGGA is recognized by mammalian hnRNP H, the ortholog of *Drosophila* Glorund (Barbosa-Morais et al. 2006), and it exhibited a strong enhancing effect on cassette exon inclusion (Figure 2.10.B, lane 5). A novel fly-specific ESE, CGGATG, also showed a stimulatory effect. Out of seven ESEs tested, six exerted a statistically significant stimulatory effect on cassette exon inclusion (t-test, $p<0.05$) (Figure 2.10.C). ESEs identified near short and long introns showed no difference in activity in the splicing reporter, even though the introns surrounding the cassette exon are both long.

[2]Most of this subsection was written by Julie L. Aspden.

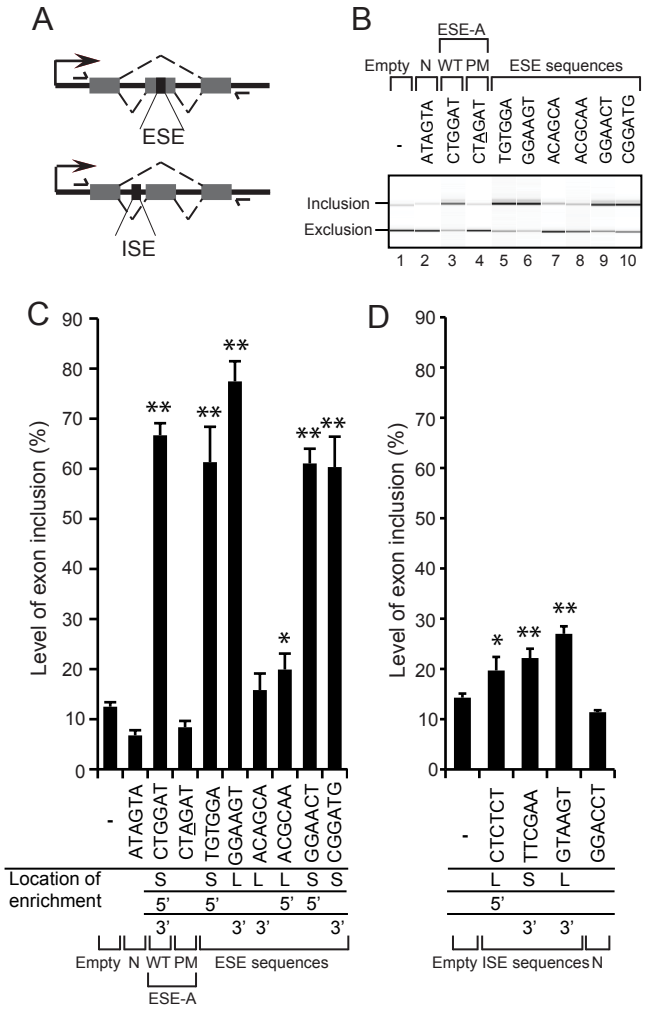


Figure 2.10: **Predicted ESEs and ISEs exhibit stimulatory activity in mini-gene reporter assay.** Figure was created by Julie L. Aspden. (A) Schematic of mini-gene reporter with an alternative cassette exon. ESEs and ISEs were cloned into the cassette exon (101 nt) and upstream intron (811 nt), respectively, as indicated (downstream intron is 252 nt). RT-PCR primers designed to the plasmid backbone are shown as arrows. (B) Semi-quantitive RT-PCR Bioanalyzer 2100 image indicates that ESEs stimulate cassette exon inclusion compared to empty vector (-) and negative control hexamers (PM and N). Wild type (WT) ESE-A is shown alongside a single point mutant (PM), which is underlined. A neutral hexamer control is also shown (N). (C and D) RT-PCR quantitation of cassette exon inclusion levels with mini-genes containing putative ESEs (C) and ISEs (D). Error bars represent mean ± SD of three independent experiments. Asterisks indicate significant differences from empty vector control (t-test; *$p < 0.05$ and **$p<0.005$). The location at which each SRE was identified is indicated under the bar chart, from long (L) or short (S) introns, at 5' or 3' splice sites.

The activities of three ISEs were tested when inserted in the intron near the 3' splice site of the cassette exon (Figure 2.10.A). All had small but significant effects on exon inclusion (t-test, $p<0.05$) (Figure 2.10.D). A neutral hexamer (N), having a different sequence composition from identified ISEs, exhibited background levels of exon inclusion, indicating that ISE effects are specific. One tested ISE matches the consensus 5' splice site sequence, GTAAGT, for *D. melanogaster* (Schwartz et al. 2008). Addition of this splice site sequence did not introduce a cryptic splice site. It has been shown that, in some circumstances, neighboring splice sites can assist in splice site recognition (Chiara and Reed 1995; Hicks et al. 2010). One of the exceptionally conserved ISEs, CTCTCT, was also tested and had significant enhancement activity. Interestingly, TTCGAA, which was identified from short introns, was just as active as predicted ISEs from long introns even though the mini-gene has long introns. Two of the ISEs tested were found near 3' splice sites but were still active near the 5' splice site in the reporter. Two other putative ISEs, identified from earlier iterations of RESCUE analysis, but not above the final cut-off, were also tested in the reporter and stimulated cassette exon inclusion (data not shown). The difference in the magnitude of enhancement between the tested ESEs and ISEs may be due to effects from local sequence context or relative position (Goren et al. 2006; Zhang et al. 2009) (Figure 2.10.D).

2.3 Discussion

I used the RESCUE method (Fairbrother et al. 2002) to predict 99 ESEs and 231 ISEs that were overrepresented near weak splice sites of constitutive introns in *D. melanogaster*. Within the set of computationally predicted SREs, I identified binding sites of multiple *Drosophila* and mammalian splicing regulators, implicating putative cognate binding proteins for the enhancer sequences. I found many SR and hnRNP binding sites within the set of ESEs and ISEs, giving further evidence that these proteins can act as enhancers and bind to both exons and introns. Seven ESEs and three ISEs were tested *in vivo* for enhancer activity when introduced in a mini-gene reporter assay, and all but one showed a statistically significant enhancement of splicing.

I identified putative SREs separately near short and long introns and found that the majority of enhancer sequences were specific to each intron class, suggesting on average, genome-wide differences in splicing regulation that depend on intron length. Splice site recognition is thought to occur through the definition of introns or exons, depending on intron length. Perhaps the distinct regulatory sequences found near different length introns are associated with factors preferentially used for intron or exon definition. However, when putative SREs were tested for their ability to stimulate cassette exon inclusion of a mini-gene reporter, where the introns were long, there was no difference in activity between those SREs found near long and short introns. This may be the result of the different context in which the SREs were tested from where they were found.

A previous study of SREs cautions that many computational predictions have been "too successful" because now at least 75% of a typical human exon sequence can be shown to influence splicing (Zhang et al. 2009). Our study indicates which of these many SREs are particularly relevant by identifying SREs that overlap between *Drosophila* and vertebrates. I found that a significant portion (58%) of fly putative enhancer sequences were identical to human, mouse, or pufferfish enhancer sequences. Moreover, a substantial fraction (20%) of fly ESEs were identical to ESEs found in all three vertebrate species, highlighting enhancer sequences whose function has been maintained throughout evolution.

In addition to RESCUE-identified ISEs, I also made use of 15 insect species to report a set of intronic sequence elements phylogenetically conserved at a greater evolutionary depth than previous analyses (Siepel et al. 2005; Kabat et al. 2006; Yeo et al. 2007; Voelker and Berglund 2007). Some of these conserved intronic hexamers may not be involved in splicing; however, 73 of these sequences were also identified as ISEs using the RESCUE approach. The hexamers CTCTCT and TTATAA were highlighted as exceptionally conserved, since they were identified as ISEs through the RESCUE method in fly and three vertebrates and were also enriched in conserved intronic regions of the *Drosophila* genome. The heaxmer CTCTCT is predicted to be recognized by the splicing regulator PTB, which is itself conserved between fly and vertebrates (Barbosa-Morais et al. 2006). I did not find a previously reported cognate binding protein for TTATAA, yet this orphan putative regulatory sequence appears important due to its conservation. Given that the sequence is a palindrome, perhaps it is acting through RNA secondary structure. The sequence may also act through splicing regulation at the DNA level by affecting transcription rates (Kornblihtt 2006) or chromatin states (Schwartz and Ast 2010). Most identified SREs are likely binding sites for *trans*-acting regulatory proteins; however, some may regulate splicing through these alternative mechanisms.

This study presents the most comprehensive computational analysis of splicing enhancer sequences in *Drosophila melanogaster* to date, and it has revealed splicing regulatory elements whose function is conserved across metazoan evolution. Since splicing patterns can differ between tissue type and developmental stages, it is also necessary to study splicing regulation in diverse cellular contexts (Matlin et al. 2005; Zhang et al. 2009), taking into account the SREs' role in the pantheon of splice affecters.

2.4 Methods

Intron coordinates in the *D. melanogaster* genome

The *D. melanogaster* genome sequence and annotations from FlyBase release 5.4 (Tweedie et al. 2009), which includes 37,803 constitutively spliced introns, were used for the RESCUE method. Using modified scripts from the *Drosophila* Exon Database (Lee et al.

2004), constitutively spliced introns were identified as intron coordinates that are present in all isoforms of the same gene.

Measurements of splice site strength with MaxEntScan

MaxEntScan was used to assess how well a sequence conforms to the well-established 5'ss or 3'ss consensus motif (Yeo and Burge 2004; Schwartz et al. 2008). MaxEntScan scoring was done by Anna I. Podgornaia. This score was taken as an indication of splice site strength. The 5'ss sequence is defined as position [-3,+6] and the 3'ss sequence at position [-20, +3], relative to the exon-intron junction. MaxEntScan models short sequence motifs and accounts for relationships between adjacent and non-adjacent nucleotide positions.

Defining short and long introns in *D. melanogaster*

A histogram was created from the lengths of all constitutive introns in FlyBase r5.4. By visual inspection, I identified a sharp peak in the distribution of lengths with most introns $\leq$80nt in length ("short") and a tail corresponding to introns longer than 80nt ("long").

RESCUE-ESE and RESCUE-ISE method

Code used to implement this method was contributed by Anna I. Podgornaia. The frequency of all 4096 possible hexamers was determined, using a sliding 6 bp window and allowing overlaps, in each of the following locations: exonic sequence, intronic sequence, near a weak 5'ss or 3'ss, and near a strong 5'ss or 3'ss. Sequences within 100 bp of the exon-intron boundary, excluding the nucleotides [-3, +6] relative to the 5' ss and [-20, +3] relative to the 3' ss, were used for the RESCUE method (Fairbrother et al. 2002). If the intron or exon was <100 bp, then the entire intron sequence, excluding splice sites, was used. Hexamer frequencies were calculated separately for exonic sequences near short and long introns and intronic sequences within short and long introns. When identifying ESEs near 5'ss, the length of the downstream intron was used to separate between short and long introns, while the length of the upstream intron was used to separate by length when identifying ESEs near 3'ss.

ΔEI, Δ5WS, and Δ3WS scores were calculated for each hexamer using the formula as described in the Supporting Online Material for Fairbrother et al. (2002). Hexamers with ΔEI and Δ5WS, or Δ3WS, scores above 2.5 (p-value < 0.01), were selected as potential 5' and potential 3' ESEs, respectively. Therefore, the p-value for significance of each putative enhancer is less than 10-4, given that the significance threshold for the two independent scores has a p-value of 0.01. A similar procedure was performed for identifying ISEs where a ΔIE, Δ5WS, and Δ3WS was calculated for each hexamer found in intron sequences.

Positional biases of enhancers near weak splice sites

To determine if enhancer sequences had a positional bias, relative to the splice sites, I counted the frequency of each set of enhancers at positions near the splice sites, distal to the splice sites, and at intermediate distances to the splice sites.

Exon and intron sequence within 100 bp of the intron-exon boundary was used for the analysis. If the exon or intron was <100 bp, the entire sequence was used. Due to the varying lengths caused by shorter introns or exons, each region was divided into four equal length windows (proximal to the splice site, distal to the splice site, and two intermediate windows) and the frequency of enhancers was divided by the length of each window to get a proportion of enhancers found in each window. As a control, the nucleotides in each sequence window were randomly shuffled and enhancer frequencies were calculated from these shuffled sequences. Enhancer frequencies were obtained separately in regions near weak splice sites and strong splice sites. Frequency of enhancers were only determined in the sequence region from which they were identified as enhancers. For example, in exonic sequences 100 bp upstream of 5' splice sites, with a short downstream intron, the frequency of ESEs found near 5' splice sites near short introns was determined.

Clustering hexamers

The following manipulations were done separately for hexamers in the different intron length and splice site groups. An edit distance (number of insertions, deletions, and substitutions) was calculated between all possible pairs of hexamers, using the Levenshtein distance algorithm (Böckenhauer and Bongartz 2007; Crooks et al. 2004; Schneider and Stephens 1990). The MATLAB 'linkage' function was used to generate a hierarchical cluster tree from the unweighted average distances, such that sequences with the lowest edit distances would fall into the same cluster. The tree was then visualized using the MATLAB 'dendrogram' function with the 'colorthreshold' parameter. In the original RESCUE-ESE study, a dissimilarity cutoff of 2.7 was used to select clusters of hexamers (Fairbrother et al. 2002). I increased the cutoff up to 3.0 whenever it led to the inclusion of several additional hexamers to any cluster. Clusters composed of four or more hexamers were aligned using ClustalW and a sequence logo was generated for each cluster using WebLogo (Schneider and Stephens 1990; Crooks et al. 2004; Larkin et al. 2007). A majority of the code used to cluster the hexamers was created by Anna I. Podgornaia.

Identifying high affinity binding sites from SELEX-derived binding matrices

The SELEX-derived binding affinities for B52, PSI (Amarasinghe et al. 2001), hrp36, hrp38, hrp40, and hrp48 (Blanchette et al. 2009) in *D. melanogaster* and SF2, SC35,

SRp40, and SRp55 (Smith et al. 2006; Cartegni et al. 2002) in human were used to determine high affinity binding sites for each protein. The SF2 binding motif from functional SELEX of the IgM- and BRCA1-derived mini-gene was used. The average and standard deviation of position weight matrix (pwm) scores against all exons and introns in FlyBase r5.4 were used to calculate a Z-score for a given hexamer. When comparing against hrp36, hrp38, hrp40, hrp48, and SRp55 pwms, hexamers with a Z-score ≥ 2 (p-val ≤ 0.05; two-tailed) were considered high affinity binding sites. The binding site for PSI, SF2, SC35, and SRp40 is greater than 6nt; therefore, hexamers were compared against all sub-hexamer windows within the matrix. A Bonferroni-correction for the multiple sub-hexamer windows was used to maintain an overall p-val ≤ 0.05 for matches to these binding sites. The exact binding sequences for B52 are reported in (Shi et al. 1997); therefore exact matches to the conserved 17 nt core of the B52 binding site was used to identify binding sites.

Identifying previously published enhancer sequences in vertebrates

Hexamers identified as ESEs and ISEs in human, mouse, or *Fugu* were identified through queries to the ACESCAN2 web server (http://genes.mit.edu/acescan2/index.html).

Identification of conserved intronic hexamers

Anna I. Podgornaia contributed to code used to implement this method. Conserved regions of the genome were defined by PhastCons conserved elements that had a transformed log-odds score greater than 0 (Siepel et al. 2005) and coordinates (dm3) were downloaded from the UCSC Genome Browser public MySQL server (Rhead et al. 2010) (http://genome.ucsc.edu). The frequency of hexamers within 100nt of either splice site of constitutive introns was calculated. A χ^2 statistic with Yate's continuity correction was computed for each hexamer using a two-by-two contingency table as performed with conserved sequences in humans (Yeo et al. 2007). The table for each hexamer compared (a) the number of times the hexamer occurred in PhastCons elements within 100 bp of a splice site versus the number of times all other hexamers occurred in conserved elements, (b) the number of times the hexamer occurred within 100 bp of a splice site versus the number of times all other hexamers occurred within 100 bp of a splice site. Only hexamers with counts greater than 10 were selected for testing. Hexamers with Bonferroni-corrected p-values less than 0.001 were selected as significantly enriched in conserved intronic regions.

Reporter plasmid construction

Reporter construct was created by Julie L. Aspden. This description was written by Julie L. Aspden. The mini-gene reporter (obtained from M. Blanchette, Stowers Institute)

was prepared from exons 1, 2 and 3 of the *Drosophila pep* gene (*CG6143*) fused to EGFP in pMT/V5 (Invitrogen). Oligo pairs containing ESEs were ligated into NheI digested plasmid, 17 bp into exon 4, resulting in the addition of 12 bp, 6 bp of which corresponded to ESEs. ISEs were inserted into a BglII site 84 bp upstream of the cassette exon.

Tissue culture, DNA transfections, RNA purification and RT-PCR

All experimental validation was done by Julie L. Aspden. This description was written by Julie L. Aspden. *Drosophila* Schneider (S2) cells were grown in standard tissue culture conditions at 26°C with M3 supplemented with 5% fetal bovine serum. Plasmid DNA was transfected using Effectene (Qiagen), according to manufacturer's instructions. After 24 hr, CuSO4 was added to a final concentration of 0.5 mM. After a further 48 hr, cells were harvested and total RNA purified. Samples were DNaseI-treated followed by nucleic acid purification. RNA was subjected to reverse transcription using SuperScript II Reverse Transcriptase (Invitrogen) with oligo d(T)15. PCR was performed on the resulting cDNA with HotStar polymerase (Qiagen) and primers designed to anneal to the vector backbone (forward: cgtagaatcgagaccgagg, reverse: gctcctcgcccttgctca). PCR products were examined using a 2100 Bioanalyzer (Agilent) and subsequently quantitated using 2100 Expert Software.

Chapter 3

Identification and quantification of alternative splicing events given RNA-Seq data

3.1 Introduction

At the time our group first received RNA-Seq data to examine alternative splicing, there were no available computational tools to identify and quantify alternative splicing. Here, I describe my approach to first identify splice junction reads in the data and then use those reads to identify alternative splicing events. The RNA-Seq data initially used for method development were from untreated S2-DRSC cells and S2-DRSC cells treated with dsRNA against the *pasilla* gene, a putative splicing regulator (results are discussed in 4.3 on page 53). Excerpts from Brooks et al. 2011 are included and modified in this section. Those excerpts were co-written by Brenton R. Graveley with input from Li Yang, Michael O. Duff, Kasper D. Hansen, Sandrine Dudoit, and Steven E. Brenner and these authors made significant contribution to the design of the method.

Current versions of the software package, JuncBASE, that implements the methods described in this chapter, can be found here: http://compbio.berkeley.edu/proj/juncbase. Although the methods were developed for use of RNA-Seq data from *D. melanogaster*, JuncBASE can be used to examine RNA-Seq data from any organism with a sequenced reference genome.

3.2 Method for aligning RNA-Seq reads to splice junctions

3.2.1 Obtaining splice junction sequences and alignment parameters

A database of 58,212 annotated and 221,388 unannotated (novel) splice junction sequences was created from a merged transcript annotation of FlyBase r5.11 and a modENCODE annotation, MB5 (www.modencode.org). 215,757 of the unannotated splice junctions were generated by joining every annotated exon with all possible downstream exons within the same gene. An additional set of 5,631 novel junction sequences were created by joining every pair of exons from different gene loci that were $\leq$2kb away. Our reads were 37 bp; therefore, splice junction sequences were 62 bp long (31 bp on either side of the splice junction) to ensure a $\geq$6 bp overhang of the read mapping from one side of the junction onto the other. When aligning to our longer 75 bp reads, splice junction sequences are 138 bp. This yielded an even coverage of alignment positions across all splice junctions, given that we allowed up to two mismatches in the alignment (Figure 3.1.A). There is a decrease in coverage for overhang lengths $\leq$ 5nt as a result of reads no longer uniquely aligning to a junction. A 6nt overhang should be used if 2 mismatches are allowed, regardless of the lenght of the reads. Recent updates to the methodology now includes additional splice junctions involving one annotated splice site and the other splice site a possible novel or cryptic splice site (GT or AG dinucleotide) that is < 2kb away. I also create a set of all possible novel splice site pairings as long as the distance between them is <2kb away; however, practically it is difficult to align against these sequences due to their enormous size.

3.2.2 Removing potential false positive alignments

As our splice junction dataset contains nearly four times as many predicted junctions as annotated junctions, I assessed the criteria that could be used to distinguish between true splice junctions and false positive splice junctions. To do this, I generated a set of 5,409,600 splice junctions joining each annotated 5' splice site with 50 randomly drawn annotated 3' splice sites located on a different chromosome and from each annotated 3' splice site with 50 randomly drawn annotated 5' splice sites from a different chromosome. Alignments to these junctions are considered false positives, as such junctions are thought to rarely exist when compared to annotated junctions. Comparison of the alignment results of one lane of data containing 6.2 million paired-end reads to the genome and either the annotated or random splice junctions revealed that the false positive rate could be greatly reduced (0.006% false positive) by requiring at least three different start positions (offsets) for reads spanning the junction (Figure 3.1.B). Our most current method to remove potential

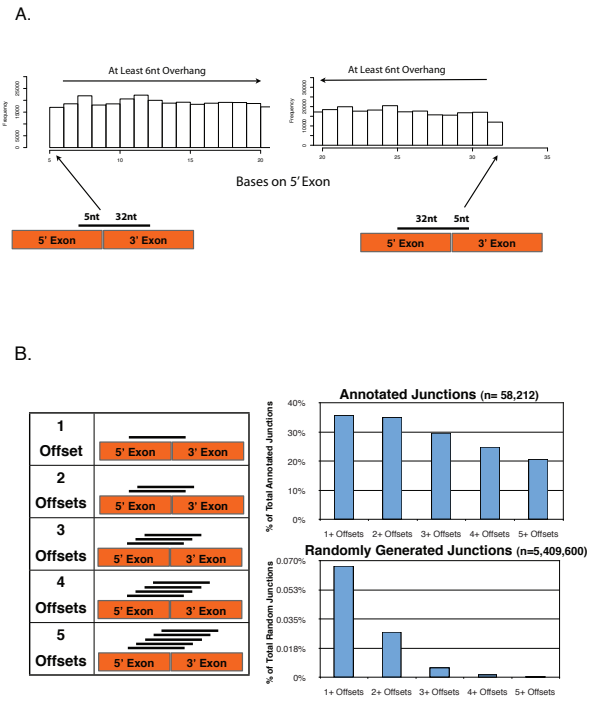


Figure 3.1: **Analysis of optimal overhang for splice junction alignments.** (A) Distribution of overhang positions $\geq$5 nt. A histogram of the number of uniquely aligned reads across all annotated junctions is shown. An even distribution of read alignments across all base positions occurs if at least a 6nt overhang is enforced. (B) Distinguising true junction from false positive alignments. To reduce the number of false positive junction, as determined by randomly generated junctions, a total of 3 alignment start positions (offsets) were required to consider a junction to be truly present.

false positives is by calculating a Shannon entropy score for every junction, taken from the empirical probability of a read aligning to a particular start position across the junction.

$p(i)$ = number of reads at offset i / total reads to junction

Entopy of junction =

$$\sum_i p(i) log_2 p(i)$$

This approach was originally developed by Michael O. Duff and has now been incorporated into my own pipelines. The entropy score cutoff will depend on the depth of sequencing and is typically identified by aligning the reads against a set of randomly generated junctions or the full set of all possible junctions formed from using an annotated splice site to a cryptic splice site and comparing with alignments against a set of annotated junctions.

3.3 Junction Based Analysis of Splicing Events (JuncBASE)

JuncBASE (Junction Based Analysis of Splicing Events) is a series of Python scripts that takes RNA-Seq aligment files in SAM format to identify alternative splicing events, calculate exon exclusion and inclusion counts for each splicing event, and to identify significantly affected events. This is updated from a previous version that used Bowtie (Langmead et al. 2009), spa.pl, and exon_hitter.pl (Brooks et al. 2011; McManus et al. 2010) output files as input. Input was changed to SAM format to maintain consistency with a standardized format and to create a more user friendly version. It is important to note that all splice junctions used to identify significantly affected splicing events are filtered to have at least 3 distinct offsets aligning to the junction (or an entropy score above a given cutoff). Figure 3.2 contains diagrams of exclusion and inclusion isoforms for the eight types of alternative splicing that were examined. Junctions and exon reads that are part of an exclusion isoform are depicted in blue and reads from inclusion isoforms are in red. For both the untreated and *ps*(RNAi) samples, I counted the reads that aligned to the inclusion isoform or the exclusion isoform. For the paired-end alignments, if both ends of a read aligned to an unique regions of an isoform (e.g., one read within a cassette exon and the other read to a junction that includes the cassette exon), the count was only incremented by one.

This recapitulates key aspects of the approach described in Wang et al. (2008) with several distinctions. First, to identify retained intron events, I found that counting read alignments throughout the entire intron (as evidence for the inclusion of the intron) was confounded in cases where additional splice sites reside within the intron; therefore, only reads spanning the exon-intron boundaries of the 5' and 3' splice sites were used as evidence for retention

of the intron (red bars in Figure 3.2). Moreover, each end of the intron was tested separately for intron retention and then p-values were combined for a final p-value associated with the retained intron event (see subsection 3.3.7 on page 46). The method to examine retained intron was co-developed with Kasper D. Hansen. I also extended the Wang et al. method to identify coordinate cassette exons—two or more exons skipped or included as a group. Finally, I did not examine tandem 3' UTR events (alternative polyadenylation) seeing that there were few reads that gave direct evidence of a poly(A) site and that there are no exon reads or junction reads that are specific to the exclusion isoform (alternative polyadenylation analysis done by Kasper D. Hansen, see Methods).

What follows are details of how JuncBASE analyzes each alternative splicing events.

3.3.1 Cassette exons

For every exon in the transcript annotation set, I searched for splice junctions that gave evidence for the skipping of the exon—termed exclusion junctions. An exclusion junction has splice sites flanking an intron that fully contains the exon; thus, when this intron is spliced the exon is necessarily skipped. Splice junction reads that aligned to the cassette exon were considered inclusion junctions. Reads mapping to all inclusion junctions and reads fully contained within the cassette exon were used for the inclusion counts. If there was more than one exclusion junction (in the case that either of the splice sites corresponding to the exclusion junction are also part of an alternative 5' or 3' splice site), the sum of counts for all exclusion junctions was used for the total exclusion counts.

3.3.2 Mutually exclusive exons

Mutually exclusive exons were identified as two or more non-overlapping exons that formed junctions with the same upstream and downstream exon. If a junction supported skipping of all exons in the event, it was still classified as mutually exclusive. Initial versions of JuncBASE removed cases where the exons within a group had a junction read connecting them. Through analysis of additional RNA-Seq data, I found that this removed genuine cases of mutually exclusive events that also had minor coordinate cassette exon isoforms. Due to the mixing of isoforms that are mutually exclusive types and coordinate cassette exon types, I updated the algorithm to proportionally assign reads to each event, using junction reads specific to each isoform type. An example is shown in Figure 3.3. If there were three or more exons in a group, all possible inclusion and exclusion isoforms were treated as separate events. For example, if Exon 3, Exon 4, and Exon 5 are mutually exclusive, one event would treat Exon 3 as the inclusion isoform and the amalgamation of Exons 4 and 5 as the exclusion isoform. Another event would treat Exon 4 as the inclusion isoform and the amalgamation of Exons 3 and 5 as the exclusion isoform. A third event would treat Exon 5 as the inclusion isoform and the amalgamation of Exons 3 and 4 as the

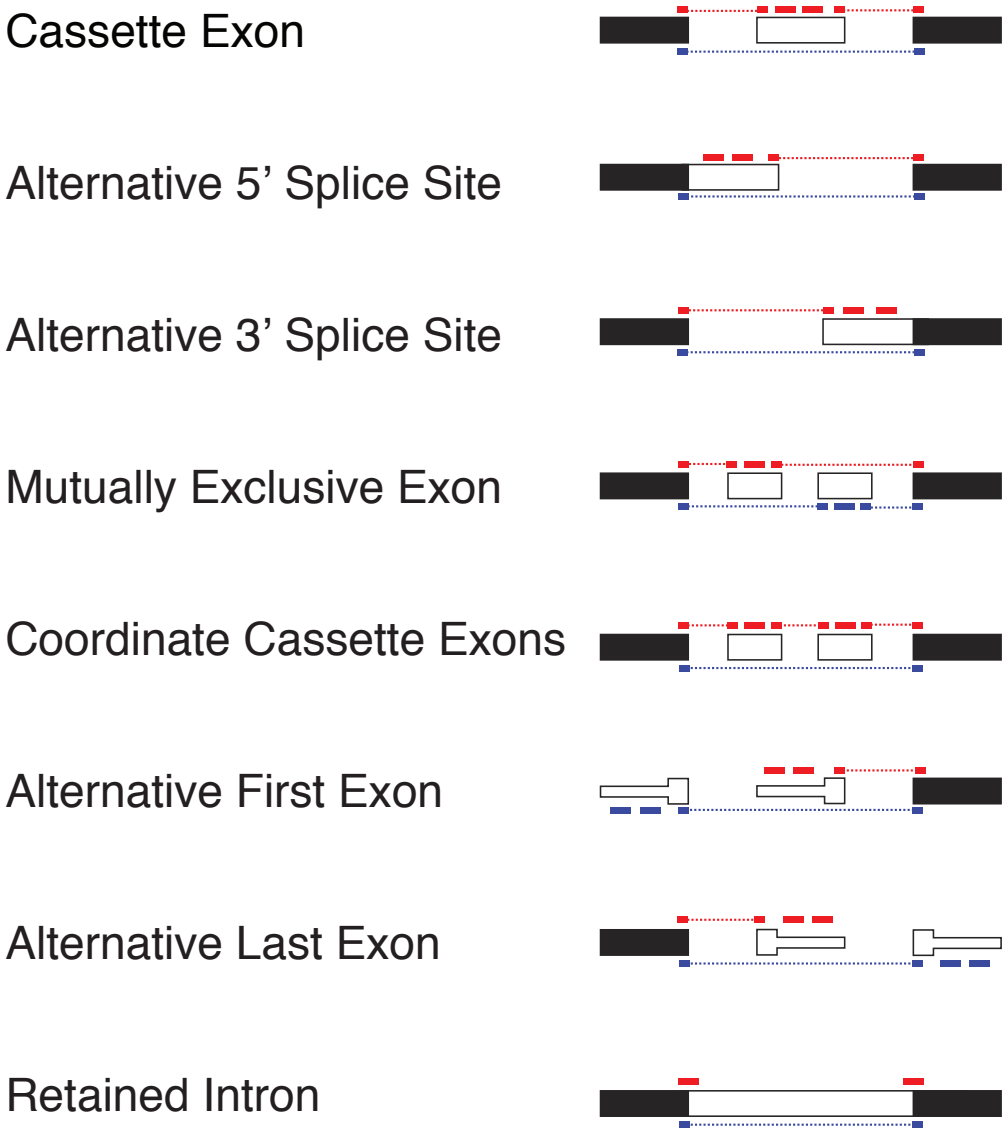


Figure 3.2: **Reads supporting presence of inclusion or exclusion isoform of each type of alternative splicing.** White boxes, alternative exons. Black boxes, constitutive exons. Thinner portions of alternative exons indicate UTR regions.

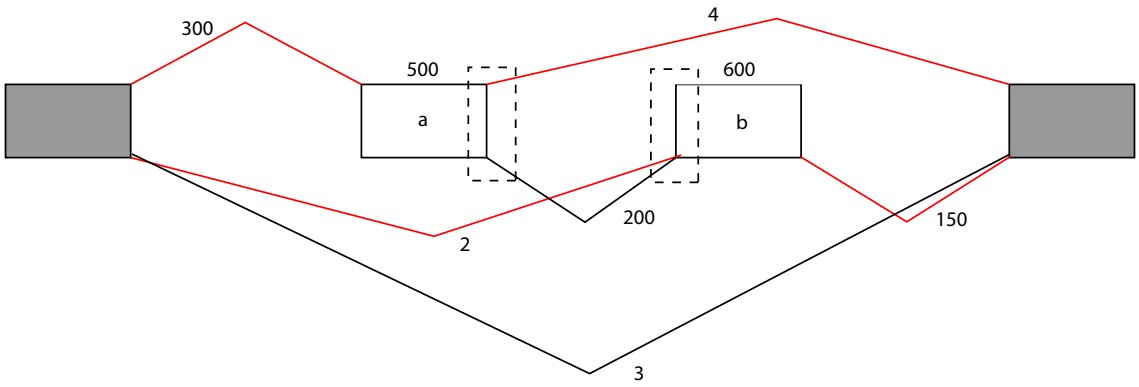


Figure 3.3: **Example of alternative splicing event that includes a mixture of mutually exclusive exons and coordinate cassette exons.** Here, the coordinate cassette exon is the major form. Red lines, mutually exclusive isoform junctions. Dashed box, junctions used for proportional estimates. This is a hypothetical event.

exclusion isoform. Junction reads and reads fully contained within the exons were used for counts.

3.3.3 Coordinate cassette exons

Groups of two or more consecutive and non-overlapping exons that were fully contained in an intron and wholly spliced out (inferred by a splice junction), and thus had evidence of skipping, were identified. From these, I identified cases where all consecutive exons within a group had a junction read connecting them, and classified these as coordinate cassette exons. As mentioned in the explanation of mutually exclusive exons, there may be a mixture of isoform types and therefore reads are proportionally assiged to the coordinate cassette exons based on the junctions that are specific to this alternative splice form. The sum of all connecting splice junctions and reads completely contained within the exons were counted toward the inclusion isoform. Reads aligning to the exclusion junction were counted toward the exclusion isoform.

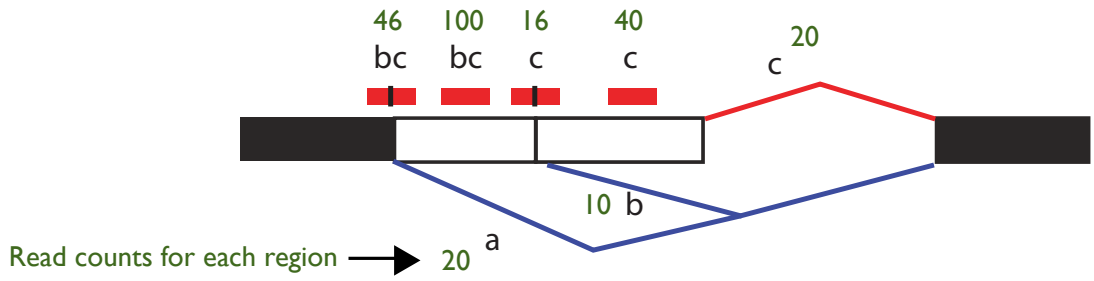
3.3.4 Alternative 5' splice site and alternative 3' splice site

I identified instances where two or more introns (inferred from splice junctions) had the same start (or end) position. Next, the strand of the event is given by the SAM alignment

or inferred based on splice site sequence. Introns were classified as part of an alternative 5' splice site or alternative 3' splice site event based on the strand orientation of the intron. If the adjacent exons of the alternative 5' (or 3') splice sites were overlapping then they were considered alternative 5' and 3' splice sites. If there were three or more introns involved in an event, each intron was treated as the inclusion isoform from a separate event, with all others treated as exclusion isoforms. For example, if there were three introns involved in an alternative 3' splice site choice, they will be treated as three separate alternative 3' splice site events. The three junctions would have the same start position, but three different end positions: (a),(b),(c), where (c) is the splice site that would remove the longest intron. One alternative 3' splice site event would assign reads supporting splice site (a) to the inclusion isoform count and counts supporting (b) and (c) to the exclusion isoform count. Another alternative 3' splice site event would add reads supporting splice site (b) to the inclusion isoform and counts supporting (a) and (c) to the exclusion isoform and so forth. For alternative 5' and 3' splice sites, splice junction reads were used for counts. In addition, reads fully contained within the alternative region of an inclusion isoform were used for the inclusion counts. Reads that aligned to the boundary of the alternative region and the constitutive region of the event, were added to inclusion counts. Recent improvements to JuncBASE includes better quantification in cases of three or more splice site choices. Exon body reads that are shared by different isoforms are proportionally assigned based on the estimates of their relative abundance from splice junction reads. An example is shown in Figure 3.4.

3.3.5 Alternative first exons and alternative last exons

I identified instances where two or more introns (inferred from splice junctions) had the same start (or end) position. Introns that were already part of a cassette exon event, a mutually exclusive event, or a coordinate cassette exon event were removed from this list. Introns were classified as part of an alternative first or alternative last exon event based on the strand orientation of the intron. If the adjacent exons of the alternative 5' (or 3') splice sites were non-overlapping and at least one adjacent exon was a first (or last) exon, then the event was categorized as an alternative first (or last) exon. Similar to alternative 5' and alternative 3' splice sites, if there were three or more introns involved in an event, each intron was treated as the inclusion isoform from a separate event, with all the others treated as exclusion. Splice junction reads and reads fully contained within alternative first or last exons were used (Figure 3.2)



Isoform	Count
a	20
b	(46 * (10/(10+20)) + (100 * (10/(10+20))) + 10
c	(46 * (20/(10+20)) + (100 * (20/(10+20))) + 16 + 40 + 20

Figure 3.4: **Read assignments of shared regions for alternative 5' and 3' splice sites based on relative proportion of isoforms.** The example shown is for a hypothetical alternative 5' splice site event. Read counts for each region is indicated. Here junction reads are represented by the red and blue lines. Red bars with black lines indicate reads at an exon-intron boundary.

3.3.6 Fisher's exact test to identify significantly affected alternative splicing events

Read counts from either the inclusion or exclusion isoforms in the untreated or RNAi sample were used to create a 2 x 2 contingency table (e.g., tables shown in Figure 4.3 on page 57). A Fisher's exact test was performed for each event. In the *pasilla* study, a total of 2,324 tests were performed from the seven event classes described above. A cutoff corresponding to a Benjamini-Hochberg corrected p-value of 0.05 was applied to each of the above seven event classes. Between a choice of a Bonferroni or a Benjamini-Hochberg (BH) mutiple testing correction, the less stringent BH cutoff incorporated additional cassette exon events that appeared valid based on visual inspection of the read alignments. The shift in the proportion of inclusion reads (inclusion reads/(exclusion + inclusion reads)) in each sample was used to determine the shift in the direction of the splicing event. For example, if there was an increase in the percent of reads that supported the inclusion isoform in the RNAi sample, then that event was considered to be normally repressed by PS.

3.3.7 Identifying significantly affected retained intron events

Kasper D. Hansen contributed to the development of this analysis. Every confident junction in the data was examined for evidence of intron retention. For quantifying the retention (inclusion) of the intron, I did not use all reads falling within the intron coordinates, because the intron could contain, for example, internal cassette exons. I found that the most informative reads to quantify the inclusion of an intron were reads that aligned to the intron-exon junction of both the 5' and 3' end of the intron. However, this could lead to some introns with both alternative 5' and 3' splice sites to be misclassified as retained introns. The splice junction reads formed from the splicing of the intron were counted towards the exclusion isoform. I found that simply summing the inclusion counts from the 5' and 3' end of the intron-exon junction could be confounded with alternative 5' and 3' splice sites. Therefore, I applied Fisher's exact test separately at the 5'-end and the 3'-end and combined the resulting two p-values as depicted in Figure 3.5. All events with a Benjamini-Hochberg corrected p-value ≤ 0.05 (based on N = 1,568 for the *pasilla* study) were considered affected.

3.3.8 Identifying significantly affected junctions that are not classified in an event type

To identify potentially unclassified alternative splicing events, read counts to each junction were compared to read counts to all other mutually exclusive junctions on the same strand. When an exon-exon junction is observed, an intron is removed and other introns

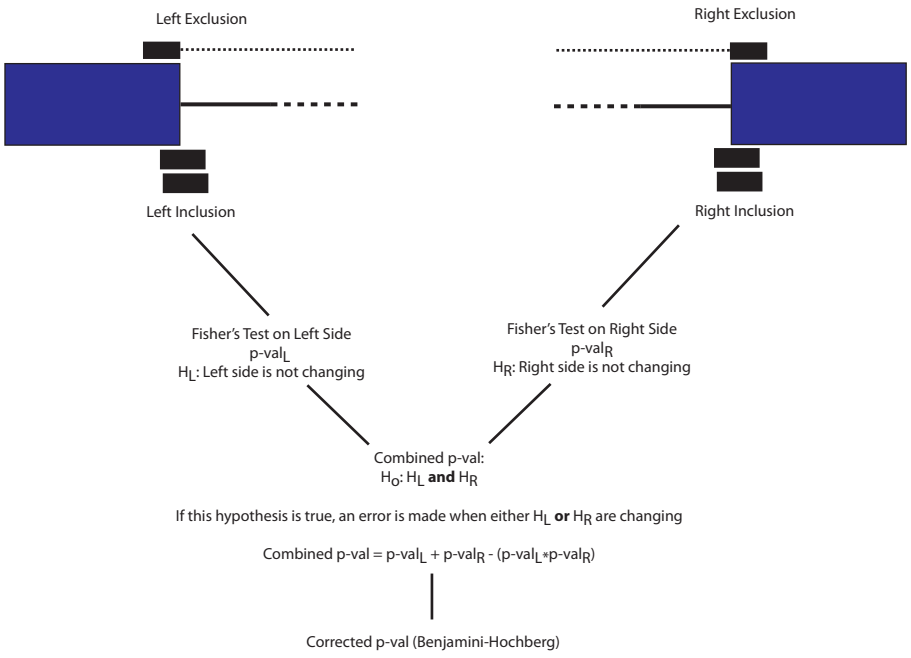


Figure 3.5: **Method used to identify retained intron events.**

that overlap the removed intron (an alternative intron selection) cannot also be removed; therefore, introns that are overlapping are mutually exclusive. Counts from each junction and its mutually exclusive junctions were used from both the untreated and *ps*(RNAi) samples to perform Fisher's exact test to identify significantly changing events. Seventeen junctions that were significantly changing (Benjamini-Hochberg corrected p-value ≤ 0.05), but not part of any classified event, were identified in the *pasilla* study.

3.3.9 Tandem 3' UTRs (alternative polyadenylation)

The analysis to examine poly(A) containing reads was performed and written by Kasper D. Hansen. He attempted to identify tandem 3' UTRs by two methods, but did not find sufficient evidence for a genome-wide analysis of alternative polyadenylation. To detect novel alternative poly(A) cleavage sites, he identified reads that had at least 6 consecutive As (or Ts) at the end (or beginning) of the read, and which thus potentially represent direct sequencing the poly(A) taily. In total, 56,828 reads (before mapping) from all the ps(RNAi) samples contained evidence of a poly(A) tail and 58,636 reads from the untreated samples. These were deemed insufficient for making significant conclusions, as some of these reads may be genomically-encoded adenine stretches while others may not map uniquely; amongst those which do represent true mappable poly-A sites, they will be distributed across all polyadenylated transcripts. In addition, tandem 3' UTR events do not contain exon or junction reads that are unique to the exclusion isoform, which makes it difficult for a reliable quantification of the two isoforms.

3.3.10 Obtaining a non-redundant set of alternative splicing events

Two or more splicing events were considered redundant if they contained overlapping intron coordinates. More specifically, cassette exon, alternative 5' splice site, alternative 3' splice site, coordinate cassette exons, alternative first exon, alternative last exon, and retained intron events were considered redundant if their exclusion (skipping) introns overlapped. Mutually exclusive exons were considered redundant if any intron overlapped with another event. Events within each class of alternative splicing were considered redundant and the event with the lowest p-value was included in the non-redundant set.

3.4 Discussion

The benefit of an exon-centric method to analyze alternative splicing events from RNA-Seq data is the reduction of quantification error due to unclear assignment of reads to multiple full length isoforms. The difficulty with quantifying full length transcripts is due to a limitation of the technology. Even with paired-end reads, it is difficult to sequence a

fragment that contains exons at the beginning and end of a very long transcript since library sizes tend to be moderate (about the size of one exon). Advances in sequencing technology such as longer reads with the same throughput or strobed reads (sequencing more than two segments of a cDNA) will improve full isoform quantification. This will also advance our understanding of splicing regulation because we can ask if multiple alternative splicing events in a transcript are non-independent of each other.

Further advancement to the method of JuncBASE includes better use of paired-end information. Given a relatively narrow insert length distribution of paired-end reads, one can probabilistically assign a read to an inclusion or exclusion isoform, even if the sequenced ends align to exons that are shared between both isoforms. Another method called MISO (Katz et al. 2010) performs this type of quantification on alternative splicing events. The major benefit of JuncBASE is that alternative splicing events are directly inferred from the reads and then quantified, while MISO requires an *a priori* set of alternative splicing events.

Chapter 4

Identifying *trans*-acting splicing regulators, their target exons, and associated RNA maps

4.1 Introduction

To gain a better understanding of splicing regulation it is important to first identify all proteins that are regulating splicing, identify the target splicing events that are affected by each regulator, and determine if the protein is acting to activate or repress splicing. Examining splicing regulation at a more global level has revealed interesting patterns of regulation such as shared and distinct regulated events (Blanchette et al. 2005) as well as RNA maps (Zhang et al. 2008; Yeo et al. 2009; Xue et al. 2009; Llorian et al. 2010; König et al. 2010; Warzecha et al. 2010; Taliaferro et al. 2011; Aznarez et al. 2008). This chapter describes work on the study of 58 splicing regulators in *D. melanogaster*.

The second section shows my analysis of binding site enrichment of the hrp36, hrp38, hrp40, and hrp48 splicing regulators near affected splicing events determined by splice junction microarrays. This work was published in Blanchette et al. 2009. The list of affected events and the SELEX-derived binding motifs for each regulator was determined by Marco Blanchette.

We were particularly interested in exploring the conservation of the splicing code between distantly related organisms. As a first step in this process, we sought to generate an RNA map of Pasilla (PS) (Seshaiah et al. 2001), the D. melanogaster ortholog of Nova1/2. To identify PS-regulated exons, I used RNA-Seq to identify splicing events that changed upon depletion of PS by RNAi. We conclude that the RNA map of PS and Nova1/2 is highly conserved between mammals and insects. [1]

[1]This paragraph and a majority of the section on Pasilla was co-written by Brenton R. Graveley with

The last section describes ongoing analysis of RNAi depletion of 57 proteins, including hrp36, hrp38, hrp40, and hrp48, to measure their effect on alternative splicing using RNA-Seq.

4.2 RNA maps for hrp36, hrp38, hrp40, and hrp48[2]

4.2.1 Results

Knowing the target transcripts and specific splice junctions affected by RNAi of each of the four hnRNP proteins and the RNA binding site motifs from the SELEX experiments (determined by Marco Blanchette), we asked if there was enrichment of SELEX motifs near the affected splice junctions. Here, the affected splice events were grouped according to splicing patterns (cassette exon, competing 5' splice sites and competing 3' splice sites). This analysis looked for the enrichment of preferred binding site scoring matrices within windows adjacent to affected splice events compared to splice events not affected by the hnRNP. Positions enriched with hnRNP binding sites are shown in Figure 4.1. More details of binding enrichment can be found in Supplementary Figure 8 in (Blanchette et al. 2009). These findings suggest a preferential binding site location for a given factor controlling a specific splice event, as seen previously (Aznarez et al. 2008; Blanchette et al. 2005; Castle et al. 2008; Licatalosi et al. 2008; Ule et al. 2006; Zhang et al. 2008). However, none of the enrichments were significant after using a conservative Bonferroni correction for >2,000 comparisons between affected splicing events and controls. Thus, while the target transcripts affected by RNAi against a given hrp protein appear to be enriched in the SELEX motifs for that factor, there does not seem to be a strict fixed location at which a given factor acts on a specific class of splicing events (Figure 4.1).

4.2.2 Discussion

Other studies have shown a complex relationship between the location of splicing control elements and their activities (Goren et al. 2006; Wang et al. 2006) and this lack of a strict spacing of regulatory elements may reflect differences in the mechanisms of action of the *Drosophila* hnRNP A/B family of proteins (Cartegni et al. 2002) and the KH domain splicing factor, Nova-1 (Ule et al. 2006). This variability in the spacing of regulatory elements is also common in transcriptional enhancers and silencers (Levine and Tjian

contributions by Li Yang, Michael O. Duff, Kasper D. Hansen, Sandrine Dudoit, and Steven E. Brenner. The work was published in (Brooks et al. 2011).

[2]Text from this section was co-written by Marco Blanchette and Donald C. Rio with contributions from Steven E. Brenner, Richard Green, Stewart MacArthur, and Michael B. Eisen. Text was taken from excerpts in (Blanchette et al. 2009).

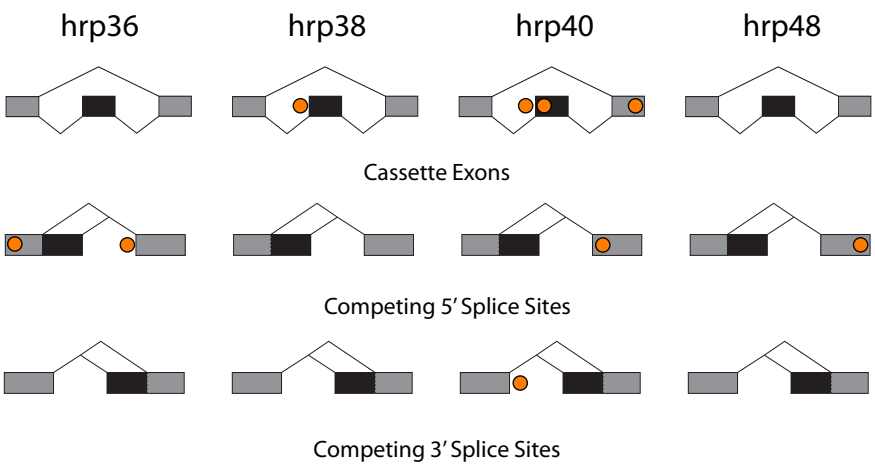


Figure 4.1: **Locations of SELEX-derived binding site enrichment near hrp36, hrp38, hrp40, and hrp48 affected splicing events.** Circles show positions that had an enrichment of SELEX motifs in affected splicing events (two-sided Wilcoxon ranked-sum test, $p < 0.01$ without multiple trials correction for >2000 trials) using any of several measures for enrichment.

2003). My bioinformatic analysis also does not take into account the possible role of weak or non-consensus binding sites, which could be important in the context of cooperative interactions with other splicing factors and might explain the lack of significant enrichment of strong binding sites within fixed windows near affected splice junctions. There were very few cassette exon, competing 5' splice sites, and competing 3' splice sites that were identified to be affected by each of the regulators and this may also contribute to a lack of power to detect an enrichment of binding motifs. By analyzing targets of these four factors using RNA-Seq instead of microarrays, we may observe more affected events and be able to detect a stronger signal.

4.3 Pasilla[3]

4.3.1 Results

Transcriptome analysis of untreated and *ps*(RNAi) S2 cells

To identify regulatory targets of PS, S2-DRSC cells were cultured in biological quadruplicate with the presence of a 444 bp dsRNA fragment corresponding to the *ps* mRNA sequence. Experimental work including RNAi depletion, RNA extraction, and library preparation was done by Li Yang. In parallel, untreated S2-DRSC were cultured in biological triplicate to serve as a control. After 4 days of treatment, total RNA was isolated. Semi-quantitative RT-PCR was used to demonstrate that the *ps* mRNA levels were ~60% lower in the *ps*(RNAi) cells than in the untreated S2 cells. The efficiency of RNAi was also confirmed by the RNA-Seq data as the number of normalized reads (fragments per kilobase of exon model per million mapped reads, FPKM (Trapnell et al. 2010) for *ps* mRNA FPKM were ~4.7-fold lower in the *ps*(RNAi) cells than untreated S2 cells. No other genes containing a KH or RRM domain had a significant change in gene expression. RNA-Seq libraries were prepared from each biological replicate by performing two rounds of poly(A)+ enrichment, RNA fragmentation, random hexamer primed cDNA synthesis, linker ligation, PCR enrichment, and size selection. The libraries were sequenced using both single read and paired-end methodology with read lengths between 37 nt and 45 nt. Most libraries were sequenced in 2-6 separate lanes (technical replicates), with the exception of one *ps*-RNAi sample (supplementary information in Brooks et al. 2011). Paired-end reads had an approximate insert size of 175 ± 50 bp. For consistency, all reads were trimmed to 37 nt from the 3'-end prior to alignment.

Our mapping strategy involved simultaneously aligning the reads to the genome sequence and splice junction sequences. Using our alignment parameters, I observed a high correlation between biological (r2=0.88-0.89) and technical (r2=0.92-1.0) replicates of our samples when comparing the number of reads that aligned to splice junctions (Brooks et al. 2011, Supplemental Figure 5). Thus, I instituted a cutoff of at least three distinct offsets to consider a predicted splice junction as a confident splice junction. This filtering resulted in a final junction dataset of 28,926 confident junctions from the *ps*(RNAi) and control samples. Brenton R. Graveley and Michael O. Duff aligned all single reads to the combined genome and splice junctions using the parameters outlined above. For our paired-end alignments, the Spliced Paired-End Aligner (SPA) software package was used, written by Michael O. Duff, which aims to uniquely map mate-pairs to ensure consistency between mappable reads, particularly when one or both reads align to splice junctions, and to uniquely place the

[3]The following section was co-written with Brenton R. Graveley and with contributions from Li Yang, Michael O. Duff, Kasper D. Hansen, Sandrine Dudoit, and Steven E. Brenner. Text was modified from (Brooks et al. 2011).

pairs in instances where the individual reads could not be mapped uniquely. SPA treats each read of the mate-pair as a single read and is aligned using Bowtie (Langmead et al. 2009) to the combined genome and splice junction sequences and all mapping positions for reads that can be mapped up to 10 possible locations are reported. SPA then processes the Bowtie output to identify mate-pair combinations in which both reads map to the same chromosome, to opposite strands, are oriented towards one another, and are less than 200kb apart (the size of the largest annotated *D. melanogaster* intron).

Out of 47.5 million paired reads, 30.2 million (64%) were uniquely aligned given all three criteria. Importantly, the distribution of the distance between mate-pairs was consistent with the insert size selected during the library preparation (Brooks et al. 2011, Supplemental Figure 2). The remaining reads were further analyzed to include cases where one read could be uniquely aligned, but the other read had no valid alignment. These are most likely instances where the unalignable read had a high error rate. This step "rescued" an additional 9.4 million (20%) reads from the paired-end sequence data that were treated as single reads. In summary, our alignment strategy yielded 115.8 million uniquely mapped 37 nt read sequences from the untreated and *ps*(RNAi) samples, of which ~5% (5.76/115.8 million) map to splice junctions.

Identification of PS-affected splicing events

To identify changes in splicing upon depletion of PS, I used JuncBASE (Junction Based Analysis of Splicing Events), which takes as input genome coordinates of all annotated exons and all confidently identified splice junctions (including annotated and novel junctions) to find sets of exons and junctions that can be classified as one of eight types of alternative splicing events: cassette exons, alternative 5' splice sites, alternative 3' splice sites, mutually exclusive exons, coordinate cassette exons, alternative first exons, alternative last exons, and retained introns. JuncBASE is described in more detail in 3.3 on page 40. After identifying each splicing event, JuncBASE calculates the sum of single reads, mate-pairs, and splice junction reads that align to either the inclusion or exclusion isoforms in both the untreated and *ps*(RNAi) samples, to determine if a shift in splicing has occurred upon depletion of PS. Mate-pairs aligning to an isoform are treated as one event giving evidence for that isoform, instead of considering each mate as an independent read. A Fisher's exact test was performed on 2 x 2 contingency tables comprised of these read counts (inclusion vs. exclusion, untreated vs. *ps*(RNAi)).

From the 2 x 2 tables of counts constructed for each splicing event, I were able to classify 494 splicing events from 323 genes that changed significantly in the *ps*(RNAi)sample (corrected p-val ≤ 0.05; Brooks et al. 2011, Supplemental Figure 7, Supplemental Datasets 1-3). Within each of the eight types of alternative splicing, I identified a non-redundant set of splicing events with no overlapping introns. If two events had an overlapping intron, the event with the lowest p-value was kept. From the non-redundant set, I identified 405 total splicing events affected by *ps*(RNAi) (Figure 4.2).

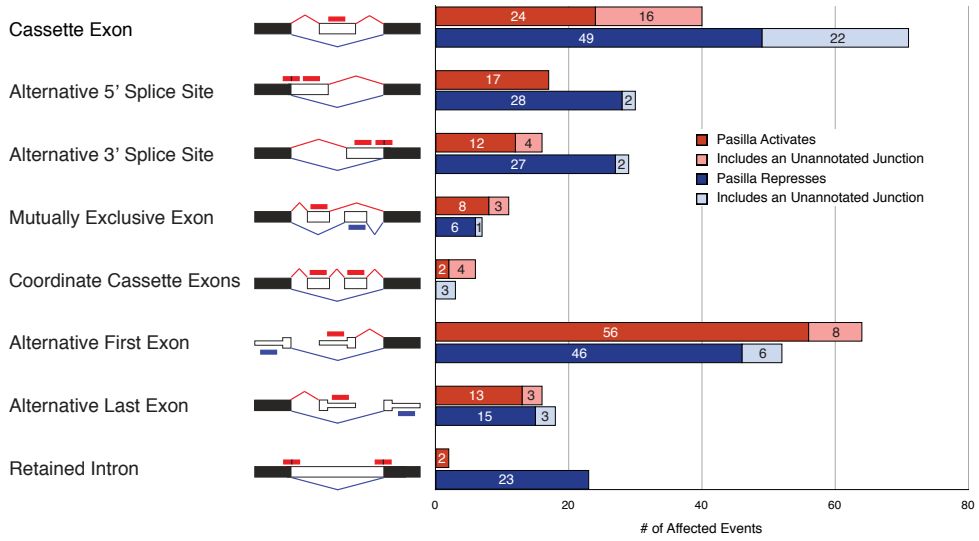


Figure 4.2: **405 PS-regulated pre-mRNA processing events.** (Black boxes) Constitutive regions; (white boxes) alternative regions. (Red lines) Splice junctions for the inclusion isoform; (blue lines) junctions for the exclusion isoform. (Red bars) Exonic reads that support the inclusion isoforms; (blue bars) exonic reads that support the exclusion isoforms; (red bars with black line) reads that support the inclusion isoform, but have shared portion with the exclusion isoform. Thinner portions of the boxes in alternative first exons and alternative last exons correspond to UTRs.

Semi-quantitative RT-PCR experiments (done by Li Yang) validated all 16 tested splicing events I identified by RNA-Seq (Figure 4.3, Additional validations in (Brooks et al. 2011, Supplemental Figure 8). I did not observe a general decrease in gene expression upon RNAi, as noted in previous studies (Blanchette et al. 2005), and moreover, our approach to identify changes in splicing accounts for any change in overall expression (row and column sums of the 2 x 2 contingency tables). Figure 4.3 highlights three examples of PS-regulated splicing events that were identified and validated. The first example involves the adjacent *sesB* and *Ant2* genes (Figure 3A). In untreated cells, the first exon of *sesB* is most frequently spliced to the downstream constitutive *sesB* exon (top isoform). However, in *ps*(RNAi) cells, splicing is strongly switched to favor the expression of two different isoforms. The first involves splicing of the first exon of *sesB* to the first exon of *Ant2* (bottom isoform). The second isoform involves an increased inclusion of an alternative cassette exon (green) in *sesB*. A second example of a PS-regulated splicing event involves a cassette exon in the *bmm* gene (Figure 4.3.B). In this case, the exon is included in ~50% of the transcripts in untreated cells as determined by RNA-Seq, but nearly constitutively included in the *ps*(RNAi) cells. The final example involves the *dre4* gene where a cassette exon is activated by PS and therefore skipped more frequently in the *ps*(RNAi) cells than in the untreated cells (Figure 4.3.C).

The analysis method is quite sensitive, as I was able to identify PS-affected splicing events even when an isoform is normally expressed at a low level. For example, in untreated cells there were 865 reads supporting the inclusion of a validated PS-affected cassette exon in the *cg* gene but only 9 reads supporting the exclusion of the exon (Brooks et al. 2011, Supplemental Figure 8B). However, in *ps*(RNAi) cells, the exon is constitutively included as there are 831 reads supporting inclusion of the exons and no reads supporting exclusion. These results indicate that this exon is normally repressed by PS at a very low level. The change in inclusion of this exon in the *ps*(RNAi) cells is observed in the RT-PCR validation experiment (Brooks et al. 2011, Supplemental Figure 8B). While splicing events such as this can be detected when measuring both exon and splice-junction reads, they can be missed if only exon reads are considered.

Interestingly, while most (327 of 405) of the PS-affected splicing events contained entirely annotated junctions, 19% (77 of 405) of the affected splicing events involved unannotated splice junctions. Strikingly, 90% (69 of 77) of these affected, unannotated splice junctions were expressed in untreated S2-DRSC cells while only 10% (8 of 77) of the unannotated splice junctions are exclusively expressed in the *ps*(RNAi) cells(Figure 4.2). An example involving these unannotated splice junctions is found in the *trol* gene in which a group of 9 contiguous exons, which are annotated as being constitutive, are coordinately skipped in untreated cells (with 52 reads spanning the skipping junction and only 136 reads total for all 9 exons and inclusion junctions connecting these exons) but coordinately included in the *ps*(RNAi) cells (where no reads support skipping of these exons but 3,573 reads support inclusion of the 9 exons) (Brooks et al. 2011, Supplemental Figure 8A).

I next analyzed several general features of the set of PS affected splicing events. I find

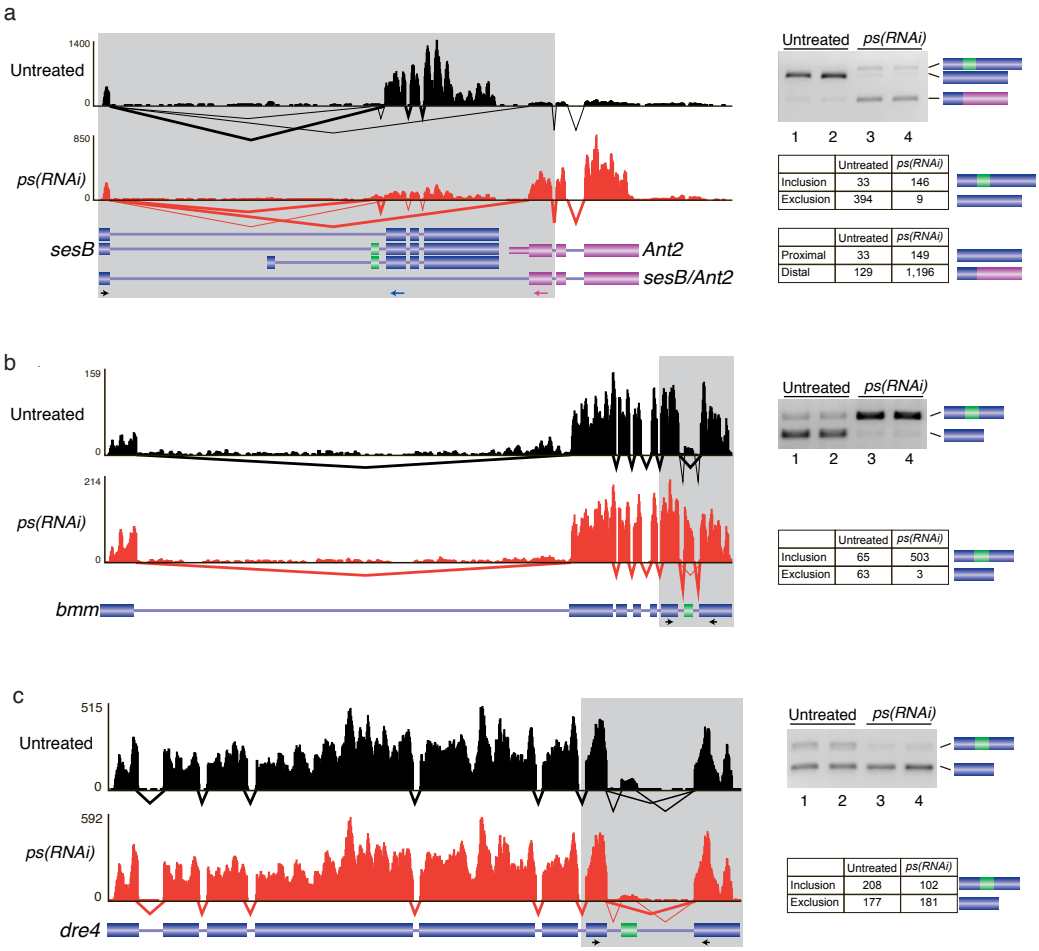


Figure 4.3: **Examples of validated PS-regulated splicing events.** All validations were done by Li Yang. Figure created by Li Yang and Brenton R. Graveley. Alternative splicing events in the *sesB/Ant2* (A) *bmm* (B) and *dre4* (C) genes were identified from the RNA-seq data and validated by RT-PCR. In each case, the RNA-seq coverage and splicing patterns for both the untreated (black) and *ps*(RNAi) (red) samples are shown along with the annotated transcript models. The RT-PCR validation assays were performed in biological duplicates for both the untreated (lanes 1,2) and *ps*(RNAi) samples (lanes 3,4). The number of read counts supporting each splicing event in each sample is indicated in the tables on the right.

that PS predominantly functions as a splicing repressor, as a majority (60%) of the affected splicing events I identified were repressed by PS. Of the splicing events that only involve splicing (and not differential promoter or poly(A) site use) PS regulated the greatest number of cassette exons (111), roughly equal numbers of alternative 5' or 3' splice sites (47 and 45, respectively), and relatively few intron retention events (25), mutually exclusive exons (18) and coordinately regulated cassette exons (9). However, when considering what fraction of expressed alternative splicing events are affected by PS, the greatest fraction of mutually exclusive splicing events (62%) were affected, and between 10% and 20% of cassette, alternative 5' and 3' splice sites, and only 1.5% of intron retention events (Figure 4.4). Perhaps surprisingly, I find that PS affected a significant number (116) of alternative first exon events and a smaller number (34) of alternative last exon events. However, it is unclear if these events are changing due to a direct or secondary effect of PS on the coordination of splicing with either transcription or polyadenylation.

A conserved Pasilla/Nova-RNA map

Mammalian Nova1/2 is known to bind directly to sequences that match the consensus motif YCAY and has the highest affinity for YCAY repeats (Jensen et al. 2000). The amino acids of Nova1/2 that contact RNA in a sequence-specific manner (Lewis et al. 1999; Jensen et al. 2000) are conserved in PS, suggesting that PS also recognizes YCAY motifs. Consistent with this, biochemical experiments done by Jung Park confirm that recombinant PS can bind to YCAY containing RNA (Figure 4.5).

To investigate potential PS binding sites in silico, I identified overrepresented hexamers in conserved sequences within and 150 nt upstream and downstream of the set of affected cassette exons compared to a set of unaffected cassette exons (corrected p-val ≥ 0.95). The top five hexamers, in order, were CACCAC, CCACCA, CAACAA, AACAAC, ACAACA (p-value < 2.0e-4, not correcting for 4,096 tested hexamers). Consistent with the known RNA binding preference of Nova1/2 and PS, the top two hexamers contain a YCAY sequence. Nova1/2 has previously been shown to preferentially repress downstream splice sites and activate upstream splice sites (Ule et al. 2006; Licatalosi et al. 2008). We were interested in determining whether a similar RNA map exists for PS. Using the binding model introduced for Nova targets (Ule et al. 2006) (see Methods), I tested for the enrichment of conserved clusters of YCAY motifs in 45 nt sliding windows across introns and exons of PS-activated and PS-repressed splicing events as well as unaffected splicing events to serve as a control (Figure 4.6, Brooks et al. 2011, Supplemental Figure 11); which yielded 452 tests across cassette exons, alternative 5' splice sites, alternative 3' splice sites, alternative first exons, alternative last exons, and retained introns. I observed that positions upstream and within PS-repressed cassette exons had a higher average conserved YCAY cluster score than the control exons; however, only one position within the alternative exon had an overrepresentation of YCAY clusters with an uncorrected p-val < 0.01 for 452 tests. I observe that positions downstream of PS-activated cassette exons had an

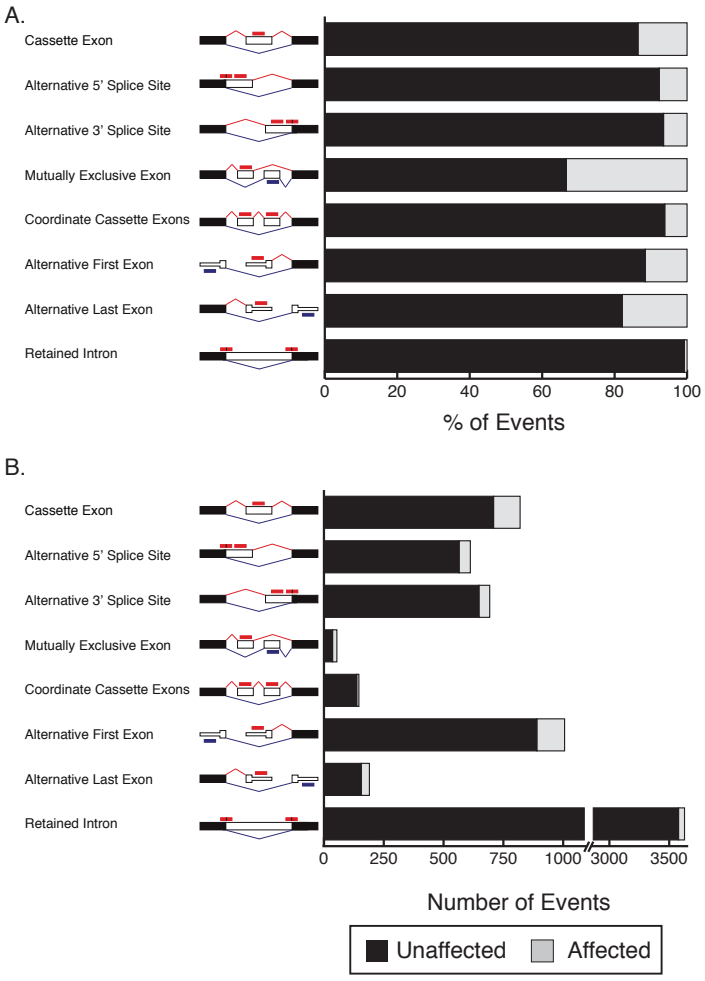


Figure 4.4: **Types of splicing events affected by Pasilla.** (A) The percentage of each class of expressed splicing event that are unaffected and affected in the *ps*(RNAi) samples. (B) The number of splicing events of each class that are unaffected in the *ps*(RNAi) sample.

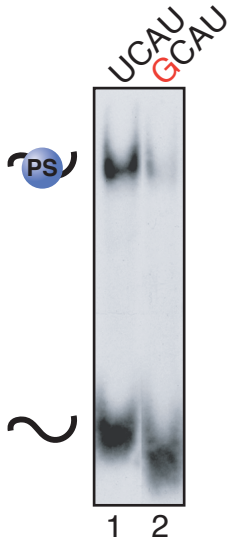


Figure 4.5: **Pasilla binds to YCAY containing RNA.** This experiment was done by Jung W. Park. Figure created by Jung W. Park and Brenton R. Graveley. RNAs containing three repeast of UCAU or GCAU were incubated in the presence of recombinant PS and the reactions resolved on a non-denaturing polyacrylamide gel. The locations of the unbound RNA and the RNA-protein complex are indicated.

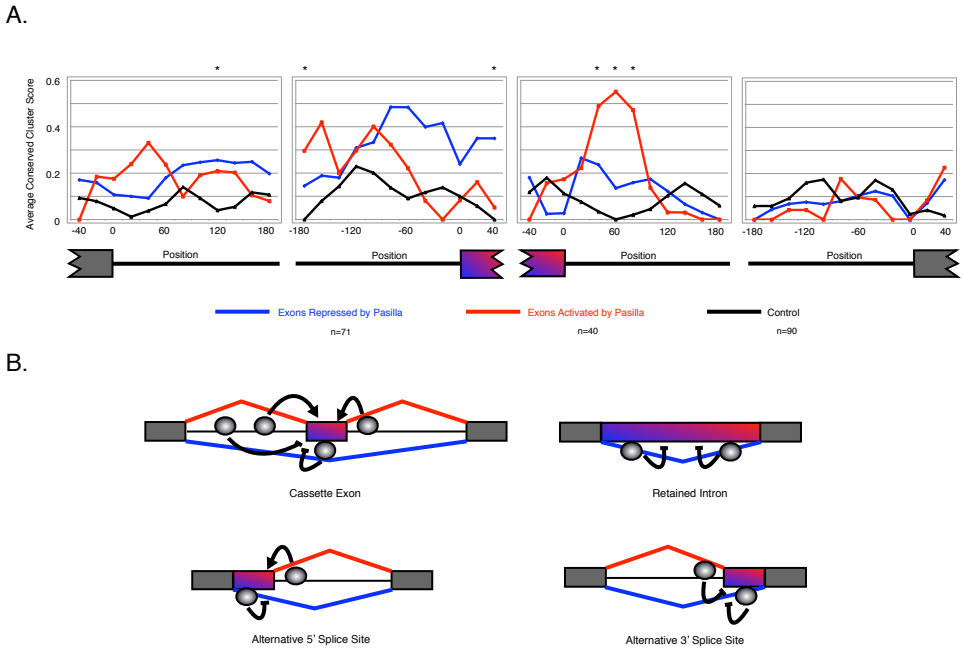


Figure 4.6: **A Pasilla RNA-map.** (A) Each position in the graph represents the average conserved YCAY cluster score, within a centered sequence window of 45nt. The conserved YCAY cluster score was calculated for cassette exons that are activated by PS, repressed by PS, and unaffected cassette exons (Fisher's exact test, Benjamini-Hochberg adjusted p-value ≥ 0.95). Only regions adjacent to introns >400nt were used for scoring. Positions with enriched YCAY cluster scores are indicated by an asterisk (Wilcoxon-rank sum test, uncorrected p-value < 0.01). (B) Positions near cassette exon events, alternative 5' splice site events, alternative 3' splice site events, and retained intron events with an enrichment of YCAY clusters. Gray spheres indicate the relative positions containing enriched binding sites. Detailed plots of average conserved YCAY cluster scores are shown in Supplemental Figure 9 of Brooks et al. 2011.

overrepresentation of conserved YCAY clusters, with a peak enrichment 60nt downstream of the cassette exon (uncorrected p-val < 0.01). I also find conserved YCAY clusters further upstream of PS-activated exons and conserved YCAY clusters near upstream constitutive exons of PS-repressed exons (uncorrected p-val <0.01).

The locations of these enriched YCAY motifs near PS-regulated cassette exons are analogous to the locations of the YCAY motifs near Nova-regulated exons (Ule et al. 2006). The conservation of the regulatory "map" is not due to the conservation of Nova and PS target genes. Out of 47 mouse genes that were identified as targets of Nova (Ule et al. 2006), 33 had at least one *Drosophila* ortholog. Of the 33 *Drosophila* orthologs, 23 were expressed in the S2 cell line. Amongst these 23 Nova target genes with an S2-expressed *Drosophila* ortholog, only 4 were also a target of PS. In addition to the regulatory map for cassette exons, I created maps for other alternative splicing events. There is an enrichment of YCAY motifs near PS-regulated alternative 5' splice sites (Bonferroni-corrected p-val < 0.05) and 3' splice sites (uncorrected p-val < 0.01) and the enrichment near alternative 5' splice sites is significant given multiple-testing correction for 452 tests (Figure 4.6 and Supplemental Figure 11 from Brooks et al. 2011). The alternative 5' and 3' splice site map of PS is consistent with the asymmetric action of Nova that was observed in the validated targets of Nova (Ule et al. 2006). I identified 23 introns that were retained significantly more often in the absence of PS. Positions adjacent to both the 5' and 3' splice sites of these introns had a significant enrichment of conserved YCAY clusters (Bonferroni-corrected p-val < 0.05). Therefore, the effect of PS on alternative splicing is not only dependent on its position with respect to a splice site, but also depends on the context of the type of alternative splicing event. However, a common pattern observed in cassette exon events, alternative 5' splice sites, alternative 3' splice sites, and retained intron events, was the presence of YCAY binding sites within the alternative exon (or portion of an exon) that is normally repressed in the presence of PS (Brooks et al. 2011, Supplemental Figure 11). Finally, the Nova RNA map contains an enrichment of YCAY motifs near the proximal poly(A) site in cases of alternative polyadenlyation events regulated by Nova. While I observe an enrichment of YCAY motifs near PS-regulated alternative last exons in an analogous position, the enrichment is not statistically significant.

Although the general RNA regulatory map of Nova and PS are quite similar, the scoring method used to predict Nova exon targets was insufficient to distinguish PS-regulated exons from control exons or the direction of regulation (Figure 4.7). Perhaps, this is due to the lack of a significant silencer region directly upstream of the alternative exon (NISS2) and the lack of a significant enhancer region further downstream of the alternative exon (NISE3) that were identified near Nova-regulated exons.

4.3.2 Discussion

I have identified 405 splicing events that are affected by RNAi depletion of PS. Interestingly, I found that 19% of the PS-regulated alternative splicing events contain an

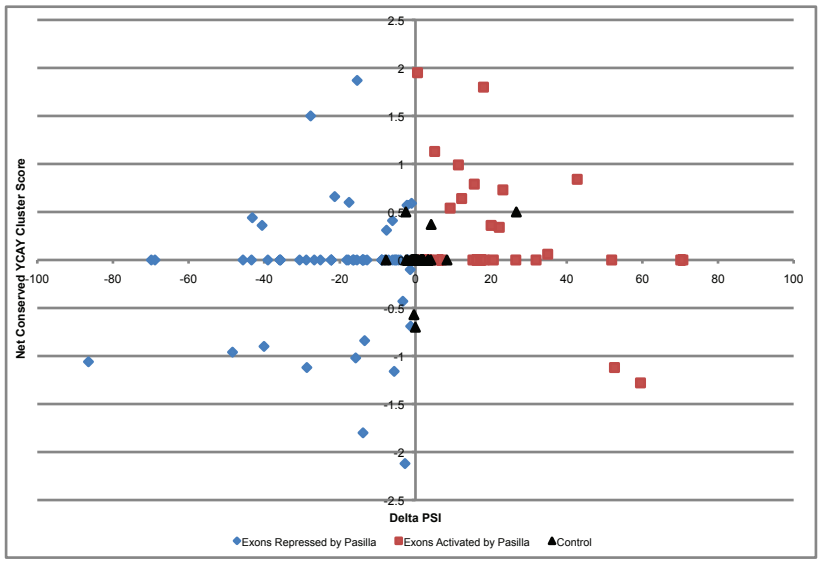


Figure 4.7: **YCAY cluster score versus change in ΔΨ.** Net conserved YCAY cluster score versus change in percent spliced in (ΔΨ) for all cassette exon events that were repressed by Pasilla (n=71) or activated by Pasilla (n=40). The control exons (n=90) are cassette exons that had no change in splicing (Benjamini-Hochberg corrected p-val > 0.95)

unannotated splice junction. The large proportion of unannotated splicing events in the affected set of exons demonstrates the benefits of using RNA-Seq as a method to detect regulated alternative splicing events over other methods such as microarrays, which rely on pre-defined splicing events to probe. Previous studies using splice junction microarrays to identify alternative splicing events in *Drosophila* found that 29-35% of their affected junctions were annotated as constitutive (Blanchette et al. 2005, 2009), which suggests that many of these junctions were participating in unannotated alternative splicing events. This work only looked at potential splice junctions formed between annotated splice sites; therefore, future work could identify even more cases of alternative splice junctions if unannotated splice sites are used. The large number of unannotated splice junctions I identified also indicates that the *D. melanogaster* transcript annotations remain incomplete. More importantly, the fact that 90% of these unannotated splice junctions are observed in untreated S2-DRSC cells, and are not specific to the *ps*(RNAi) cells, indicates that these junctions are normally expressed and are not aberrant splicing events induced by depleting a splicing regulatory protein.

By calculating conserved YCAY cluster scores across affected cassette exons, alternative 5' splice sites, and alternative 3' splice sites that were activated or repressed by PS, I found that the PS RNA map recapitulates the major features of the Nova RNA map. Both Nova1/2 and the *Drosophila* ortholog, PS, appear to activate splicing of upstream exons and repress splicing of downstream exons or exons they directly bind to; however, further

studies profiling transcriptome-wide PS binding are necessary to confirm the direct targets, as was confirmed with Nova2 (Licatalosi et al. 2008). Ule et al. proposes a molecular mechanism for the action of Nova (Ule et al. 2006); however, further work will need to be performed to determine if the same molecular mechanism is used by PS. The PS RNA map is not similar to patterns of regulation identified for non-homologous proteins such as PTB (Xue et al. 2009) and HNRNPC (König et al. 2010), further supporting an ancestraly conserved mechanism of splicing regulation for the orthologous proteins Nova and PS.

Although the PS and Nova RNA maps are similar to the mammalian Fox-1/2 map, a recent study suggests that these maps are in part similar due to the proteins combinatorial affects on a subset of alternative splicing events (Zhang et al. 2010). The four shared target genes of Nova and PS (out of 23 Nova target genes with at least one expressed *Drosophila* ortholog) are *cac*(*CG1522*), *cora*(*CG11949*), *msn*(*CG16973*), and *caki*(*CG6703*). According to their GO annotations, *cac* has voltage-gated calcium channel activity and *caki* has calmodulin-dependent protein kinase activity and is also involved with cell adhesion. Both *cac* and *caki* have GO annotations involved with adult locomotory behavior, which suggests a possible ancestral role for a neurological function of PS and Nova. Although there are a few shared target genes, the RNA-maps generated from the targets of Nova and from the targets of PS are based on almost entirely distinct sets of genes. This further supports the role of *cis*-acting binding sites in driving the evolution of alternative splicing, a result seen from a previous study examining orthologous Nova-regulated exons in vertebrates (Jelen et al. 2007). These results suggest a general evolutionary model common to alternative splicing, transcription regulation, and miRNAs (Meireles-Filho and Stark 2009; Shomron et al. 2009). In general it appears that targeting molecules retain their sequence specificity over long periods of evolutionary time, and changes in the regulation, including large-scale changes in targets, occur through gains and losses of *cis*-acting binding sites.

While there was little overlap in Nova and PS targets based on our studies in S2 cells, we cannot assume that Nova and PS do not have shared targets in other contexts. Our study was performed in cell culture while the Nova studies were performed in mouse brain. In the context of a different cell line or in *ps* mutant flies, we might observe an increase in overlapping genes; however, a previous study of *ps* mutants (Seshaiah et al. 2001) suggests that PS may have a different physiological function than Nova. Homozygous *ps* mutants showed strong defects in salivary gland development, but neurological defects were not observed (Seshaiah et al. 2001). Nonetheless, a GO analysis of the set of PS-affected genes, using Funcassociate (Berriz et al. 2009), identified overrepresented terms corresponding to neuronal functions (regulation of neurogenesis, locomotion, regulation of axonogeneis, chemosensory behavior, etc.) as well as sexual reproduction (anatomical structure morphogenesis, gamete generation, oogenesis, etc) and the cytoskeleton (cytoskeletal protein binding, actin binding, cytoskeletal protein binding, etc.) (adjusted p-value < 0.01). While the expression pattern of PS and the *ps* mutant phenotype suggests that PS has an important role in salivary gland development, our results suggest that PS may also have

additional roles not only in sexual reproduction and cytoskeleton dynamics but also in neural function, like Nova1/2. Thus, despite little overlap in the regulatory targets, the regulatory mechanisms and physiological functions of orthologous splicing regulators may be conserved.

In conclusion, I have identified and classified hundreds of alternative splicing events that are affected by one splicing regulator and have shown that the overall RNA map relating the position of binding sites for the factor to its affect on splicing is conserved from an insect to mammals. Future studies that deplete other splicing regulators in *Drosophila* may also reveal regulatory maps that can perhaps be applied to mammalian splicing regulators as well.

4.4 Regulatory targets of 57 proteins

4.4.1 Results

Identification of known or putative splicing regulators

We were interested in identifying targets of known splicing regulators, but also in finding out if other RNA binding proteins have global effects on splicing, if they were previously unknown to. We were particularly interested in proteins that contained an RRM or KH domain, which are thought to bind to RNA in a sequence specific manner. I created a full set of putative splicing regulators to examine based on literature searches or previous knowledge of function (Barbosa-Morais et al. 2006; Park et al. 2004; Blanchette et al. 2005, 2009), adding on proteins that contained an RRM or KH domain. RNAi was found to be most efficient in S2 cells (data not shown) and therefore, I only examined proteins that were shown to be expressed in S2 cells (see Methods). This narrowed down the list to 181 potential proteins to examine. I prioritized the proteins from this list based on previous knowledge of their function, homology to know regulators, or if the protein had a common name, indicating some knowledge of function from either homology or phenotype. Our group has examined the top 58 proteins from the 181 potential proteins. One of the 58 proteins listed, *ps (CG8144)*, was examined in a pilot study and results are presented in chapter and section 4.3 on page 53.

Based on homology or published literature, each of the selected proteins were categorized as an SR protein, hnRNP, part of the core spliceosome, part of the exon junction complex or involved in nonsense-mediated decay (EJC-NMD), having evidence of splicing regulation, or having no prior evidence of splicing regulation (Table 4.1).

CG number	Synonyms	RRM or KH?	Category	Aligned Reads
CG10203	xl6, 2/30, 9G8, Dxl6, RBP1-like, TN166	r	SR	27,059,564
CG10851	B52, SRp55, E(Dfd), E726, MabB52, RRM8, Rbp8	r	SR	40,726,336
CG17136	Rbp1, RRM1	r	SR	21,199,162
CG1987	Rbp1-like, i159	r	SR	23,424,420
CG4602	Srp54, p75	r	SR	34,641,388
CG5442	SC35	r	SR	40,686,826
CG5655	Rsf1, ROX21	r	SR	57,170,053
CG13425	ROK, bl, Hrb57A, Q18	k	hnRNP	48,385,744
CG10377	Hrb27c, 10280, Hrb48, Hrp48, RRM7, Rbp7, linha, p50	r	hnRNP	18,352,508
CG12749	RO87F, Hrb87F, hrp36, hrb2, hrb85CD, Hrb87F, Hrb87Fa, P11, Q14, Q16, hrp40, p38	r	hnRNP	21,184,672
CG16901	RO40, sqd, CG17791, Hrp40, Hrb87, Hrb87Fb, RRM3, hnRNP 1	r	hnRNP	22,104,329
CG17838	ROQR, Al945337	r	hnRNP	22,507,820
CG31000	heph, CG2094, CG2290, PTB, ema	r	hnRNP	21,233,917
CG5099	MUSa, msi, DMSIDNA, SIDNA	r	hnRNP	20,776,134
CG6946	ROFH, glo	r	hnRNP	24,044,390
CG7437	PCB, mub	r	hnRNP	25,561,241
CG9373	ROM	r	hnRNP	58,735,925
CG9983	ROA1, Hrb98DE, hrp38, Pen9, Hrb1, p9	r	hnRNP	21,736,670
CG18426	RY1, ytr, b(2)gcn, bgcn		hnRNP	29,467,916
CG30122	ROU, CG18610, CG5477, LD11002		hnRNP	33,269,298
CG5836	SF1, p70	k	Core	23,983,907
CG3312	Rnp4F, 4F-rnp, rnp-4f	r	Core	35,175,798
CG8749	snRNP70K, 70K, U1 70K, U1 70K snRNP, U1 snRNP, U1-70K, snRNP, snRNP27D	r	Core	27,180,736
CG10279	Rm62, 4136, DmRH8, Dmp68, Lip, p68		Core	35,649,044
CG1646	Prp39		Core	39,850,917
CG6227	Prp5, DmRH29		Core	22,443,206
CG6841	Prp6		Core	33,743,278
CG8241	Prp22, pea		Core	25,713,056
CG1101	Aly, REF1	r	EJC	26,267,072
CG16788	RnpS1, DEK	r	EJC	22,371,356
CG8781	tsu, Y14	r	EJC	34,856,806
CG1559	Upf1		EJC	27,604,682
CG3584	qkr58E-3, KEP1	k	Prior	27,874,478
CG7878	DmRH26	k	Prior	52,621,232
CG8144	ps, CG16765, CG16776, CG18509	k	Prior	57,837,992
CG8912	Psi	k	Prior	30,996,531
CG10128	tra2	r	Prior	40,122,191
CG10328	nonA-l, Z, LD09360	r	Prior	22,752,024
CG11266	CAPER, HCC1	r	Prior	25,766,498
CG18350	Sxl, CG33070, Fl, Mex156, Su(da), Sx1, Sxl-2	r	Prior	42,139,744
CG4262	elav, 44C11, 9F8A9, EC7, Elav-9F8A9, elav-1, elav-2, elav-3, fliJ	r	Prior	41,447,570
CG5422	Rox8, DmTIAR, RRM12	r	Prior	42,685,633
CG6049	Tat-SF1	r	Prior	24,836,500
CG9412	rin, nts, rasp	r	Prior	35,309,209
CG7971	SR300		Prior	20,510,698
CG8019	hay, DhR25, DhXPB, DmXpB, ERCC3, TFIIH, XPB, hwr, i50, nc2		Prior	30,554,615
CG3249	PKAAP	k	No Prior	46,118,222
CG33106	mask, CG18671, CG31138, CG6268, CG6313	k	No Prior	30,318,389
CG4816	qkr54B, SAM, SAM50, p62GAP	k	No Prior	20,176,326
CG5170	Dp1, DDP1	k	No Prior	27,102,731
CG5821	qkr58E-2	k	No Prior	29,973,777
CG6203	Fmr1, AT24755, EP(3)3517, FMRP, FXR, dFMR, dFXR1, fmr	k	No Prior	30,216,799
CG6779	RpS3	k	No Prior	37,568,498
CG31716	NOT4, CG5244, CG5251	r	No Prior	31,577,806
CG32423	alan-shepard, CG10647, CG10649, CG10668	r	No Prior	31,632,015
CG4760	bol, Boule, CG4727	r	No Prior	35,380,355
CG4878	eIF3-S9	r	No Prior	28,001,239
CG8636	eIF3-S4	r	No Prior	37,325,780

Table 4.1: **Putative splicing regulators to examine in S2 cells.** The 58 proteins listed were sequenced and the number of reads aligned are indicated. See Table 1.2 on page 9 for more details. Categorization is shown; however all categories are not mutually exclusive. Core; part of the core spliceosome. EJC; exon junction complex or nonsense-mediated decay. Prior, No Prior; Prior or no prior evidence for a role in splicing regulation.

Transcriptome analysis of cells individually depleted of 57 proteins

RNAi depletion of each splicing regulator gene was done individually, in biological duplicate. Some biological replicates, but not all, were sequenced in technical replicate in order to obtain at least 20 million reads for the combined biological replicates. The total amout sequenced for RNAi of CG10377 did not meet this threshold. The efficiency of the knockdown was verified using RT-PCR before creating libraries for sequencing RNA. Each of the 57 samples, along with untreated cells, were sequenced using single-end reads of 75-80 bp. RNAi depletions, RNA extraction, RT-PCR validations, and sequencing libraries were done by Li Yang. Additional RT-PCR validations and sequencing libraries were done by Gemma E. May. All reads were trimmed to 75 bp for consistency.

Brenton R. Graveley aligned the reads to the genome and a set of annotated (MB8, modencode.org) and novel splice junction sequences using Bowtie (Langmead et al. 2009). Alignment criteria was identical to single read mapping performed in the *pasilla* study (section 4.3 on page 53). Splice junction sequences were generated by me. Novel splice junctions included all novel combinations of exon-exon junctions within the same gene and different genes, but within 2kb. Additional novel junctions were derived from an annotated splice site and an unannotated splice site (GT or AG dinucleotide) within 2kb. When combining all alignments to all samples, only junctions with a Shannon entropy score cutoff of 3 were considered confidently present and thus used in subsequent analyses (Shannon entropy measure described in 3.2.2 on page 38). We obtained 18M-59M alignments for each sample (Table 4.1).

Junction and genome alignments were run through Cufflinks (Trapnell et al. 2010) by Brenton R. Graveley and Michael O. Duff to identify unannoated internal exons that may be present in our data, particularly since we are depleting splicing regulators. Using a reference annotation of all exons identified by a large study of the fly developmental transcriptome (Graveley et al. 2011), I identified 6,312 unannotated internal exons from the Cufflinks novel transcripts, 23 of which are completely non-overlapping with any previously annotated exon.

Read alignments for each sample, the annotated transcripts, and the novel internal exons determined by Cufflinks were used as input to JuncBASE (detailed in chapter and section 3.3 on page 40) to identify and quantify eight types of alternative splicing events in each of the 58 samples (57 proteins + untreated). Because each sample was not sequenced at the same depth, I normalized the inclusion and exclusion counts of each event by the the number of reads aligned relative to the sample with the fewest number of aligned reads. For example, the sample with the fewest aligned reads is CG10377 with 18,352,508 alignments and the sample with the most number of aligned reads is CG9373 with 58,735,925 reads; therefore, all counts in the CG9373 sample are divided by 3.2004 (58,735,925/18,352,508 = 3.2004) to normalize for the increase in the number of aligned reads in that sample.

An advantage of examining the effects of a large number of proteins on alternative splicing is that we can also normalize for the effects that RNAi itself may have on splicing. For

each identified alternative splicing event, I used the median total expression across the 57 RNAi samples and the median "percent spliced in" (Ψ) value to obtain a median exclusion and inclusion counts for the event. For example, if the median total count of an event is 100 and the median Ψ is 25%, I consider the median inclusion read count to be 25 and the median exclusion read count to be 75. If a sample had zero counts for both isoforms, the event was considered to be not expressed in that sample and was not used to calculate the median reference.

To determine if there was a significant change in splicing upon knockdown of a regulator, a Fisher's exact test was performed on a 2x2 table consisting of the exclusion and inclusion counts from the RNAi sample and the exclusion and inclusion counts from the median reference. In cases where an alternative splicing event had more than two isoforms, every isoform was treated as an inclusion isoform, while the sum of all other isoforms was treated as the exclusion isoform, in separate 2x2 tables. After a Benjamini-Hochberg multiple testing correction for the number of events and number of samples tested, I identified 1,784 splicing events that were affected by the knockdown of at least one protein. Details of this approach are described in 3.3 on page 40. I further considered splicing events with a $\Delta\Psi$ change >10% to focus on splicing events that are more likely to have a greater biological effect (n=1,315). The most significant of any overlapping, and therefore redundant, splicing event was retained, which yielded a set of 879 events affected by knockdown of at least one protein. (Figure 4.8; see Methods).

As evidence that the median reference is an appropriate control, I compared the median value to the untreated sample and found only three alternative first exon events that were significantly different. For comparison, xl6 had the least number of affected events in an RNAi sample (seven), affecting alternative first exon events, a cassette exon event, and an alternative 3' splice site.

As expected, the proteins that affected the most splicing events are considered part of the core spliceosome (Figure 4.8): Rm62, pea, and snRNP70K. Also consistent with their previously known function, a larger proportion of the affected events are retained introns which indicates a general defect in regulating intron removal (Figure 4.8). Interestingly, for all proteins, a large proportion (16-83%) of all affected alternative splicing events are alternative first exons—RNA processing events that depend on a change in transcriptional promoter usage. Considering all junctions and exons annotated or observed in the RNAi data, the type of alternative splicing event that is most prevalent is coordinate cassette exons ("Total AS" in Figure 4.8), even though only a small fraction of coordinate cassette exons are differentially spliced when depleting regulators. The prevalence of the coordinate cassette exons may be driven by many low-abundant junctions observed in the RNA-Seq data that skip multiple exons, as we do not observe a large proportion of these events in the transcriptome of developing flies (5 on page 82). Further work is required to investigate these skipping junctions.

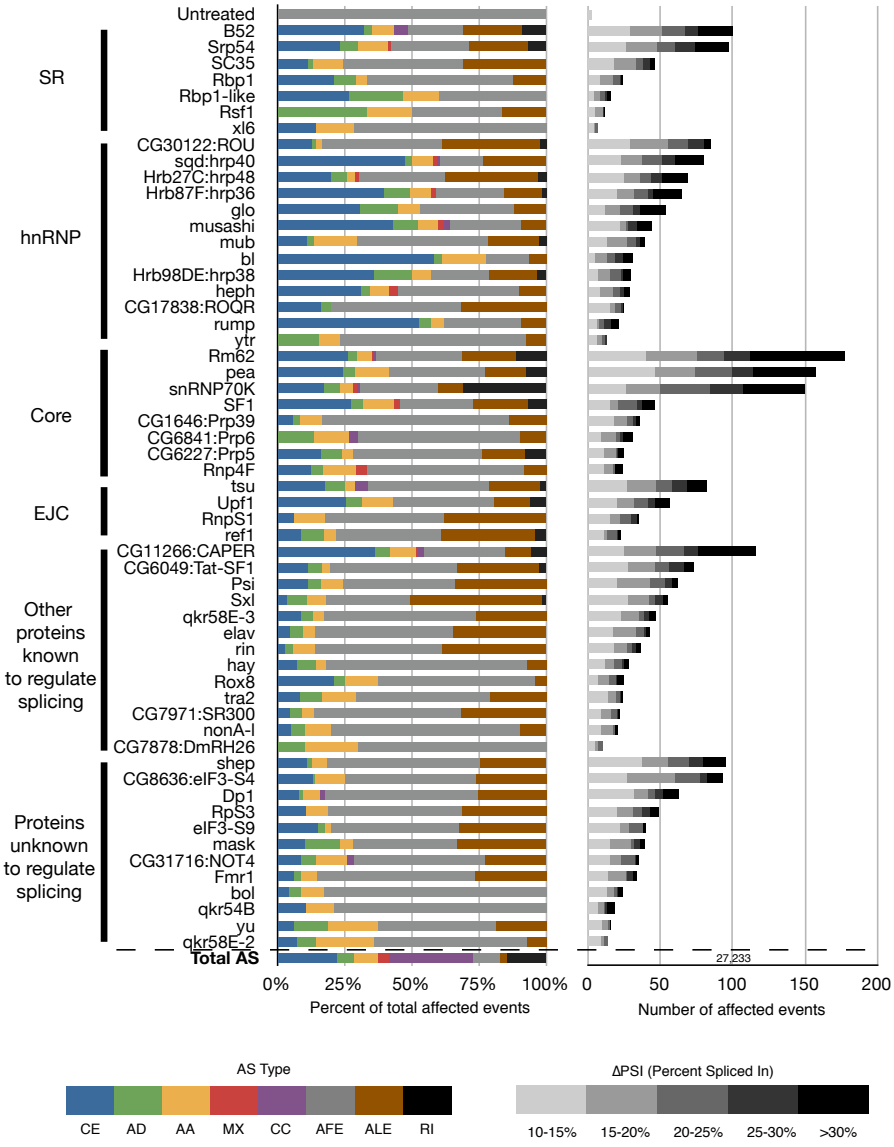


Figure 4.8: **Events affected by depletion of 57 proteins.** The number of events affected by depletion of each protein is shown along with the magnitude of the ΔPSI change relative to the median reference counts. From the set of all affected events, the proportion of each type of alternative splicing (AS) event is shown. CE, cassette exon. AD, alternative donor (alternative 5' splice site). AA, alternative acceptor (alternative 3' splice site). MX, mutually exclusive exons. CC, coordinate cassette exons. AFE, alternative first exons. ALE, alternative last exons. RI, retained intron.

Depletion of proteins have specific effects on splicing

I next examined the extent to which affected splicing events are shared or specific to each protein regulator. In Figure 4.9, I show that the majority of the splicing events are only affected by one factor. The alternative first and last exon events tend to be affected by multiple proteins compared to the other types of events; however, $\Delta\Psi$ values tend to be low and represent more subtle effects. There is a bias in the experiment to observe events that are dependent on only one regulator since we are only performing a single knockdown. Events that depend more strongly on multiple factors may only be observed if we perturb multiple factors at once.

4.4.2 Discussion

As this is an ongoing study, there are many remaining questions and analyses. We are interested to see cases in which these proteins are activating or repressing splicing and if there are events co-regulated by proteins in correlated or opposing directions. Although it is unclear if we will be able to determine why we observe so many affected alternative first exon events upon depleting splicing reguators, the prevalence indicates that these RNA binding proteins may have a larger role in RNA processing events other than splicing regulation. Studies have shown a 5' bias in intron position relative to the total gene length for eukaryotes with very few introns (Mourier and Jeffares 2003). Perhaps, there is an ancient role for these RNA binding proteins in creating a feedback mechanism to the transcriptional machinery that is coupled with splicing of a more 5' intron.

I intend to perform motif analysis on the set of affected events. In addition to looking for enriched sequence motifs near the affected events, I will also look for position specific enrichment near regulated exons to identify RNA maps.

4.5 Methods[4]

Identifying *D. melanogaster* proteins with an RRM or KH domain

All *D. melanogaster* protein sequences were obtained from Uniprot (UniProt Consortium 2011; Jain et al. 2009). Each sequence was searched against Pfam database (Finn et al. 2010) using hmmpfam (hmmer.org) for the presence of the Pfam domains RRM_1, RRM_2, RRM_3, KH_1, KH_2 with the default cutoff score. To find additional proteins that may have been missed by Pfam, each sequence was also compared against SMART (Letunic et al. 2009) domains RRM, RRM_1, and KH and against the Prosite (Hulo et al.

[4]Modified excerpts from (Blanchette et al. 2009) and (Brooks et al. 2011) are included in this section.

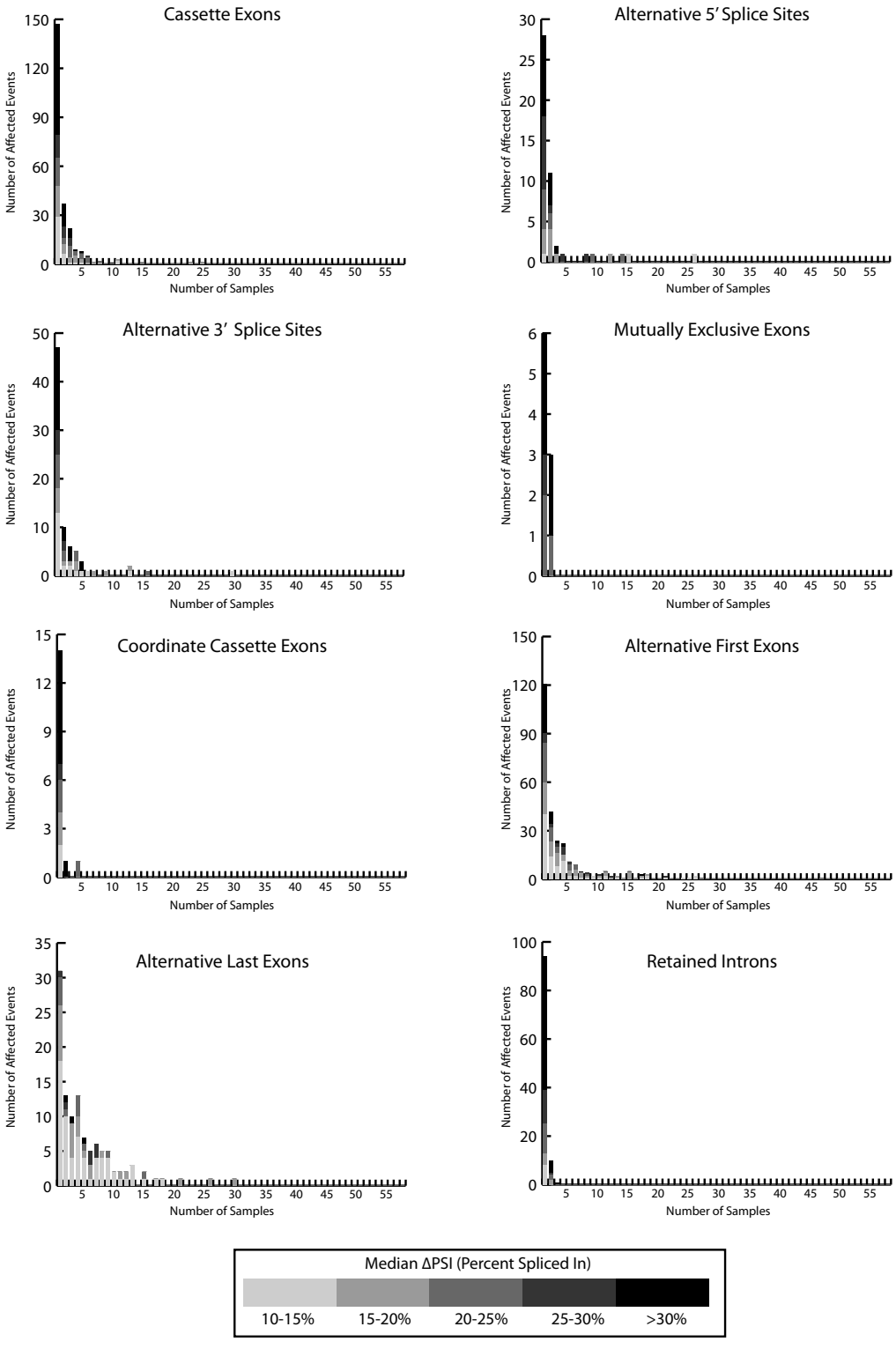


Figure 4.9: **Specific and shared effects by 57 proteins.**

2006) domains RRM (PS50102), KH_TYPE_1(PS50084), KH_TYPE_2(PS50823). The SMART hmms were scanned with hmmsearch (hmmer.org) and Prosite domains with the local version of ScanProsite, ps_scan.pl (Gattiker et al. 2002).

Determining expression of putative splicing regulator genes from Affymetrix tiling array data

38bp Affymetrix tiling arrays were used to hybridize RNA from S2-DRSC cells by members of the modENCODE Consortium and data published in Cherbas et al. 2011. Transcribed fragments (transfrags) from the array data were determined by Cherbas et al. using a bandwidth of 0, maxgap 90, and minrun 50 and were used to define transcribed regions of the S2-DRSC by me. Genes with at least 10% transfrag coverage were considered expressed. For genes not passing the transfrag coverage cutoff, probe intensities were viewed by eye and additional genes were called expressed.

Expression and purification of recombinant PS

Experiments performed by Jung Park. A full-length PS cDNA was cloned into pTrc-His2, expressed in E. coli and purified using Ni-NTA chromatography (QIAGEN) followed by DEAE chromatography.

Gel shift assay

Assay was performed by Jung Park. RNAs containing either three UCAU or GCAU repeats were radiolabeled with 32P-UTP during in vitro transcription. These RNAs were incubated with recombinant PS at room temperature for 20-30 minutes and the mixture was loaded onto 5% non-denaturing gel electrophoresis. The gel was dried and analyzed by phosphorimager and Optiquant software (Perkin Elmer, CA).

Identifying changes in overall gene expression

FPKM values were obtained for all genes in the untreated and *ps*(RNAi) samples using Cufflinks (Trapnell et al. 2010). Cufflinks runs were done by Michael O. Duff and Brenton R. Graveley. A Fisher's test was applied to identify genes with significantly different expression between untreated and *ps*(RNAi) given a Benjamini-Hochberg corrected p-value ≤ 0.05 for 7,834 tests. No genes containing a KH or RRM domain were found to be significantly changing; however, RNA binding proteins eIF-4a and CG10630 had significant, relatively small changes in gene expression (15% increase and 30% decrease,

respectively) compared to the 60% decrease in gene expression of Pasilla. eIF4a had an FPKM of 2,683 in Untreated and 2,366 in PS-RNAi. CG10630 had an FPKM of 165 in Untreated and 236 in the knockdown. It is unclear if the small changes in gene expression relative to the difference in Pasilla would have an effect on splicing.

RNA interference

RNA interference was performed by Li Yang, essentially as described previously (Park and Graveley 2005; Park et al. 2004). This description of the experimental methods was written by Li Yang and Brenton R. Graveley. A vector encoding double-stranded RNA for *ps* was generated as described previously (Park and Graveley 2005; Park et al. 2004). Briefly, cDNA fragments encoding for the specific dsRNA were amplified by RT-PCR with gene-specific primers (Brooks et al. 2011, Supplemental Table 1) from total RNA isolated from S2-DRSC cells. The cDNA fragment was then cloned into the pCRII-TOPO vector (cat. no. K4600-01, Invitrogen) and sequenced to verify the identity of the insert. DNA templates were amplified with M13 forward and M13 reverse primers. PCR products were used in individual in vitro transcription reactions with the Ampliscribe High Yield Transcription SP6 (cat. no. AS3106, Epicentre) Kit and T7 kits (cat. no. AS3107, Epicentre) to generate the sense and antisense RNA strands. After DNase I digestion, the two single-stranded RNAs were annealed to generate double-stranded RNAs. Integrity of the PCR products, the single-stranded RNA transcripts, and dsRNAs were monitored by agarose gel electrophoresis. S2-DRSC cells (obtained from the *Drosophila* Genomics Resource Center at Indiana University) were cultured with Schneider's medium (Sigma/Aldrich, cat. no. S0146) plus 10% heat-inactivated FCS (HyClone, cat. no. SH30070.03) at 27 °C. One day prior to dsRNA treatment, cells were split into six-well culture dishes at a density of 1×106 cells/ml. Immediately prior to the addition of dsRNA, the culture medium was replaced with fresh Schneider's medium without FCS, followed by the addition of 20 μg of each dsRNA directly into the FCS free medium and the cells incubated for 5-hours at 27 °C. After incubation with the dsRNA, 10% FCS was added back to cell culture. After 2 days, a second dose of 20 μg of dsRNA was added to each well in the same manner as described above and the cells incubated for 2 additional days after the re-addition of 10% FCS. After the dsRNA treatment, total RNA was isolated using TRIzol reagent (Invitrogen) according to the manufacturer's directions. Parallel dsRNA treatments and total RNA preparations were performed independently for each replicate. Untreated S2-DRSC cells were used as a reference. To monitor the level of mRNA depletion, primer sets that amplify regions of the targeted mRNAs outside of the dsRNA region were used for RT-PCR amplification, and compared with the results from the untreated cells.

Deep sequencing

All sequencing libraries were prepared by Li Yang and Gemma E. May with the mRNA-Seq Sample Prep Kits (Illumina, P/N 1004814) according to the manufacturer's instructions. This description of library preparation was written by Li Yang and Brenton R. Graveley. Briefly, poly(A)+ RNA was purified from total RNA with oligo-dT magnetic beads. The poly(A)+ RNA was fragmented using divalent cations under elevated temperature, followed by first and second strand cDNA synthesis primed with random hexamers. The cDNA fragments were end-repaired using T4 DNA polymerase and Klenow DNA polymerase, and phosphorylated at their 5´ ends with T4 polynucleotide kinase. After adding 'A' bases to the 3'-end of the DNA fragments, Illumina adaptor oligonucleotides were ligated to the ends and ~300 bp fragments were isolated from an agarose gel, enriched by PCR amplification, and gel-purified again. The samples were quantitated using a Nanodrop, loaded onto a flow-cell for cluster generation, and sequenced on an Illumina Genome Analyzer II using either single-read or paired-end protocols (Illumina).

Validation

Validations performed by Li Yang and Gemma E. May. This description was written by Li Yang and Brenton R. Graveley. Alternative splicing events identified by analysis of the RNA-Seq data were validated by RT-PCR. Briefly, PCR primers were designed to amplify multiple isoforms with different sizes. By comparing the splicing patterns between untreated cells and *ps*(RNAi) treated cells, the data obtained with Illumina sequencing were substantially confirmed for all genes tested by RT-PCR (Figure 4.3) with gene specific primers.

Transcript annotations

For the *pasilla* RNA-Seq study, coding and non-coding transcript annotations were obtained from FlyBase r5.11 (Tweedie et al. 2009) and MB5 (www.modencode.org) and merged into a non-redundant set, allowing a 10 nt difference in the start and end coordinate of the first and last exon, respectively. Gene loci were inferred from a set of non-redundant transcripts by combining all transcripts with overlapping exons into a single gene locus. For the study of 57 additional proteins, the MB8 (www.modencode.org) annotation was used.

Splice junction sequences

A database of 58,212 annotated and 221,388 unannotated (novel) splice junction sequences was created. 215,757 of the unannotated splice junctions were generated by joining every

annotated exon with all possible downstream exons within the same gene. An additional set of 5,631 novel junction sequences were created by joining every pair of exons from different gene loci that were $\leq$2kb away. A separate database of 5,409,600 random splice junctions was created by joining each annotated 5' splice site with 50 randomly drawn annotated 3' splice sites located on a different chromosome and from each annotated 3' splice site with 50 randomly drawn annotated 5' splice sites from a different chromosome. All splice junctions contained 31 nt or 76 nt of exon sequence on either side of the junction in order to force an an alignment overhang of at least 6 nt from one side of the splice junction to the other.

Single read sequence alignments

Alignments performed by Michael O. Duff and Brenton R. Graveley. Bowtie (Langmead et al. 2009) (with parameters: -m 1 –v 2 –best -y) was used to align the single read sequences against a combined index containing both the *D. melanogaster* genome (dm3 assembly) and the splice junctions. All reads were first trimmed from the 3' end to a total length of 37 nt. Reads that mapped uniquely with up to 2 mismatches were reported.

Paired-end sequence alignments

Paired-end alignment method was created by Michael O. Duff. This description was written by Michael O. Duff. Paired-end alignments were conducted using Spliced Paired-End Aligner (SPA) which is a custom perl script (spa.pl, Supplemental Materials in Brooks et al. 2011) that uses Bowtie to independently align each read of a mate-pair and then parses the output files to identify the optimal alignment position for each read of each mate pair. Specifically, spa.pl calls Bowtie (version 0.9.9.2) to separately align each read of the mate pair to the combined genome and splice junction database using using the parameters -v 2 -k 10 -m 10 -y -B 1 which reports all mapping locations for each read that maps with up to 2 mismatches to 10 or fewer locations using 1-based alignment coordinates. Next, spa.pl collects the genomic coordinates of the reads that map not only to the genome, but also the splice junctions, and then considers all possible combinations of the alignment positions of both reads of each mate-pair (up to 100 possibilities if both reads map to 10 locations). These possibilities are then filtered to identify mate-pair combinations in which both reads align to the same chromosome, the reads are oriented-towards on another (i.e., there is both a "forward" and "reverse" read), and the reads are located within 200 kbp of one another on the genome. In cases where exactly one combination of mapping locations fulfills all three criteria, the mapping locations of the reads are reported. Optionally, but not used for this study, in cases where more than one combination of mapping locations fulfills all three criteria, the combination with the shortest genomic location is reported. In cases where one read of the mate pair can be mapped uniquely, but the other read cannot be mapped at all, the read is harvested as a uniquely aligned single read.

Bioinformatic analysis of position specific motif enrichment in targets of hrp36, hrp38, hrp40, and hrp48, determined by splice junction microarray

Cassette exon, competing 5' splice sites, and competing 3' splice sites (collectively referred to as AS events) were inferred from FlyBase r5.4 (Tweedie et al. 2009) transcript annotations using a pipeline based on modified scripts from the *Drosophila melanogaster* Exon Database (Lee et al. 2004). For each hnRNP, affected AS events were identified by mapping affected exon-junction probes to junctions involved in each AS event. As a control, all AS events not affected by the hnRNP, but expressed in S2 cells, were identified. For each hnRNP, every upstream and downstream region near exon-intron borders of all affected and control AS events were scanned to count the number of occurrences of the SELEX motif above a Z-score of 2. The Z-score for a given sequence was calculated from the mean and standard deviation of motif log-odd scores derived from scans of all gene loci in FlyBase r5.4. Counts were done in three ways. The first two used non-overlapping windows of 20nt or 50nt, indexed from the splice site. For the third analysis, as an alternative to using specified window lengths, each region was divided into three equal-sized partitions and motif counts in each partition were normalized by the length of the partition. To control for differences in length of the overall exon or intron, only regions of similar length were compared between the affected and control sets (e.g., short introns (<80nt) in the affected events were only compared with corresponding short introns of the control events). Exon lengths were separated into short (<125nt) medium (125-350nt) and long (>350nt) sizes, where sizes were determined by the length distribution of all exons in FlyBase r5.4; intron sizes were <80nt, 80-350nt, >350nt, respectively, determined by the length distribution of all introns. To assess significance I used a Wilcoxon rank-sum test using the exactRankTests package in R, and then applied a conservative Bonferroni correction for 2,365 trials. These trails represent the three analysis methods, the three different lengths that each exon and intron were grouped in, and the multiple windows in each exon and intron examined for each hnRNP. After Bonferroni correction, no window had a statistically significant greater average count of the SELEX motif in the affected AS events compared to the control AS events; the lowest p-value from all trials was 0.001. I present enrichment without the Bonferroni correction in Figure 5 and Supplementary Figures 8a-i in (Blanchette et al. 2009)(positions with uncorrected p<0.01).

Overrepresented hexamers near *ps*-affected cassette exons

Any overlapping phastCons conserved element within each PS-affected cassette exon and 150 nt into the flanking introns were extracted from the UCSC Genome Browser MySQL database (http://genome.ucsc.edu/, April 2006 Assembly) (Chiaromonte et al. 2002; Kent et al. 2003; Schwartz et al. 2003; Blanchette et al. 2004; Siepel et al. 2005; Rhead et al. 2010). This was also done for the set of PS-unaffected cassette exons (those with a p-val $\geq$

0.95). The proportion of hexamer sequences within the affected sequences and the control sequences were used to perform a Z-test for the difference in population proportions. No hexamer was significant given a Bonferroni-corrected p-value of 0.05 for 4,096 tests. The top five scoring hexamers are reported; they have a raw p-value < 2.0e-4.

Calculation of conserved YCAY motif clusters

Conserved YCAY motifs were searched near cassette exons, alternative 5' splice sites, alternative 3' splice sites, alternative first exons, alternative last exons, and retained intron events. For each AS event, a set of control exons were identified as those events with a corrected p-value ≥ 0.95. The average conserved YCAY cluster score was calculated in 45 nt windows with a step of 20 nt near each alternative splicing event, similar to what was described in (Ule et al. 2006). To ensure that I was implementing the correct YCAY conserved cluster score method, I reanalyzed the mouse exon sequences reported to be regulated by the Nova splicing regulator (Ule et al. 2006). After personal communications with Jernej Ule and Robert Darnell, I realized that the YCAY conserved cluster score, calculated using predefined scores as described in the Supplemental Methods of Ule et al. (2006) ("predefined score method"), was slightly different from the score actually used for analyses and to generate figures in that paper ("counting method"). The software used for the analyses and figures in Ule et al. (2006), implementing the counting method, is available at http://splicing.rockefeller.edu/map/. That software counts the number of times the following motifs occurred in a given window:

- YCAY(N)6YCAY
- YCAYCAY
- YCAY(N)2YCAY,
- YCAYCAY(N)6YCAY,
- YCAY(N)6YCAYCAY
- YCAY(N)2YCAY(N)6YCAY
- YCAY(N)6YCAY(N)2YCAY
- YCAYCAYCAY,
- YCAY(N)2YCAY(N)2YCAY

My implementation of both predefined scores and counting methods on the mouse Nova-regulated exons from Ule et al. (2006) is included in Figure 4.10. The method using

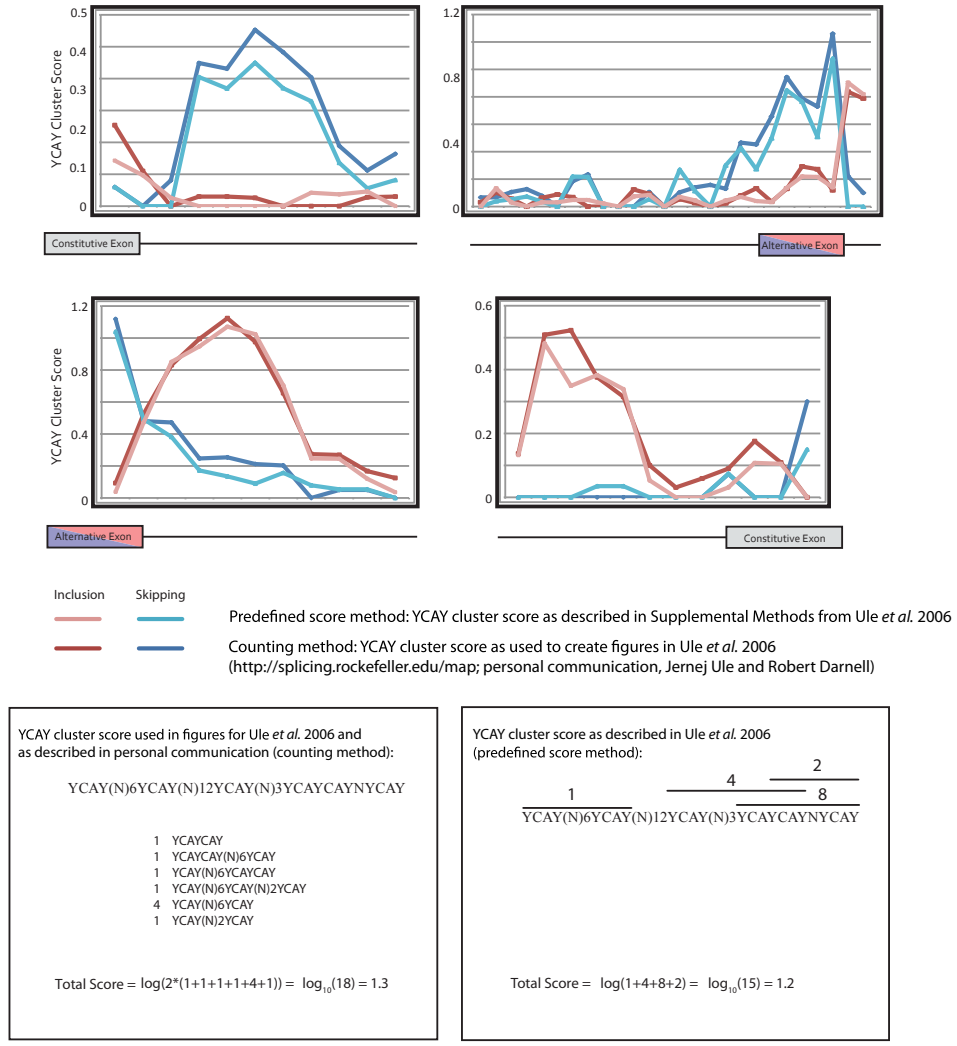


Figure 4.10: **Scoring vs. Counting method for calculating a YCAY cluster score.** Comparing the "Scoring" method and "Counting" method for calculating a YCAY cluster score on reported (Ule et al. 2006) mouse Nova-regulated exons.

predefined scores does not change the overall conclusions of the study; therefore, we decided to use the predefined scores method exactly as described in Ule et al. (2006) for our study on the Pasilla targets. *D. simulans, D. sechellia, D. yakuba*, and *D. erecta* were considered closely related species and *D. pseudobscura, D. ananassae, D. grimshawi, D. mojavensis, D. persimilis, D. virilis,* and *D. willistoni* as distant species. Thus, the conserved cluster score for a given window near PS-regulated events was: Conserved S = 2*MIN S(*D. melanogaster, D. simulans, D. sechellia, D. yakuba, D. erecta*) + AVERAGE S(*D. melanogaster, D. simulans, D. sechellia, D. yakuba, D. erecta*) + AVERAGE S (*D. pseudobscura, D. ananassae, D. grimshawi, D. mojavensis, D. persimilis, D. virilis, D. willistoni*) Where S in a given window is the log10 of the sum of occurrences of the predefined motifs described in Ule et al. (2006).

If a cassette exon had multiple flanking introns, the longest intron was taken. For alternative 5' and 3' splice sites, the cluster scores were calculated near the constitutive splice site as well as near the alternative splice sites. If an alternative 5' or 3' splice site had multiple exclusion introns, the longest one was chosen. For alternative first exon events, the cluster scores are calculated near the constitutive splice site, near both alternative splice sites, and near the transcriptional start site; similarly for alternative last exon events, except near the poly(A) site. For retained intron events, cluster scores were calculated near both 5' and 3' splice sites. Plots of the average conserved cluster score were made only from introns that were $\geq$ 400 nt.

To identify specific windows that had a significant enrichment of YCAY motifs, I performed a Wilcoxon-rank sum test on every window for every event, which yielded 452 tests. Positions with a p-value $\leq$ 0.01 are analogous to positions of conserved YCAY clusters near mouse Nova target cassette exons. Moreover, I report positions that have a significant p-value given a more stringent Bonferroni correction of 0.05 upstream of alternative 5' splice sites and within retained introns.

Calculating a net conserved YCAY conserved cluster score

In Ule et al. 2006, a net conserved YCAY cluster score, S, for a particular exon was given by:

Net conserved S = $\frac{1}{2}$(MAX(NISE1, NISE2, NISE3, SUM(NISE2, NISE3)*$\frac{2}{3}$) -

MAX(NISS1, NISS2, NESE))

Where NISE1, NISE2, and NISE3 were Nova intronic splicing enhancer regions, NISS1, NISS2 were Nova intronic splicing silencer regions, and NESE was a Nova exonic splicing enhancer region. Given the YCAY cluster scores in windows surrounding the PS-affected cassette exons, I only identified significant enrichment at positions analogous to NISE1, NISE2, NISS1, and NISS2. I therefore modified the net conserved cluster score to correspond with only positions that were significantly enriched with conserved YCAY clusters near PS-affected. My modified cluster score is:

Net conserved S = $\frac{1}{2}$(MAX(NISE1:-180, NISE2:+40, NISE2:+60, NISE2:+80) -

MAX(NISS1:+120, NESS1:+40))

The relative positions correspond to the locations of significant enrichment indicated by an asterisk in Figure 4.6. Scores were calculated using only introns > 400nt.

Total number of alternative splicing events

The number of possible alternative splicing events was identified by using all junctions in the merged FlyBase r5.11 and MB5 annotation plus any additional confident novel junctions as input into JuncBASE for the *pasilla* study. For the study involving depletion of 57 proteins, all junctions and exons identified in (Graveley et al. 2011) and confident junctions from the RNAi study were used as input to JuncBASE to determine the set of all alternative splicing events. An alternative splicing event was considered "expressed" if at least one junction from any isoform was present in both the untreated and *ps*(RNAi) samples.

Identifying *Drosophila* orthologs of Nova-regulated genes

A list of cassette exons regulated by Nova, in mouse, were obtained from Supplemental Table 1 in Ule et al. Nature, 2006. Each gene was queried in TreeFam (Li et al. 2006; Ruan et al. 2008) to identify the orthologous *Drosophila* gene(s). Out of 47 genes regulated by Nova, 33 had at least one *Drosophila* ortholog in the TreeFam database. To determine if these *Drosophila* orthologs were expressed in the S2-DRSC cell line, I identified genes that had at least one confident junction as the threshold for expression.

GO enriched terms

Funcassociate 2.0 (Berriz et al. 2009) was used to identify GO terms that were enriched in the set of PS target alternative splicing events. FlyBase CG names from all 323 PS-affected genes were used as the query set. A "gene space" set can be used in Funcassociate, to control for genes that might be enriched due to a bias in expression of the gene in the cell type. For this background set of genes, I used all genes that had at least one confident junction as a threshold for expression.

Obtaining a non-redundant set of affected splicing events

For each alternative splicing event analyzed upon knockdown of 57 proteins, 58 Fisher's tests were performed to compare the median reference exclusion and inclusion counts

to every exclusion and inclusion count in RNAi samples and the untreated sample. A combined p-value was determined for each event in order to rank the significance of each event. The combined p-value for an event was equal to the product of the 58 Benjamini-Hochberg-corrected p-values that were lower than 0.05. For each type of alternative splicing event, if a subset of the events overlapped in coordinate space, the event with the lowest combined p-value was retained in the final analysis.

Chapter 5

Alternative splicing changes throughout 30 *D. melanogaster* developmental time points

5.1 Introduction

To understand the biological processes involved in the development of an organism from a single-celled zygote to a full adult, it is important to identify changes in gene expression that correspond to morphogenic changes. The identification of developmentally regulated genes in *D. melanogaster* has led to a further understanding of the function of genes (Arbeitman et al. 2002; Stolc et al. 2004; Manak et al. 2006). In addition, tracking temporal expression changes of gene regulators with their targets may uncover gene regulatory networks or identify additional correlated target genes (e.g., Wang et al. 2004a; Qian et al. 2003). Previous studies have shown dynamic gene expression changes during the developmental life cycle of *Drosophila melanogaster* (Arbeitman et al. 2002; Stolc et al. 2004; Manak et al. 2006); however, these studies examined the transcriptome using microarrays and could not observe the full extent of developmentally regulated alternative splicing. By examining changes in alternative splicing during a developmental process, we can begin to understand the functional roles of these alternative splicing events. We can also gain a better understanding of the role of alternative splicing in regulating biological processes. Additionally, by looking at correlated gene expression changes of splicing regulators with changes in their target exons (determined by RNAi studies) in normal biological condition, we can further support the role of the protein in regulating splicing of its targets.

5.2 Results

5.2.1 Strategy for characterization of the transcriptome

This subsection describes the general strategy for identifying transcripts throughout development. The work was done by multiple individuals as part of the modENCODE consortium and published in Graveley et al. (2011)[1]. The following is a modified excerpt from that publication, primarily written by Brenton R. Graveley and Susan E. Celniker.

To discover new transcribed features and comprehensively characterize their expression dynamics throughout development, my collaborators conducted complementary tiling microarray and RNA-Seq experiments using RNA isolated from 30 whole-animal samples representing 27 distinct stages of development. These included 12 embryonic samples collected at 2-h intervals for 24 h, six larval, six pupal and three sexed adult stages at 1, 5 and 30 days after eclosion. To attain single-nucleotide resolution and to facilitate the analysis of alternative splicing, they performed non-strand-specific poly(A)+ RNA-Seq from all 30 samples generating a combination of single and paired-end ~75-bp reads on the Illumina Genome Analyser IIx platform (short poly(A)+ RNA-Seq). For the short poly(A)+ RNA-Seq experiments, the same library was sequenced in different labs; however, these technical replicates were highly correlated. The results described here are derived from merging all lanes sequenced from each biological replicate. To identify primary transcripts and non-coding RNAs, the 12 embryonic time points were also interrogated with strand-specific 50-bp sequence reads from partially rRNA-depleted total RNA on the Applied Biosystems SOLiD platform. To improve connectivity, mixed-stage embryos, adult males and adult females were used to generate ~250-bp reads on the Roche 454 platform (non-strand-specific long poly(A)+ RNA-Seq). In total, they generated 176,962,906,041 bp of mapped sequence representing 1,266-fold coverage of the genome and 5,902-fold coverage of the annotated *D. melanogaster* transcriptome.

5.2.2 Discovery and dynamics of alternative splicing

The following subsection was written by Brenton R. Graveley and Susan E. Celniker, with input by me and other co-authors. Work performed in this section was done by me, unless otherwise specified. To characterize constitutive and alternative splicing, 71,316

[1]Co-authors in Graveley et al. (2011): Brenton R. Graveley, Joseph W. Carlson, Michael O. Duff, Jane M. Landolin, Li Yang, Carlo G. Artieri, Marijke J. van Baren, Nathan Boley, Benjamin W. Booth, James B. Brown, Lucy Cherbas, Carrie A. Davis, Alex Dobin, Renhua Li, Wei Lin, John H. Malone, Nicolas R. Mattiuzzo, David Miller, David Sturgill, Brian B. Tuch, Chris Zaleski, Dayu Zhang, Marco Blanchette, Sandrine Dudoit, Brian Eads, Richard E. Green, Ann Hammonds, Lichun Jiang, Phil Kapranov, Laura Langton, Norbert Perrimon, Jeremy E. Sandler, Kenneth H. Wan, Aarron Willingham, Yu Zhang, Yi Zou, Justen Andrews, Peter J. Bickel, Steven E. Brenner, Michael R. Brent, Peter Cherbas, Thomas R. Gingeras, Roger A. Hoskins, Thomas C. Kaufman, Brian Oliver, and Susan E. Celniker

Splicing event	FlyBase r5.12	modENCODE	New events	Short poly(A)+ RNA-Seq	Significantly changing
Cassette exons	793	2,717	2,014	2,369	1,539
Alternative 5' splice sites	843	5,192	4,599	4,583	3,142
Alternative 3' splice sites	879	6,253	5,505	5,579	3,242
Mutually exclusive exons	229	251	123	228	226
Coordinate cassette exons	301	1,227	979	992	467
Alternative first exons	1,767	4,936	3,442	4,473	3,996
Alternative last exons	227	604	432	553	471
Retained/unprocessed introns	1,434	2,679 (5,667)	1,275 (4,263)	2,439 (35,641)	868 (8,998)
Total	6,437	23,859 (26,847)	18,369 (21,478)	21,216 (54,418)	13,951 (22,081)

Table 5.1: **Classification of alternative splicing events.** The number of retained/unprocessed introns in parentheses indicates the total number identified, whereas the number not in parentheses indicates the subset of identified events that have been validated by cDNA sequences or FlyBase 5.12 annotations. Here, modENCODE refers to splicing events identified using the FlyBase r5.12 and all sequencing data described in subsection 5.2.1.

splice junctions were identified, of which 22,965 were new discoveries. Of the new splice junctions, 26% were supported by multiple experimental data types and 74% by only one data type, primarily short poly(A)+ RNA-Seq. A total of 102,026 exons were identified. The analysis to determine the total number of splice junctions and exons was done by my collaborators; however, I contributed to this work (see Methods). To examine splicing dynamics throughout development, I categorized all splicing events into the common types of alternative splicing events (Table 5.1). I identified 23,859 splicing events, of which 18,369 were new or recategorized, a threefold increase from annotated splicing events. An additional 2,988 intron-retention events were identified from the short poly(A)+ RNA-Seq data, and are yet to be supported by other experimental data. In all, 7,473 genes contain at least one alternative splicing event, which is 60.7% of the 12,295 expressed multi-exon genes—also a threefold increase in the fraction of genes with alternatively spliced transcripts. Although smaller than the fraction of human genes with alternatively spliced transcripts (95%) (Wang et al. 2008), a larger proportion of *Drosophila* genes encode alternative transcripts than was previously known. Of the new alternative exons, 8,226 were previously annotated as constitutive. As previously observed (Philipps et al. 2004), annotated cassette exons, and their flanking introns, are more highly conserved than annotated constitutive exons (Figure 5.1). To identify features of my newly discovered alternatively spliced exons, Brenton Graveley looked at the conservation of sequence within and surrounding the new cassete exons as well as the phase of the exons (Figure 5.1.a). The newly discovered cassette exons are more highly conserved than the new constitutive exons, although both classes are less conserved than the corresponding class of annotated exons. New cassette exons that were previously annotated as constitutive exons are the most highly conserved set of exons. Annotated and new cassette exons show a strong tendency

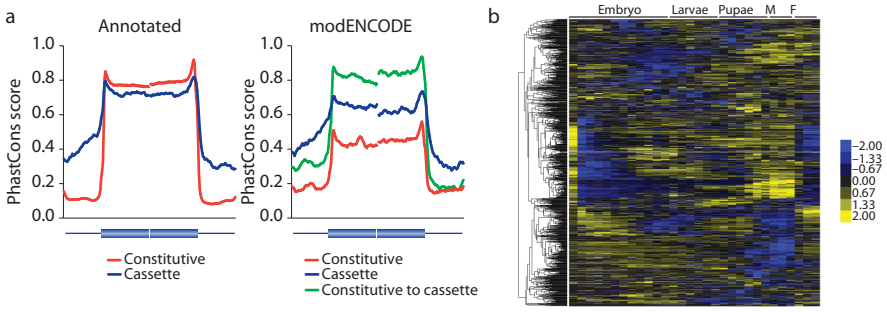


Figure 5.1: **Developmentally regulated splicing events.** Figure created by Brenton R. Graveley. a, Sequence conservation across 14 related insect species of internal constitutive and cassette exons > 50nt that were annotated or new discoveries. Amount of sequence conservation was determined by a PhastCons score (Siepel et al. 2005; Rhead et al. 2010) (http://genome.ucsc.edu). Conservation analysis done by Brenton R. Graveley. (Annotated constitutive, n=26,127; annotated cassette, n=438; modENCODE cassette, n=173; modENCODE constitutive, n=306; FlyBase 5.12 constitutive to modENCODE cassette, n=304.) b, Clusters of regulated cassette exon events during development. The scale bar indicates Z-scores of Ψ for each event.

to preserve reading frame, indicating that these transcripts increase protein diversity (data not shown; Graveley et al. 2011). Both annotated and new cassette exons tend to be shorter than their constitutive counterparts, although both sets of new exons tend to be shorter than annotated exons.

To assess the extent of splicing variation I calculated the 'percent spliced in' or Ψ (Wang et al. 2008) for each splicing event in each sample as well as the switch score (ΔΨ) by determining the difference between the highest and lowest Ψ values across development (ΔΨ = Ψ max − Ψmin). This revealed a very smooth distribution of Ψ among all events, indicating that the splicing of most exons is fairly constant whereas only a minority change markedly (Figure 5.2) The analysis to look at the distribution of ΔΨ scores was done by Brenton Graveley. Only 831 splicing events have a △Ψ value >90. Further statistical analyses identified 15,847 (67%) alternative splicing events that change significantly throughout development. Hierarchical clustering of cassette exon events revealed the dynamic nature of splicing throughout development (Figure 5.1.b), as exemplified by Cadherin-N (CadN), a gene with three sets of mutually exclusive exons. In each set, one exon is preferentially included in early embryos, the other in late embryos, with a smooth transition between the two. My analysis also identified groups of exons that have coordinated splicing patterns (Figure 5.1.b). A set of 55 genes contain exons that are preferentially included in early embryos, late larvae, early pupae and females but skipped in all other stages. Gene Ontology (GO) analysis of these genes indicates that many encode proteins involved in epithelial cell-to-cell junctions. GO analysis of genes that contain exons preferentially included during late pupal and adult stages indicates that

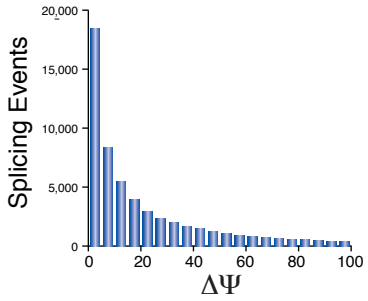


Figure 5.2: **Distribution of ΔΨ changes throughout development.** Figure created by Brenton R. Graveley.

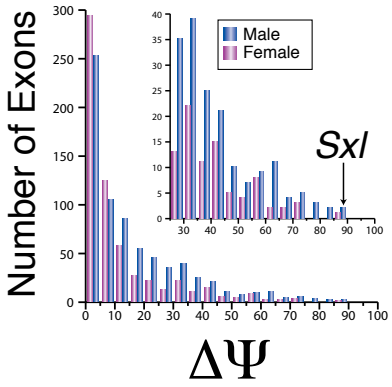


Figure 5.3: **ΔΨ values of sex-biased alternative splicing.** Figure created by Brenton R. Graveley using data from my analysis of sex-biased alternative splicing. One of the most sex-biased alternative splicing events is splicing of Sxl, the master regulators of sex determination in *Drosophila melanogaster.*

many encode proteins that are part of neuronal synapses.

Sex-biased alternative splicing

Sex determination in *Drosophila* is mediated by a cascade of regulated alternative splicing events involving Sex lethal (Sxl), transformer (tra), male-specific lethal 2 (msl-2), doublesex (dsx) and fruitless (fru) that specify nearly all physical and behavioural dimorphisms between males and females as well as X chromosome dosage compensation (Sánchez 2008). Using my analysis of sex-biased alternative splicing events, Brenton Graveley confirmed the sex-biased splicing of Sxl ($\triangle\Psi$ = 89.6), tra ($\triangle\Psi$ = 39.2), dsx ($\triangle\Psi$ = 59.7) and fru ($\triangle\Psi$ = 100). In addition to the canonical sex-determination cascade, I identified 119 strongly sex-biased splicing events ($\triangle\Psi > 70$).

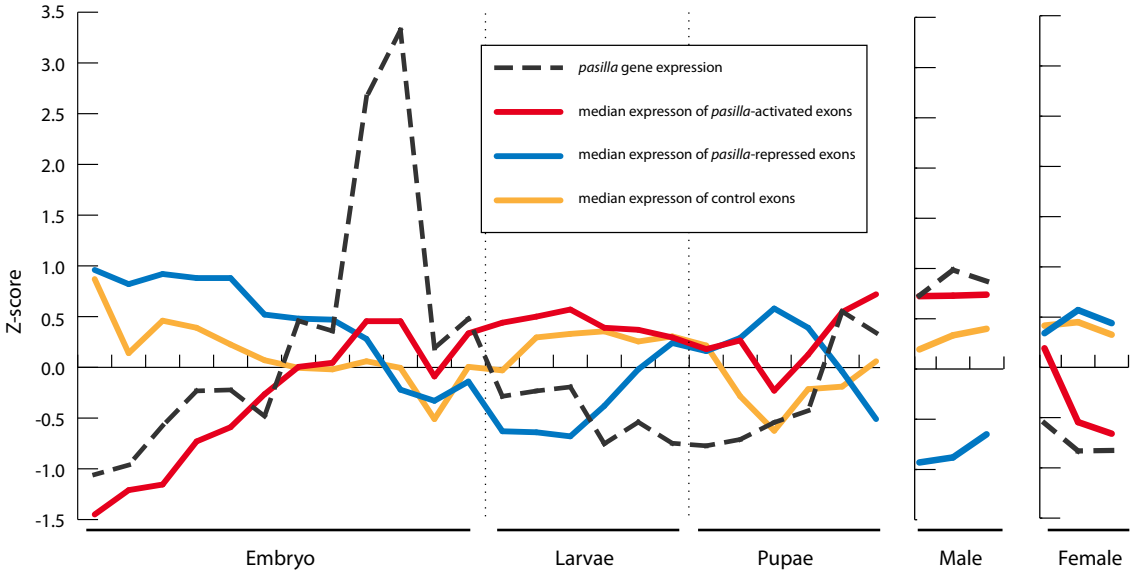


Figure 5.4: **Pasilla-regulated splicing throughout 30 developmental time points.** Dashed black line, Z-score for *pasilla* gene expression, measured as FPKM (fragments per kilobase of exon model per million reads) values using Cufflinks (Trapnell et al. 2010). Red line, median Z-score of Ψ values for cassette exons previously shown to be activated by *pasilla*. Blue line, median Z-score of Ψ values of cassette exons previously shown to be repressed by *pasilla*. Yellow line, median Z-score of Ψ values for cassette exons previously shown to be unaffected by *pasilla,* shown as a control.

5.2.3 Pasilla-regulated splicing throughout development

Based on our studies, the most well characterized splicing regulator is Pasilla (from chapter and section 4.3 on page 53). I found a total 405 splicing events affected by depletion of *pasilla*, 111 of which were cassette exons. The knockdown studies also allowed our group to determine which cassette exons Pasilla would act to activate or represss. Using this information, we were interested in looking for correlations of *pasilla* gene expression with exons that were determined to be PS-activated and anti-correlations of *pasilla* gene expression with PS-repressed exons. I observed this trend when plotting a Z-score of *pasilla* gene expression, against the median Z-score of Ψ values for *pasilla* target exons that were determined by our RNAi studies (Figure 5.4).

There is a striking difference in *pasilla* gene expression between adult males and adult female flies with corresponding differences in splicing of the target exons. This further supports a functional role for *pasilla* in sexual reproduction or gametogenesis, functions that were enriched in the full set of identified Pasilla targets. RNA-Seq studies in male and female-specific gonads may further support *pasilla*'s function. In the larvae and pupae stages, the correlation between gene expression and target splicing is not as pronounced.

This may be due to the effects of additional splicing regulators during these time points or a lag in splicing shifts occuring after increased *pasilla* expression in the late embryo. Another possibility for the lack of correlation in the larvae and puape stages may be from sequencing of whole flies. Better correlations may be observed when sequencing individual tissues.

5.3 Methods

Splice junction discovery and validation

The following description was modified from text written by Brenton R. Graveley and Susan E. Celniker. A variety of methods were used with the RNA-Seq data to predict novel splice junctions and then to evaluate the confidence level of those junctions. My collaborators aligned reads to databases of predicted splice junctions, and to models using TopHat (Trapnell et al. 2009), STAR (unpublished), and BLAT (Kent 2002). Predicted splice junctions were produced by me and is described in chapter and section 3.2.1 on page 38. Other junction alignments performed by other co-authors from (Graveley et al. 2011). To identify new splice junctions they derived a set of metrics that could be used to distinguish true splice junctions from false positive splice junctions. First, they required a minimum of 6 nt overhang across a splice junction which was determined from our studies on *pasilla* RNAi depletion (Figure 3.1 on page 39). Moreover, rather than requiring a specific number of reads to map to a given splice junction (Figure 3.1 on page 39), Michael O. Duff developed an entropy measure for reads that mapped to the splice junction (also described in 3.2.1 on page 38). The entropy score is a function of both the total number of reads that map to a given junction and the number of different offsets to which those reads map and the number that map at each offset. Thus, junctions with multiple reads mapping at each of the possible windows across the junction will be assigned a higher entropy score, than junctions where many reads map to only one or two positions. For this analysis we required that a junction have an entropy score of two or greater in at least two biological samples for junctions with canonical splice sites, and an entropy score of three or greater in at least three biological samples for junctions with non-canonical splice sites. Entropy score cutoffs were determined by examining the distribution of entropy scores of annotated junctions versus unannotated junctions. Entropy was calculated using the following equations:

$p(i)$ = number of reads at offset i / total reads to junction

Entopy of junction =

$$\sum_i p(i) log_2 p(i)$$

We produced a total of ~79,400 high confidence junctions, ~29,000 of which are novel junctions. 259,628,732 reads map uniquely to splice junctions. The set of ~79K splice junctions was further filtered by Jane M. Landolin to exclude junctions with an intron length less than 41 nt, or that inadvertently join parologous genes. Junctions were also excluded if they had non-cannonical splice sites with the exception of those that were previously annotated or had other experimental support. This yielded a final set of 67,317 high confidence splice junctions in the short poly(A)+ RNA-Seq data that correspond to 46,566 (93%) FB5.12 annotated splice junctions and 20,751 new splice junctions. In addition, they identified 32,917 splice junctions from the 454 data of which 28,860 (87.6%) correspond to FB5.12 annotated splice junctions and 4,057 new splice junctions. Members of the consortium confirmed 1,643 unannotated splice junctions by using cDNA, EST, and RT-PCR datasets.

Quantifying gene expression

Gene expression FPKM (fragments per kilobase of exon model per million reads) values were calculated by Cufflinks (Trapnell et al. 2010), using all transcripts identified in this study. Cufflinks analysis done by Michael O. Duff.

Identifying alternatively spliced exons

The Junction Based Analysis of Splicing Events (JuncBASE) package was used to identify the following alternative splicing events: cassette exon, alternative 5' splice sites, alternative 3' splice sites, alternative first exon, alternative last exon, mutually exclusive exons, intron retention and coordinate cassette exons (described in chapter and section 3.3 on page 40). These events were counted as follows:

Cassette exon: The number of cassette exons in the reference annotation or "new" is shown in Table 5.1. Each cassette exon is considered a separate cassette exon event.

Alternative 5' splice sites, alternative 3' splice sites, alternative first exons, and alternative last exon events: These alternative events were grouped by an "anchor" splice site. For example, alternative donor events were assumed to have the same acceptor site ("anchor" splice site) and grouped together in this way. The number of total splicing events in a group is equal to the number of introns involved in the group - 1. This is due to the method of calculating Ψ values, where every proximal alternative splice site (inclusion isoform) is compared to more distal splice sites (exclusion isoform(s)). When comparing events between two different annotation sets, the same group of introns, with the same anchor splice site, were compared. If both annotation sets had the same anchor point, then identical introns were not considered and the remaining intron(s) equaled the number of "novel" event(s). If the reference (FB5.12) annotation did not have the same anchor point, then the

number of new events was equal to the number of events in the "new" annotation (number of introns in the event - 1).

Mutually exclusive exons: The number of exons involved in a mutually exclusive exon event is shown in Table 1. "New" mutually exclusive exons were exons that were not part of a mutually exclusive event in the reference annotation (FlyBase r5.12). n.b. If there was evidence of an isoform that skipped all exons in a mutually exclusive event, it was still considered mutually exclusive. "Percent spliced in" (Ψ) values were calculated by treating each mutually exclusive exon as an inclusion isoform of an event and all other exons in the event as possible exclusion isoforms. If there are only two exons involved in a mutually exclusive event, then a Ψ was calculated for only one of the exons.

Coordinate cassette exons: The number of exons involved in a coordinate cassette exon event is shown in Table 1. "New" coordinate cassette exons are exons that were not part of a coordinate cassette exon event in the reference annotation (FB5.12). A Ψ was calculated for all observed coordinate cassette exon events given a set of exons. For example, there may be four consecutive exons that have evidence that all are skipped but also evidence of additional events with a subset (2 or 3 of the 4) of exons that are included while the rest are skipped.

Intron retention: The number of introns that show evidence of intron retention is given in Table 1. Intron retention events observed in the RNA-Seq data, and subsequently incorporated into the modENCODE annotation, cannot be distinguished between unprocessed RNA; therefore, I report two numbers for the modENCODE intron retention events and the short read poly(A)+ RNA-Seq data: (1) intron retention events that are present in either FlyBase or from modENCODE cDNAs (2) All intron retention events.

JuncBASE was also used to identify exons that were differentially spliced. Read counts to every exclusion and inclusion isoform from each alternative splicing event for each developmental sample were obtained from JuncBASE and subsequently used to perform pairwise comparisons of differential splicing between samples. Fisher's exact test was used to identify differential splicing between each pair of samples and a Benjamini-Hochberg correction (False Discovery Rate < 0.05) was performed for all pairwise comparisons within each type of splicing event.

Exon conservation analysis

Conservation analysis was performed and written here by Brenton R. Graveley. To analyze and compare the conservation of constitutive and cassette exons, he extracted the PhastCons scores (Siepel et al. 2005; Rhead et al. 2010) (http://genome.ucsc.edu) of each individual exon >50 bp and calculated the average PhastCons score at each nt position across the exons within each class. PhastCons is a method to score the amount of sequence conservation at positions in a multiple alignment of related species. Here, the PhastCons score is determined from a multiple alignment of *D. melanogaster* and 14 other insect species.

Z-score and clustering of alternatively spliced exons

For each alternatively spliced exon and for each sample, a percent spliced in (Ψ) was calculated: Ψ = (number of reads to inclusion isoform)/(number of reads to inclusion isoform + number of reads to exclusion isoform). The Ψ was calculated only for events that had a total read count > 25. The $\Delta\Psi$ for an event was the difference between the highest and lowest Ψ across all time points. For clustering analysis, a Z-score for each event was calculated. The Z-score for event i in time point j is:

$$\frac{\Psi_{ij} - \mu_i}{\sigma_i}$$

where μ_i and σ_i is the average and standard deviation, respectively, of Ψ values for event i in all time points j, for which the event is expressed. A negative Z-score is interpreted as an exon that is more skipped when compared to the average Ψ and a positive Z-score corresponds to an exon that is more included when compared to the average Ψ. Hierarchical clustering was performed with optimal leaf ordering (Bar-Joseph et al. 2003). Clusters were viewed with Java TreeView (http://jtreeview.sourceforge.net/). GO analysis was performed using Funcassociate 2.0 (Berriz et al. 2009).

Bibliography

Ashish Agarwal, David Koppstein, Joel Rozowsky, Andrea Sboner, Lukas Habegger, Ladeana W Hillier, Rajkumar Sasidharan, Valerie Reinke, Robert H Waterston, and Mark Gerstein. Comparison and calibration of transcriptome data from RNA-seq and tiling arrays. *BMC Genomics*, 11:383, 2010. doi: 10.1186/1471-2164-11-383.

A. K. Amarasinghe, R. MacDiarmid, M. D. Adams, and D. C. Rio. An in vitro-selected RNA-binding site for the KH domain protein PSI acts as a splicing inhibitor element. *RNA*, 7(9):1239–53, 9 2001. ISSN 1355-8382.

Michelle N Arbeitman, Eileen E M Furlong, Farhad Imam, Eric Johnson, Brian H Null, Bruce S Baker, Mark A Krasnow, Matthew P Scott, Ronald W Davis, and Kevin P White. Gene expression during the life cycle of Drosophila melanogaster. *Science*, 297(5590): 2270–5, Sep 2002. doi: 10.1126/science.1072152.

Kin Fai Au, Hui Jiang, Lan Lin, Yi Xing, and Wing Hung Wong. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res*, 38(14): 4570–8, Aug 2010. doi: 10.1093/nar/gkq211.

Isabel Aznarez, Yoseph Barash, Ofer Shai, David He, Julian Zielenski, Lap-Chee Tsui, John Parkinson, Brendan J Frey, Johanna M Rommens, and Benjamin J Blencowe. A systematic analysis of intronic sequences downstream of 5' splice sites reveals a widespread role for U-rich motifs and TIA1/TIAL1 proteins in alternative splicing regulation. *Genome Res*, 18(8):1247–58, Aug 2008. doi: 10.1101/gr.073155.107.

Ziv Bar-Joseph, Erik D Demaine, David K Gifford, Nathan Srebro, Angèle M Hamel, and Tommi S Jaakkola. K-ary clustering with optimal leaf ordering for gene expression data. *Bioinformatics*, 19(9):1070–8, Jun 2003.

Yoseph Barash, John A. Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J. Blencowe, and Brendan J. Frey. Deciphering the splicing code. *Nature*, 465(7294):53–9, 5 2010. ISSN 1476-4687. doi: 10.1038/nature09000.

Sergio Barberan-Soler, Pedro Medina, Jeffrey Estella, James Williams, and Alan M Zahler. Co-regulation of alternative splicing by diverse splicing factors in Caenorhabditis elegans. *Nucleic Acids Res*, 39(2):666–74, Jan 2011. doi: 10.1093/nar/gkq767.

Nuno L. Barbosa-Morais, Maria Carmo-Fonseca, and Samuel Aparício. Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res*, 16(1):66–77, 1 2006. ISSN 1088-9051. doi: 10.1101/gr.3936206.

Gabriel F. Berriz, John E. Beaver, Can Cenik, Murat Tasan, and Frederick P. Roth. Next generation software for functional trend analysis. *Bioinformatics*, 25(22):3043–4, 11 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp498.

Marco Blanchette, Richard E. Green, Steven E. Brenner, and Donald C. Rio. Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila. *Genes Dev*, 19(11):1306–14, 6 2005. ISSN 0890-9369. doi: 10.1101/gad.1314205.

Marco Blanchette, Richard E. Green, Stewart MacArthur, Angela N. Brooks, Steven E. Brenner, Michael B. Eisen, and Donald C. Rio. Genome-wide analysis of alternative pre-mRNA splicing and RNA-binding specificities of the Drosophila hnRNP A/B family members. *Mol Cell*, 33(4):438–49, 2 2009. ISSN 1097-4164. doi: 10.1016/j.molcel.2009.01.022.

Mathieu Blanchette, W. James Kent, Cathy Riemer, Laura Elnitski, Arian F. A. Smit, Krishna M. Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D. Green, David Haussler, and Webb Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, 14(4):708–15, 4 2004. ISSN 1088-9051. doi: 10.1101/gr.1933104.

H. J. Böckenhauer and D. Bongartz. *Algorithmic aspects of bioinformatics*. Springer-Verlag New York Inc, 2007.

Angela N. Brooks, Li Yang, Michael O. Duff, Kasper D. Hansen, Jung W. Park, Sandrine Dudoit, Steven E. Brenner, and Brenton R. Graveley. Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res*, 21(2):193–202, Feb 2011. ISSN 1549-5469. doi: 10.1101/gr.108662.110.

Douglas W Bryant, Jr, Rongkun Shen, Henry D Priest, Weng-Keen Wong, and Todd C Mockler. Supersplat–spliced RNA-seq alignment. *Bioinformatics*, 26(12):1500–5, Jun 2010. doi: 10.1093/bioinformatics/btq206.

Luca Cartegni, Shern L. Chew, and Adrian R. Krainer. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet*, 3(4): 285–98, 4 2002. ISSN 1471-0056. doi: 10.1038/nrg775.

John Castle, Phil Garrett-Engele, Christopher D. Armour, Sven J. Duenwald, Patrick M. Loerch, Michael R. Meyer, Eric E. Schadt, Roland Stoughton, Mark L. Parrish, Daniel D. Shoemaker, and Jason M. Johnson. Optimization of oligonucleotide arrays and RNA

amplification protocols for analysis of transcript structure and alternative splicing. *Genome Biol*, 4(10):R66, 2003. ISSN 1465-6914. doi: 10.1186/gb-2003-4-10-r66.

John C. Castle, Chaolin Zhang, Jyoti K. Shah, Amit V. Kulkarni, Auinash Kalsotra, Thomas A. Cooper, and Jason M. Johnson. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat Genet*, 40 (12):1416–25, 12 2008. ISSN 1546-1718. doi: 10.1038/ng.264.

Susan E Celniker, Laura A L Dillon, Mark B Gerstein, Kristin C Gunsalus, Steven Henikoff, Gary H Karpen, Manolis Kellis, Eric C Lai, Jason D Lieb, David M MacAlpine, Gos Micklem, Fabio Piano, Michael Snyder, Lincoln Stein, Kevin P White, Robert H Waterston, and modENCODE Consortium. Unlocking the secrets of the genome. *Nature*, 459(7249):927–30, Jun 2009. doi: 10.1038/459927a.

Mo Chen and James Manley. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature Reviews Molecular Cell Biology*, 10(11): 741–754, 9 2009. ISSN 1471-0072.

Lucy Cherbas, Aarron Willingham, Dayu Zhang, Li Yang, Yi Zou, Brian D. Eads, Joseph W. Carlson, Jane M. Landolin, Philipp Kapranov, Jacqueline Dumais, Anastasia Samsonova, Jeong-Hyeon H. Choi, Johnny Roberts, Carrie A. Davis, Haixu Tang, Marijke J. van Baren, Srinka Ghosh, Alexander Dobin, Kim Bell, Wei Lin, Laura Langton, Michael O. Duff, Aaron E. Tenney, Chris Zaleski, Michael R. Brent, Roger A. Hoskins, Thomas C. Kaufman, Justen Andrews, Brenton R. Graveley, Norbert Perrimon, Susan E. Celniker, Thomas R. Gingeras, and Peter Cherbas. The transcriptional diversity of 25 Drosophila cell lines. *Genome Res*, 21(2):301–14, 2 2011. ISSN 1549-5469. doi: 10.1101/gr.112961.110.

M. D. Chiara and R. Reed. A two-step mechanism for 5' and 3' splice-site pairing. *Nature*, 375(6531):510–3, 6 1995. ISSN 0028-0836. doi: 10.1038/375510a0.

F. Chiaromonte, V. B. Yap, and W. Miller. Scoring pairwise genomic sequence alignments. *Pac Symp Biocomput*, pages 115–26, 2002. ISSN 1793-5091.

Alexander Churbanov, Igor Vorechovsky, and Chindo Hicks. Computational prediction of splicing regulatory elements shared by tetrapoda organisms. *BMC Genomics*, 10(1): 508–, 2009. ISSN 1471-2164.

Gavin E. Crooks, Gary Hon, John-Marc M. Chandonia, and Steven E. Brenner. Weblogo: a sequence logo generator. *Genome Res*, 14(6):1188–90, 6 2004. ISSN 1088-9051. doi: 10.1101/gr.849004.

Debopriya Das, Tyson A Clark, Anthony Schweitzer, Miki Yamamoto, Henry Marr, Josh Arribere, Simon Minovitsky, Alexander Poliakov, Inna Dubchak, John E Blume, and John G Conboy. A correlation with exon expression approach to identify cis-regulatory

elements for tissue-specific alternative splicing. *Nucleic Acids Res*, 35(14):4845–57, 2007. doi: 10.1093/nar/gkm485.

Michelle T Dimon, Katherine Sorber, and Joseph L DeRisi. HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. *PLoS One*, 5(11):e13875, 2010. doi: 10.1371/journal.pone.0013875.

William G. Fairbrother, Ru-Fang F. Yeh, Phillip A. Sharp, and Christopher B. Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297 (5583):1007–13, 8 2002. ISSN 1095-9203. doi: 10.1126/science.1073774.

A. Fedorov, S. Saxonov, L. Fedorova, and I. Daizadeh. Comparison of intron-containing and intron-lacking human genes elucidates putative exonic splicing enhancers. *Nucleic Acids Research*, 29(7):1464, 2001. ISSN 0305-1048.

L. Fedorova and A. Fedorov. Puzzles of the human genome: Why do we need our introns? *Current Genomics*, 6:589–595, 12 2005. ISSN 1389-2029. doi: 10.2174/138920205775811416.

Robert D Finn, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E Pollington, O Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L L Sonnhammer, Sean R Eddy, and Alex Bateman. The Pfam protein families database. *Nucleic Acids Res*, 38(Database issue):D211–22, Jan 2010. doi: 10.1093/nar/gkp985.

Paul Flicek, Bronwen L. Aken, Benoit Ballester, Kathryn Beal, Eugene Bragin, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Julio Fernandez-Banet, Leo Gordon, Stefan Gräf, Syed Haider, Martin Hammond, Kerstin Howe, Andrew Jenkinson, Nathan Johnson, Andreas Kähäri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Felix Kokocinski, Gautier Koscielny, Eugene Kulesha, Daniel Lawson, Ian Longden, Tim Massingham, William McLaren, Karine Megy, Bert Overduin, Bethan Pritchard, Daniel Rios, Magali Ruffier, Michael Schuster, Guy Slater, Damian Smedley, Giulietta Spudich, Y. Amy Tang, Stephen Trevanion, Albert Vilella, Jan Vogel, Simon White, Steven P. Wilder, Amonida Zadissa, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé M. Fernández-Suarez, Javier Herrero, Tim J. P. Hubbard, Anne Parker, Glenn Proctor, James Smith, and Stephen M. J. Searle. Ensembl's 10th year. *Nucleic Acids Research*, 38(suppl 1):D557–D562, 2010. doi: 10.1093/nar/gkp972.

Alexandre Gattiker, Elisabeth Gasteiger, and Amos Bairoch. ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl Bioinformatics*, 1(2):107–8, 2002.

Amir Goren, Oren Ram, Maayan Amit, Hadas Keren, Galit Lev-Maor, Ida Vig, Tal Pupko, and Gil Ast. Comparative analysis identifies exonic splicing regulatory sequences–the

complex definition of enhancers and silencers. *Mol Cell*, 22(6):769–81, 6 2006. ISSN 1097-2765. doi: 10.1016/j.molcel.2006.05.008.

Brenton R. Graveley, Angela N. Brooks, Joseph W. Carlson, Michael O. Duff, Jane M. Landolin, Li Yang, Carlo G. Artieri, Marijke J. van Baren, Nathan Boley, Benjamin W. Booth, James B. Brown, Lucy Cherbas, Carrie A. Davis, Alex Dobin, Renhua Li, Wei Lin, John H. Malone, Nicolas R. Mattiuzzo, David Miller, David Sturgill, Brian B. Tuch, Chris Zaleski, Dayu Zhang, Marco Blanchette, Sandrine Dudoit, Brian Eads, Richard E. Green, Ann Hammonds, Lichun Jiang, Phil Kapranov, Laura Langton, Norbert Perrimon, Jeremy E. Sandler, Kenneth H. Wan, Aarron Willingham, Yu Zhang, Yi Zou, Justen Andrews, Peter J. Bickel, Steven E. Brenner, Michael R. Brent, Peter Cherbas, Thomas R. Gingeras, Roger A. Hoskins, Thomas C. Kaufman, Brian Oliver, and Susan E. Celniker. The developmental transcriptome of Drosophila melanogaster. *Nature*, 471(7339):473–479, 03 2011. URL `http://dx.doi.org/10.1038/nature09715`.

Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, John L Rinn, Eric S Lander, and Aviv Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol*, 28(5):503–10, May 2010. doi: 10.1038/nbt.1633.

Matthew W. Hahn and Gregory A. Wray. The g-value paradox. *Evol Dev*, 4(2):73–5, 2002. ISSN 1520-541X.

Martin J. Hicks, William F. Mueller, Peter J. Shepard, and Klemens J. Hertel. Competing upstream 5' splice sites enhance the rate of proximal splicing. *Mol Cell Biol*, 30(8): 1878–86, 4 2010. ISSN 1098-5549. doi: 10.1128/MCB.01071-09.

Nicolas Hulo, Amos Bairoch, Virginie Bulliard, Lorenzo Cerutti, Edouard De Castro, Petra S Langendijk-Genevaux, Marco Pagni, and Christian J A Sigrist. The PROSITE database. *Nucleic Acids Res*, 34(Database issue):D227–30, Jan 2006. doi: 10.1093/nar/gkj063.

Eric Jain, Amos Bairoch, Severine Duvaud, Isabelle Phan, Nicole Redaschi, Baris E Suzek, Maria J Martin, Peter McGarvey, and Elisabeth Gasteiger. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics*, 10: 136, 2009. doi: 10.1186/1471-2105-10-136.

Géraldine Jean, André Kahles, Vipin T Sreedharan, Fabio De Bona, and Gunnar Rätsch. RNA-seq read alignments with PALMapper. *Curr Protoc Bioinformatics*, Chapter 11: Unit 11.6, Dec 2010. doi: 10.1002/0471250953.bi1106s32.

Nejc Jelen, Jernej Ule, Marko Zivin, and Robert B. Darnell. Evolution of Nova-dependent splicing regulation in the brain. *PLoS Genet*, 3(10):1838–47, 10 2007. ISSN 1553-7404. doi: 10.1371/journal.pgen.0030173.

K. B. Jensen, K. Musunuru, H. A. Lewis, S. K. Burley, and R. B. Darnell. The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain. *Proc Natl Acad Sci U S A*, 97(11):5740–5, 5 2000. ISSN 0027-8424. doi: 10.1073/pnas.090553997.

Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–32, Apr 2009. doi: 10.1093/bioinformatics/btp113.

Jennifer L. Kabat, Sergio Barberan-Soler, Paul McKenna, Hiram Clawson, Tracy Farrer, and Alan M. Zahler. Intronic alternative splicing regulators identified by comparative genomics in nematodes. *PLoS Comput Biol*, 2(7):e86, 7 2006. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0020086.

Jennifer L. Kabat, Sergio Barberan-Soler, and Alan M. Zahler. HRP-2, the Caenorhabditis elegans homolog of mammalian heterogeneous nuclear ribonucleoproteins Q and R, is an alternative splicing factor that binds to UCUAUC splicing regulatory elements. *J Biol Chem*, 284(42):28490–7, 10 2009. ISSN 1083-351X. doi: 10.1074/jbc.M109.023101.

Auinash Kalsotra, Xinshu Xiao, Amanda J. Ward, John C. Castle, Jason M. Johnson, Christopher B. Burge, and Thomas A. Cooper. A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. *Proc Natl Acad Sci U S A*, 105(51):20333–8, 12 2008. ISSN 1091-6490. doi: 10.1073/pnas.0809045105.

Yarden Katz, Eric T. Wang, Edoardo M. Airoldi, and Christopher B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*, 7(12):1009–15, 12 2010. ISSN 1548-7105. doi: 10.1038/nmeth.1528.

Tadashi Kawashima, Matteo Pellegrini, and Guillaume F Chanfreau. Nonsense-mediated mRNA decay mutes the splicing defects of spliceosome component mutations. *RNA*, 15 (12):2236–47, Dec 2009. doi: 10.1261/rna.1736809.

W James Kent. BLAT–the BLAST-like alignment tool. *Genome Res*, 12(4):656–64, Apr 2002. doi: 10.1101/gr.229202. Article published online before March 2002.

W. James Kent, Robert Baertsch, Angie Hinrichs, Webb Miller, and David Haussler. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A*, 100(20):11484–9, 9 2003. ISSN 0027-8424. doi: 10.1073/pnas.1932072100.

Julian König, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J. Turner, Nicholas M. Luscombe, and Jernej Ule. iCLIP reveals the function

of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*, 17(7):909–15, 7 2010. ISSN 1545-9985. doi: 10.1038/nsmb.1838.

E. Koren, G. Lev-Maor, and G. Ast. The emergence of alternative 39 and 59 splice site exons from constitutive exons. *PLoS Comput Biol*, 3(5):e95, 2007. ISSN 1553-7358.

Alberto R Kornblihtt. Chromatin, transcript elongation and alternative splicing. *Nat Struct Mol Biol*, 13(1):5–7, Jan 2006. doi: 10.1038/nsmb0106-5.

Tracy L Kress, Nevan J Krogan, and Christine Guthrie. A single SR-like protein, Npl3, promotes pre-mRNA splicing in budding yeast. *Mol Cell*, 32(5):727–34, Dec 2008. doi: 10.1016/j.molcel.2008.11.013.

Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3): R25, 2009. ISSN 1465-6914. doi: 10.1186/gb-2009-10-3-r25.

M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–8, 11 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm404.

Bernett T. K. Lee, Tin Wee Tan, and Shoba Ranganathan. DEDB: a database of Drosophila melanogaster exons in splicing graph form. *BMC Bioinformatics*, 5:189, 12 2004. ISSN 1471-2105. doi: 10.1186/1471-2105-5-189.

Ivica Letunic, Tobias Doerks, and Peer Bork. SMART 6: recent updates and new developments. *Nucleic Acids Res*, 37(Database issue):D229–32, Jan 2009. doi: 10.1093/nar/gkn808.

Michael Levine and Robert Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945):147–51, Jul 2003. doi: 10.1038/nature01763.

H. A. Lewis, H. Chen, C. Edo, R. J. Buckanovich, Y. Y. Yang, K. Musunuru, R. Zhong, R. B. Darnell, and S. K. Burley. Crystal structures of Nova-1 and Nova-2 K-homology RNA-binding domains. *Structure*, 7(2):191–203, 2 1999. ISSN 0969-2126.

Heng Li, Avril Coghlan, Jue Ruan, Lachlan James Coin, Jean-Karim K. Hériché, Lara Osmotherly, Ruiqiang Li, Tao Liu, Zhang Zhang, Lars Bolund, Gane Ka-Shu Wong, Weimou Zheng, Paramvir Dehal, Jun Wang, and Richard Durbin. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*, 34(Database issue):D572–80, 1 2006. ISSN 1362-4962. doi: 10.1093/nar/gkj118.

Donny D. Licatalosi, Aldo Mele, John J. Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A. Clark, Anthony C. Schweitzer, John E. Blume, Xuning Wang, Jennifer C.

Darnell, and Robert B. Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–9, 11 2008. ISSN 1476-4687. doi: 10.1038/nature07488.

L. P. Lim and C. B. Burge. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A*, 98(20):11193–8, 9 2001. ISSN 0027-8424. doi: 10.1073/pnas.201407298.

H. X. Liu, M. Zhang, and A. R. Krainer. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes & development*, 12(13), 1998. ISSN 0890-9369. doi: 10.1101/gad.12.13.1998.

Miriam Llorian, Schraga Schwartz, Tyson A. Clark, Dror Hollander, Lit-Yeen Y. Tan, Rachel Spellman, Adele Gordon, Anthony C. Schweitzer, Pierre de la Grange, Gil Ast, and Christopher W. J. Smith. Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nat Struct Mol Biol*, 17(9):1114–23, 9 2010. ISSN 1545-9985. doi: 10.1038/nsmb.1881.

J Robert Manak, Sujit Dike, Victor Sementchenko, Philipp Kapranov, Frederic Biemar, Jeff Long, Jill Cheng, Ian Bell, Srinka Ghosh, Antonio Piccolboni, and Thomas R Gingeras. Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nat Genet*, 38(10):1151–8, Oct 2006. doi: 10.1038/ng1875.

Jeffrey Martin, Vincent M Bruno, Zhide Fang, Xiandong Meng, Matthew Blow, Tao Zhang, Gavin Sherlock, Michael Snyder, and Zhong Wang. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics*, 11: 663, 2010. doi: 10.1186/1471-2164-11-663.

Arianne J. Matlin, Francis Clark, and Christopher W. J. Smith. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*, 6(5):386–98, 5 2005. ISSN 1471-0072. doi: 10.1038/nrm1645.

C Joel McManus, Michael O Duff, Jodi Eipper-Mains, and Brenton R Graveley. Global analysis of trans-splicing in Drosophila. *Proc Natl Acad Sci U S A*, 107(29):12975–9, Jul 2010. doi: 10.1073/pnas.1007586107.

Antonio C. A. Meireles-Filho and Alexander Stark. Comparative genomics of gene regulation-conservation and divergence of cis-regulatory information. *Curr Opin Genet Dev*, 19(6):565–70, 12 2009. ISSN 1879-0380. doi: 10.1016/j.gde.2009.10.006.

Ali Mortazavi, Brian Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5(7):621–628, 5 2008. ISSN 1548-7105.

Tobias Mourier and Daniel C Jeffares. Eukaryotic intron loss. *Science*, 300(5624):1393, May 2003. doi: 10.1126/science.1080559.

Timothy W. Nilsen and Brenton R. Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–63, 1 2010. ISSN 1476-4687. doi: 10.1038/nature08909.

Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–5, 12 2008. ISSN 1546-1718. doi: 10.1038/ng.259.

Jung W. Park and Brenton R. Graveley. Use of RNA interference to dissect the roles of trans-acting factors in alternative pre-mRNA splicing. *Methods*, 37(4):341–4, 12 2005. ISSN 1046-2023. doi: 10.1016/j.ymeth.2005.07.016.

Jung W. Park, Katherine Parisky, Alicia M. Celotto, Robert A. Reenan, and Brenton R. Graveley. Identification of alternative splicing regulators by RNA interference in Drosophila. *Proc Natl Acad Sci U S A*, 101(45):15974–9, 11 2004. ISSN 0027-8424. doi: 10.1073/pnas.0407004101.

I. Pérez, J. G. McAfee, and J. G. Patton. Multiple RRMs contribute to RNA binding specificity and affinity for polypyrimidine tract binding protein. *Biochemistry*, 36(39): 11881–90, 9 1997. ISSN 0006-2960. doi: 10.1021/bi9711745.

Mihaela Pertea, Stephen M. Mount, and Steven L. Salzberg. A computational survey of candidate exonic splicing enhancer motifs in the model plant Arabidopsis thaliana. *BMC Bioinformatics*, 8:159, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-159.

Dana L Philipps, Jung W Park, and Brenton R Graveley. A computational and experimental approach toward a priori identification of alternatively spliced exons. *RNA*, 10(12):1838–44, Dec 2004. doi: 10.1261/rna.7136104.

Jeffrey A Pleiss, Gregg B Whitworth, Megan Bergkessel, and Christine Guthrie. Transcript specificity in yeast pre-mRNA splicing revealed by mutations in core spliceosomal components. *PLoS Biol*, 5(4):e90, Apr 2007. doi: 10.1371/journal.pbio.0050090.

Jiang Qian, Jimmy Lin, Nicholas M Luscombe, Haiyuan Yu, and Mark Gerstein. Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, 19(15):1917–26, Oct 2003.

Brooke Rhead, Donna Karolchik, Robert M. Kuhn, Angie S. Hinrichs, Ann S. Zweig, Pauline A. Fujita, Mark Diekhans, Kayla E. Smith, Kate R. Rosenbloom, Brian J. Raney, Andy Pohl, Michael Pheasant, Laurence R. Meyer, Katrina Learned, Fan Hsu, Jennifer Hillman-Jackson, Rachel A. Harte, Belinda Giardine, Timothy R. Dreszer, Hiram Clawson, Galt P. Barber, David Haussler, and W. James Kent. The UCSC Genome Browser database: update 2010. *Nucleic Acids Res*, 38(Database issue):D613–9, 1 2010. ISSN 1362-4962. doi: 10.1093/nar/gkp939.

Mark Robida, Vinod Sridharan, Sheridan Morgan, Timsi Rao, and Ravinder Singh. Drosophila polypyrimidine tract-binding protein is necessary for spermatid individualization. *Proc Natl Acad Sci U S A*, 107(28):12570–5, 7 2010. ISSN 1091-6490. doi: 10.1073/pnas.1007935107.

Xavier Roca, Ravi Sachidanandam, and Adrian R. Krainer. Determinants of the inherent strength of human 5' splice sites. *RNA*, 11(5):683–98, 5 2005. ISSN 1355-8382. doi: 10.1261/rna.2040605.

C. M. Romfo, C. J. Alvarez, W. J. van Heeckeren, C. J. Webb, and J. A. Wise. Evidence for splice site pairing via intron definition in Schizosaccharomyces pombe. *Mol Cell Biol*, 20(21):7955–70, 11 2000. ISSN 0270-7306.

Tanja Dorothe Rösel, Lee-Hsueh H. Hung, Jan Medenbach, Katrin Donde, Stefan Starke, Vladimir Benes, Gunnar Rätsch, and Albrecht Bindereif. RNA-seq analysis in mutant zebrafish reveals role of U1C protein in alternative splicing regulation. *EMBO J*, 4 2011. ISSN 1460-2075. doi: 10.1038/emboj.2011.106.

Scott W. Roy. Intron-rich ancestors. *Trends Genet*, 22(9):468–71, 9 2006. ISSN 0168-9525. doi: 10.1016/j.tig.2006.07.002.

Jue Ruan, Heng Li, Zhongzhong Chen, Avril Coghlan, Lachlan James M. Coin, Yiran Guo, Jean-Karim K. Hériché, Yafeng Hu, Karsten Kristiansen, Ruiqiang Li, Tao Liu, Alan Moses, Junjie Qin, Søren Vang, Albert J. Vilella, Abel Ureta-Vidal, Lars Bolund, Jun Wang, and Richard Durbin. Treefam: 2008 update. *Nucleic Acids Res*, 36(Database issue):D735–40, 1 2008. ISSN 1362-4962. doi: 10.1093/nar/gkm1005.

Lucas Sánchez. Sex-determining mechanisms in insects. *Int J Dev Biol*, 52(7):837–56, 2008. doi: 10.1387/ijdb.072396ls.

Jeremy Sanford, Pedro Coutinho, Jamie Hackett, Xin Wang, William Ranahan, and Javier Caceres. Identification of nuclear and cytoplasmic mRNA targets for the shuttling protein SF2/ASF. *PLoS ONE*, 3(10):e3369–, 10 2008. ISSN 1932-6203.

Jeremy R Sanford, Xin Wang, Matthew Mort, Natalia Vanduyn, David N Cooper, Sean D Mooney, Howard J Edenberg, and Yunlong Liu. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res*, 19(3):381–94, Mar 2009. doi: 10.1101/gr.082503.108.

T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–100, 10 1990. ISSN 0305-1048.

Schraga Schwartz and Gil Ast. Chromatin density and splicing destiny: on the cross-talk between chromatin structure and splicing. *EMBO J*, 29(10):1629–36, 5 2010. ISSN 1460-2075. doi: 10.1038/emboj.2010.71.

Schraga H. Schwartz, João Silva, David Burstein, Tal Pupko, Eduardo Eyras, and Gil Ast. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res*, 18(1):88–103, 1 2008. ISSN 1088-9051. doi: 10.1101/gr.6818908.

Scott Schwartz, W. James Kent, Arian Smit, Zheng Zhang, Robert Baertsch, Ross C. Hardison, David Haussler, and Webb Miller. Human-mouse alignments with BLASTZ. *Genome Res*, 13(1):103–7, 1 2003. ISSN 1088-9051. doi: 10.1101/gr.809403.

P. Seshaiah, B. Miller, M. M. Myat, and D. J. Andrew. pasilla, the Drosophila homologue of the human Nova-1 and Nova-2 proteins, is required for normal secretion in the salivary gland. *Dev Biol*, 239(2):309–22, 11 2001. ISSN 0012-1606. doi: 10.1006/dbio.2001.0429.

H. Shi, B. E. Hoffman, and J. T. Lis. A specific RNA hairpin loop structure binds the RNA recognition motifs of the Drosophila SR protein B52. *Molecular and cellular biology*, 17(5):2649, 1997. ISSN 0270-7306.

Noam Shomron, David Golan, and Eran Hornstein. An evolutionary perspective of animal microRNAs and their targets. *J Biomed Biotechnol*, 2009:594738, 2009. ISSN 1110-7251. doi: 10.1155/2009/594738.

Adam Siepel, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W. Hillier, Stephen Richards, George M. Weinstock, Richard K. Wilson, Richard A. Gibbs, W. James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–50, 8 2005. ISSN 1088-9051. doi: 10.1101/gr.3715005.

Philip J. Smith, Chaolin Zhang, Jinhua Wang, Shern L. Chew, Michael Q. Zhang, and Adrian R. Krainer. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum Mol Genet*, 15(16):2490–508, 8 2006. ISSN 0964-6906. doi: 10.1093/hmg/ddl171.

Viktor Stolc, Zareen Gauhar, Christopher Mason, Gabor Halasz, Marinus F van Batenburg, Scott A Rifkin, Sujun Hua, Tine Herreman, Waraporn Tongprasit, Paolo Emilio Barbano, Harmen J Bussemaker, and Kevin P White. A gene expression map for the euchromatic genome of Drosophila melanogaster. *Science*, 306(5696):655–60, Oct 2004. doi: 10.1126/science.1101312.

Charles W Sugnet, Karpagam Srinivasan, Tyson A Clark, Georgeann O'Brien, Melissa S Cline, Hui Wang, Alan Williams, David Kulp, John E Blume, David Haussler, and Manuel Ares, Jr. Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Comput Biol*, 2(1):e4, Jan 2006. doi: 10.1371/journal.pcbi.0020004.

J. Matthew Taliaferro, Nehemiah Alvarez, Richard E. Green, Marco Blanchette, and Donald C. Rio. Evolution of a tissue-specific splicing network. *Genes Dev*, 25(6): 608–20, 3 2011. ISSN 1549-5477. doi: 10.1101/gad.2009011.

Cole Trapnell, Lior Pachter, and Steven L. Salzberg. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics*, 25(9):1105–11, 5 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp120.

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–5, May 2010. doi: 10.1038/nbt.1621.

Susan Tweedie, Michael Ashburner, Kathleen Falls, Paul Leyland, Peter McQuilton, Steven Marygold, Gillian Millburn, David Osumi-Sutherland, Andrew Schroeder, Ruth Seal, Haiyan Zhang, and FlyBase Consortium. Flybase: enhancing Drosophila gene ontology annotations. *Nucleic Acids Res*, 37(Database issue):D555–9, 1 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn788.

Jernej Ule, Aljaz Ule, Joanna Spencer, Alan Williams, Jing-Shan S. Hu, Melissa Cline, Hui Wang, Tyson Clark, Claire Fraser, Matteo Ruggiu, Barry R. Zeeberg, David Kane, John N. Weinstein, John Blume, and Robert B. Darnell. Nova regulates brain-specific splicing to shape the synapse. *Nat Genet*, 37(8):844–52, 8 2005. ISSN 1061-4036. doi: 10.1038/ng1610.

Jernej Ule, Giovanni Stefani, Aldo Mele, Matteo Ruggiu, Xuning Wang, Bahar Taneri, Terry Gaasterland, Benjamin J. Blencowe, and Robert B. Darnell. An RNA map predicting Nova-dependent splicing regulation. *Nature*, 444(7119):580–6, 11 2006. ISSN 1476-4687. doi: 10.1038/nature05304.

UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res*, 39(Database issue):D214–9, Jan 2011. doi: 10.1093/nar/gkq1020.

Rodger B. Voelker and J. Andrew Berglund. A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing. *Genome Res*, 17(7):1023–33, 7 2007. ISSN 1088-9051. doi: 10.1101/gr.6017807.

Eric Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen Kingsmore, Gary Schroth, and Christopher Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 11 2008. ISSN 0028-0836.

Kai Wang, Darshan Singh, Zheng Zeng, Stephen J Coleman, Yan Huang, Gleb L Savich, Xiaping He, Piotr Mieczkowski, Sara A Grimm, Charles M Perou, James N MacLeod,

Derek Y Chiang, Jan F Prins, and Jinze Liu. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*, 38(18):e178, Oct 2010. doi: 10.1093/nar/gkq622.

Q Tian Wang, Karolina Piotrowska, Maria Anna Ciemerych, Ljiljana Milenkovic, Matthew P Scott, Ronald W Davis, and Magdalena Zernicka-Goetz. A genome-wide study of gene activity reveals developmental signaling pathways in the preimplantation mouse embryo. *Dev Cell*, 6(1):133–44, Jan 2004a.

Zefeng Wang and Christopher B. Burge. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(5):802–13, 5 2008. ISSN 1469-9001. doi: 10.1261/rna.876308.

Zefeng Wang, Michael E. Rolish, Gene Yeo, Vivian Tung, Matthew Mawson, and Christopher B. Burge. Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6):831–45, 12 2004b. ISSN 0092-8674. doi: 10.1016/j.cell.2004.11.010.

Zefeng Wang, Xinshu Xiao, Eric Van Nostrand, and Christopher B Burge. General and specific functions of exonic splicing silencers in splicing control. *Mol Cell*, 23(1):61–70, Jul 2006. doi: 10.1016/j.molcel.2006.05.018.

Claude C Warzecha, Peng Jiang, Karine Amirikian, Kimberly A Dittmar, Hezhe Lu, Shihao Shen, Wei Guo, Yi Xing, and Russ P Carstens. An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. *EMBO J*, 29(19):3286–300, Oct 2010. doi: 10.1038/emboj.2010.195.

Thomas D Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–81, Apr 2010. doi: 10.1093/bioinformatics/btq057.

Xinshu Xiao, Zefeng Wang, Minyoung Jang, and Christopher B. Burge. Coevolutionary networks of splicing cis-regulatory elements. *Proc Natl Acad Sci U S A*, 104(47):18583–8, 11 2007. ISSN 1091-6490. doi: 10.1073/pnas.0707349104.

Yuanchao Xue, Yu Zhou, Tongbin Wu, Tuo Zhu, Xiong Ji, Young-Soo S. Kwon, Chao Zhang, Gene Yeo, Douglas L. Black, Hui Sun, Xiang-Dong D. Fu, and Yi Zhang. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol Cell*, 36(6):996–1006, 12 2009. ISSN 1097-4164. doi: 10.1016/j.molcel.2009.12.003.

Gene Yeo and Christopher B. Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*, 11(2-3):377–94, 2004. ISSN 1066-5277. doi: 10.1089/1066527041410418.

Gene Yeo, Shawn Hoon, Byrappa Venkatesh, and Christopher B. Burge. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A*, 101(44):15700–5, 11 2004. ISSN 0027-8424. doi: 10.1073/pnas.0404901101.

Gene W. Yeo, Eric L. Van Nostrand, Eric L. Van Nostrand, and Tiffany Y. Liang. Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genet*, 3(5):e85, 5 2007. ISSN 1553-7404. doi: 10.1371/journal.pgen.0030085.

Gene W. Yeo, Nicole G. Coufal, Tiffany Y. Liang, Grace E. Peng, Xiang-Dong D. Fu, and Fred H. Gage. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol*, 16(2):130–7, 2 2009. ISSN 1545-9985. doi: 10.1038/nsmb.1545.

Chaolin Zhang, Zuo Zhang, John Castle, Shuying Sun, Jason Johnson, Adrian R. Krainer, and Michael Q. Zhang. Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. *Genes Dev*, 22(18):2550–63, 9 2008. ISSN 0890-9369. doi: 10.1101/gad.1703108.

Chaolin Zhang, Maria A. Frias, Aldo Mele, Matteo Ruggiu, Taesun Eom, Christina B. Marney, Huidong Wang, Donny D. Licatalosi, John J. Fak, and Robert B. Darnell. Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science*, 329(5990):439–43, 7 2010. ISSN 1095-9203. doi: 10.1126/science.1191150.

Xiang H-F Zhang and Lawrence A. Chasin. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*, 18(11):1241–50, 6 2004. ISSN 0890-9369. doi: 10.1101/gad.1195304.

Xiang H-F Zhang, Mauricio A. Arias, Shengdong Ke, and Lawrence A. Chasin. Splicing of designer exons reveals unexpected complexity in pre-mRNA splicing. *RNA*, 15(3): 367–76, 3 2009. ISSN 1469-9001. doi: 10.1261/rna.1498509.