Global Analyses of Alternative Splicing in Evolution and Nervous System Development

by

John Anthony Calarco

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Department of Molecular Genetics
University of Toronto

© Copyright by John Anthony Calarco 2011
Global Analyses of Alternative Splicing in Evolution
and Nervous System Development

John Anthony Calarco

Doctor of Philosophy

Department of Molecular Genetics
University of Toronto

2011

Abstract

Technological advancements have sparked discovery in biology, enabling important questions to be addressed experimentally at unprecedented depth and scale. One such advance, the development of large-scale approaches to study gene expression, has transformed the way we view the transcriptome. In recent years, these approaches have been applied to studies of alternative RNA splicing, a process where multiple distinct messenger RNAs can be generated from precursor transcripts to produce extensive transcriptomic diversity from a limited repertoire of genes. Global analyses have not only reinforced models initially based on single gene studies, they have also led to numerous insights into general principles governing the regulation and evolution of alternative splicing. In this thesis, I describe how I have combined both large-scale and focused approaches to study alternative splicing regulation during development and in an evolutionary context. Using microarray profiling and comparative genomics approaches, I describe the first large-scale comparative analysis of alternative splicing patterns between humans and chimpanzees. Next, I describe the discovery of a novel neural-specific RS domain splicing factor and the network of alternative exons it regulates to
promote nervous system development in vertebrates. Finally, I describe the profiling of alternative splicing patterns during *C. elegans* development using splicing microarrays and high-throughput sequencing. In this latter study, I also describe two resources that facilitate the analysis of tissue- or cell type-specific splicing events, and enable the function of isoforms to be assessed *in vivo*. Collectively, these studies have shed light on how differential regulation of alternative splicing has contributed to the evolution of complexity and diversity in biological systems.
Acknowledgments

A number of factors are required to shape a truly exceptional experience during graduate studies. In my opinion however, one of the most important factors is the strong and tireless support from the community of people surrounding a Ph.D. student.

I am indebted to my mentors Ben Blencowe and Mei Zhen for helping me grow as a biologist throughout these past few years. Their prudent guidance and input on my projects have taught me how to approach biological problems. Their immense creativity, drive, and passion for research have been infectious, igniting my own dedication to the scientific method, the pursuit of the unknown, and the training of future generations of scientists. They have also been great colleagues and friends, and the many conversations we have had, both scientific and non-scientific, will be fondly remembered. As I move into the next stage of my career, I know that I will continue to value their friendship and support.

I have been fortunate to have three exceptional committee members: Derek van der Kooy, Peter Roy, and Quaid Morris. Their input and advice concerning my research has been fantastic, and their unwaiving ability to withstand long and drawn-out committee meetings describing three projects has been appreciated. They have frequently gone beyond the normal requirements of most committee members, willing to discuss science outside of scheduled meetings, and being involved in several collaborative projects related to my own work. I am grateful for all of their support, encouragement, and help both in the lab and out.

I am also appreciative to have interacted with many talented people in the Zhen and Blencowe labs. It has been a lot of fun doing research in both groups, having get togethers with great food, and many other great times. In particular, Nick Watkins, Michelle Po, Calvin Mok, Joanne Kim, and Wesley Hung have been great colleagues at Mount Sinai and have provided much feedback and advice on my work. Additionally, Ying Wang and Hang Li have been incredibly helpful throughout my studies. At the Best Institute where I began my research journey, Susan McCracken and Emanuel Rosonina
who introduced me to experimental methods have been wonderful teachers and friends. At the Donnelly Centre, Joanna Ip, May Khanna, Deb Ray, and Yoseph Barash have been great people to discuss science and share memories with. I am lucky to have Christine Misquitta as a friend, who is a truly genuine person and has always looked out for me. Mathieu Gabut, who is an exceptional scientist and critical thinker, and a great person to have as a colleague both in and out of the lab. I would also like to thank Dave O’Hanlon, my partner in attempts of experimental world domination (we were so close...), for his wise and prudent approach to experiments (surpassed only by his knowledge of all things brewed and fermented) and friendship over the years. I have also had the pleasure of working with several undergraduate students who have now gone on to become the next generation of exciting young minds: Bushra Raj, Xinyi Tang, Valentina Slobodeniuc, and Chris Wedeles. I have learned many things from working with them and thank them for all the great experiences.

I would also like to thank the many collaborators I have interacted with during the course of my studies: Yoseph Barash, Ofer Shai, Leo Lee, and Brendan Frey from the PSI lab, whose computational expertise have complemented our experimental approaches time and time again leading to exciting science; Brenda Coles Takabe and Laura Clarke from the van der Kooy lab for their expertise in stem cells; Simone Superina, Danielle Gelinas, and Brian Ciruna for extensive help with analysis of nSR100 in zebrafish; Sepand Mavandadi and Quaid Morris for many stimulating interactions and the development of the PATA algorithm; Mario Caceres, Todd Preuss, Yi Xing, and Chris Lee for their help in the human and chimpanzee profiling project; and Arun Ramani and Andy Fraser for the excellent collaboration during the C. elegans project. Arun in particular has become a close friend and colleague whom I respect with the highest regard both academically and personally.

I would also like to thank friends and family for their support. To the Saltzman family, for their compassionate, caring nature, and for their love and encouragement. To my parents, and for everything they have done for me, words cannot properly express my appreciation. I would not be who I am today without their love, support and everything
that they have taught me. I owe much to my brother and his profound influence on me. His creative spirit, and more importantly his friendship, continues to be an inspiration.

I would like to especially thank my grandparents, to whom this work has been dedicated to from the moment it began years ago. Their bravery in coming to a new world, and the many sacrifices they made out of love for their family, have created the opportunities that I have today. I will never forget all that they have done, and will always keep my memories of them close to my heart.

Finally, I would like to thank Arneet, who has been the most important collaborator in my life. Together, we have shared many joyous occasions together, and supported and comforted each other through each of our challenges. It has been both a unique and amazing experience to have her in my life, and I look forward to our future adventures together.
Table of Contents

Acknowledgments ........................................................................................................ iv

List of Tables .................................................................................................................. xii

List of Figures ................................................................................................................ xiii

List of Appendices .......................................................................................................... xvi

Introduction .................................................................................................................. 1

1.1 The discovery of split genes and pre-mRNA splicing ................................................. 2
  1.1.1 Mechanism of pre-mRNA splicing ....................................................................... 3
  1.1.2 E and A complexes ............................................................................................. 4
  1.1.3 B, B*, and C complexes ..................................................................................... 6

1.2 Alternative splicing .................................................................................................... 8

1.3 Genome-wide approaches facilitating the study of AS .............................................. 9

1.4 Cis and trans determinants of alternative splicing .................................................... 13
  1.4.1 Core splicing signals .......................................................................................... 13
  1.4.2 Exonic and intronic splicing enhancers and silencers ....................................... 13
  1.4.3 Secondary structure ......................................................................................... 15
  1.4.4 Core spliceosomal proteins .............................................................................. 17
  1.4.5 SR and SR-related proteins ............................................................................. 18
  1.4.6 heterogeneous nuclear ribonucleoproteins ...................................................... 19
  1.4.7 RNA binding proteins and non-coding RNA .................................................. 21

1.5 AS regulatory networks and the nervous system ....................................................... 22
  1.5.1 Nova-1 and Nova-2 .......................................................................................... 23
  1.5.2 PTB and PTBP2 ............................................................................................... 25
  1.5.3 Fox splicing regulators ..................................................................................... 27
  1.5.4 Sam68 ............................................................................................................. 29
  1.5.5 Additional neural-specific splicing regulators .................................................. 31

1.6 Coordination of gene regulatory machineries ............................................................ 32

1.7 Evolutionary features of alternative splicing ............................................................ 34

1.8 Alternative splicing in C. elegans ............................................................................ 38
Global Analysis of Alternative Splicing Differences Between Humans and Chimpanzees

1.9 Overview of thesis research

2.1 Abstract

2.2 Introduction

2.3 Materials and Methods

2.3.1 Comparative genomic analysis of human and chimpanzee exons/flanking introns

2.3.2 Tissue samples, cell lines, and cell culture

2.3.3 Microarray hybridization, data extraction, and analysis

2.3.4 RT–PCR assays and quantification

2.3.5 Plasmids

2.3.6 Transfection and Western blotting

2.3.7 Microarray data analysis

2.3.8 Estimate for the percent of multi-exon human and chimpanzee genes with pronounced splicing level differences.

2.4 Results

2.4.1 Analysis of human and chimpanzee splicing patterns by comparative genomic and microarray profiling strategies

2.4.2 Role of nucleotide substitutions in generating differences between human and chimpanzee splicing patterns

2.4.3 Detection of alternative splicing differences between humans and chimpanzees by microarray profiling

2.4.4 Evidence for stabilizing selection pressure acting to preserve the majority of splicing levels of orthologous human, chimpanzee, and mouse alternative exons

2.4.5 Primarily nonoverlapping subsets of genes have evolved splicing and transcript level differences between humans and chimpanzees

2.4.6 Alternative splicing differences between humans and chimpanzees affect transcripts from genes associated with diverse cellular functions

2.4.7 Alternative splicing differences are consistent between human and chimpanzee individuals

2.4.8 Examples of validated human and chimpanzee alternative splicing differences

2.4.9 Expression of the functional splice variant of GSTO2 differs between humans and chimpanzees at the protein level

2.5 Discussion
Regulation of vertebrate nervous system alternative splicing and development by nSR100

3.1 Abstract ........................................................................................................... 75

3.2 Introduction ...................................................................................................... 76

3.3 Materials and Methods .................................................................................... 78
  3.3.1 Computational identification and characterization of RS domain proteins .... 78
  3.3.2 Expression profiling of computationally-defined mouse RS domain genes ... 79
  3.3.3 Tissue samples, cell lines and cell culture ..................................................... 79
  3.3.4 Lentivirus production, infection, and generation of stable cell lines ............. 80
  3.3.5 Plasmids ....................................................................................................... 80
  3.3.6 Genome-wide analyses of nSR100 alternative splicing regulation ............... 82
  3.3.7 RT-PCR assays ............................................................................................. 84
  3.3.8 Immunofluorescence, antibodies, and western blotting ............................. 84
  3.3.9 Expression and purification of recombinant PTB, nPTB, SF2/ASF and nSR100 .. 84
  3.3.10 RNA gel mobility shift assays .................................................................... 86
  3.3.11 Preparation of whole cell in vitro splicing extracts and in vitro splicing ....... 87
  3.3.12 In vitro and in vivo crosslinking and immunoprecipitation (CLIP) .............. 88
  3.3.13 Zebrafish embryo maintenance, imaging, in situ hybridization and injections 89
  3.3.14 Generation of probes for whole-mount RNA in situ hybridization ............. 89
  3.3.15 Morpholino injections, RNA injections, and oligonucleotide sequences ...... 90
  3.3.16 siRNA and minigene reporter transfections .............................................. 90

3.4 Results ............................................................................................................. 91
  3.4.1 A genome-wide screen identifies a vertebrate and neural-specific SR-related protein 91
  3.4.2 Distribution of nSR100 ................................................................................ 95
  3.4.3 nSR100 is required for neural cell differentiation ........................................ 97
  3.4.4 nSR100 regulates a network of brain-specific alternative exons ................. 100
  3.4.5 An nSR100 regulatory code ......................................................................... 104
  3.4.6 nSR100 promotes nPTB protein expression ............................................... 107
  3.4.7 nSR100 promotes neural-specific exon inclusion in vitro ............................ 113
  3.4.8 A role for nSR100 in nervous system development ..................................... 115

3.5 Discussion ....................................................................................................... 119
  3.5.1 A neural- and vertebrate-specific activator of alternative splicing ................ 119
  3.5.2 Global and local roles for nSR100 in neural-specific alternative splicing ....... 119
  3.5.3 Functions of an nSR100-regulated neural exon network ............................... 121
Resources for in vivo analysis of alternative splicing in Caenorhabditis elegans 123

4.1 Abstract 124

4.2 Introduction 125

4.3 Materials and Methods 127

4.3.1 Strains and maintenance 127
4.3.2 RNA extraction, purification, and cDNA synthesis 127
4.3.3 Microarray design, hybridization and analysis 127
4.3.4 cDNA labeling, microarray hybridization and analysis 128
4.3.5 Mapping and analysis of RNA-Seq data 128
4.3.6 Defining a set of annotated transcript variants in C. elegans 130
4.3.7 Filtering array events for analysis 130
4.3.8 Computational analysis of alternative exon properties 131
4.3.9 Estimates of frequency of AS and developmentally-regulated AS 132
4.3.10 Motif analysis 132
4.3.12 RT-PCR assays 133
4.3.13 Plasmid and fosmid constructs and microinjection 133
4.3.14 Microinjection and generation of integrated lines 134
4.3.15 Bichromatic reporter imaging 135
4.3.16 Synaptic development and behavioral assays 135

4.4 Results 136

4.4.1 Large-scale identification of splice variants in C. elegans by RNA-Seq and microarray profiling 136
4.4.2 Identification of novel splicing events 141
4.4.3 Genome-wide analysis of temporally regulated alternative exon usage 142
4.4.4 The C. elegans Splice Browser 146
4.4.5 Versatile bichromatic reporters reveal cell type-specific AS patterns 146
4.4.6 Functional investigations of splice variants using isoform-specific transgenes 152

4.5 Discussion 156

Conclusions and Future Directions 159

5.1 Further dissection of AS differences between humans and chimpanzees 160

5.2 Detailed investigations into the mechanism of regulation by nSR100, its regulatory network, and its role in nervous system development 161

5.3 Neuron- and developmental stage-specific splicing regulation in C. elegans 163
5.4 Promises and challenges ahead: tissue-specific splicing regulatory networks ________ 165
  5.4.1 Examining connections among splicing factors ____________________________ 165
  5.4.2 The generation of predictive models governing tissue-specific splicing__________ 166

5.3 Concluding remarks _______________________________________________________ 167

References _________________________________________________________________ 168
List of Tables

Table 2.1: Data from the comparative genomic analysis of orthologous exon regions between humans and chimpanzees. ......................................................53

Table 2.2: Genes associated with different cellular functions with experimentally validated splicing level differences between humans and chimpanzees .................61

Table 2.3: Splicing differences found to be human- or chimpanzee-specific from RT-PCR assays of macaque tissue RNA as an outgroup comparison. .........................63

Table 3.1: Overrepresented Gene Ontology (GO) terms associated with genes containing nSR100 regulated alternative exons. ...........................................104

Table A1.1: Relevant information pertaining to the human, chimpanzee and macaque specimens used in this study. .................................................................197
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Step-wise assembly of the spliceosome.</td>
<td>6</td>
</tr>
<tr>
<td>1.2</td>
<td>Technologies for the genome-wide analysis of alternative splicing.</td>
<td>12</td>
</tr>
<tr>
<td>1.3</td>
<td>Cis- and trans-determinants of alternative splicing.</td>
<td>15</td>
</tr>
<tr>
<td>2.1</td>
<td>Strategies used to identify alternative splicing differences between humans and chimpanzees.</td>
<td>52</td>
</tr>
<tr>
<td>2.2</td>
<td>Microarray profiling reveals the extent of alternative splicing differences between human and chimpanzee orthologous exons, and the degree of divergence between splicing patterns over different evolutionary time periods.</td>
<td>56</td>
</tr>
<tr>
<td>2.3</td>
<td>Transcript and AS level differences between humans and chimpanzees involve largely distinct subsets of genes.</td>
<td>59</td>
</tr>
<tr>
<td>2.4</td>
<td>Examples of alternative splicing differences between humans and chimpanzees confirmed by semiquantitative RT–PCR and sequencing.</td>
<td>62</td>
</tr>
<tr>
<td>2.5</td>
<td>Comparisons of AS levels in RNA samples from macaque, human, and chimpanzee individuals.</td>
<td>64</td>
</tr>
<tr>
<td>2.6</td>
<td>Alternative splicing patterns are similar in cell lines from individuals within a species but consistently differ between species.</td>
<td>66</td>
</tr>
<tr>
<td>2.7</td>
<td>Increased skipping of exon 4 in GSTO2 transcripts reduces overall levels of functional GSTO2 protein in humans compared with chimpanzees.</td>
<td>69</td>
</tr>
<tr>
<td>3.1</td>
<td>Identification and Expression Characteristics of nSR100</td>
<td>93</td>
</tr>
<tr>
<td>3.2</td>
<td>Multiple alignment of nSR100 ORF sequences from different vertebrate species.</td>
<td>94</td>
</tr>
<tr>
<td>3.3</td>
<td>nSR100 Is Required for Neural Cell Differentiation</td>
<td>96</td>
</tr>
<tr>
<td>3.4</td>
<td>Knockdown of SRm160 but not nSR100 dramatically affects Neuro2a cell numbers.</td>
<td>98</td>
</tr>
</tbody>
</table>
Figure 3.5 Knockdown of nSR100 impairs neurite outgrowth in Neuro2a cells 99

Figure 3.6: nSR100 Regulates a Network of Neural-Specific Alternative Splicing Events 101

Figure 3.7: nSR100-regulated splicing patterns are established prior to induction of neurite outgrowth in Neuro2a cells. 103

Figure 3.8: Elements of an nSR100 Regulatory Code 105

Figure 3.9: nSR100 Functions by Promoting nPTB Expression and by Binding to its Regulated Target Transcripts 109

Figure 3.10. Additional nSR100-regulated splicing events are also regulated by PTB and nPTB. 111

Figure 3.11: PTB and nPTB associate with Daam1 upstream intronic RNA in vitro 112

Figure 3.12: nSR100 alone or in combination with PTB or nPTB does not associate strongly with the Daam1 intronic regulatory sequence, but can overcome the repressive activity of nPTB in vivo. 113

Figure 3.13: nSR100 Regulates Vertebrate Nervous System and Sensory Organ Development 116

Figure 3.14: Model for nSR100-Dependent Regulation of Neural-Specific Alternative Splicing 121

Figure 4.1: Outline of RNA-seq and microarray approaches for profiling AS in C. elegans. 137

Figure 4.2: Identification of known and novel AS events 141

Figure 4.3: Identification of temporally regulated AS events. 143

Figure 4.4: AS events differentially regulated during development 144

Figure 4.5: Bichromatic reporters reveal cell type-specific AS 148

Figure 4.6: Bichromatic reporters identify complex AS patterns in the nervous system 150

Figure 4.7: unc-16 exon 14 is excluded in GABAergic motor neurons 151
Figure 4.8: Isoform-specific analysis of the cle-1 gene 153
Figure 4.9: Isoform-specific analysis of the mod-1 gene 155
Figure A2.2: Additional alternative splicing events regulated by nSR100. 201
Figure A2.3: Additional motifs significantly overrepresented in the introns flanking nSR100-regulated alternative exons. 202
Figure A2.4: The nucleotide sequence of mouse Daam1 exon 16 and 296 nucleotides of upstream intronic sequence. 203
Figure A2.5: ZnSR100 pre-mRNA splicing is blocked in embryos treated with splice blocking morpholino oligonucleotide (MOSpl). 203
Figure A2.6: Phenotype of nSR100 MOSpl morphant embryos. 204
Figure A2.7: Morphology and islet-1 expression in control and nSR100 MOSpl embryos at 24hpf. 204
Figure A2.8: Injection of ZnSR100 mRNA rescues neurite outgrowth defects in ZnSR100 MOSpl morphant embryos. 205
Figure A3.1: Analysis of how read depth affects junction identification 208
Figure A3.2: Novel AS events identified by RNA-Seq 209
Figure A3.3: Motifs identified in the upstream region of temporally-regulated AS events 210
Figure A3.4: Motifs identified in the downstream region of temporally-regulated AS events 211
List of Appendices

Appendix 1 ................................................................. 196
Appendix 2 ................................................................. 199
Appendix 3 ................................................................. 207
Appendix 4 ................................................................. 213
Chapter 1

Introduction

1.1 The discovery of split genes and pre-mRNA splicing

The 1950s and 1960s witnessed seminal discoveries in our molecular understanding of DNA and how information within the genetic code is interpreted. During this time, Francis Crick formulated the “central dogma” of molecular biology, which provided a framework for how genetic information flows between the key biological polymers DNA, RNA, and protein (Crick, 1958). Given the knowledge of the time, Crick was not far off in his model, which was later revised to account for the discovery of reverse transcription in the laboratories of Howard Temin and David Baltimore (Baltimore, 1970; Crick, 1970; Temin and Mizutani, 1970). The conceptualization of information flow presented in the central dogma has been important, because it presciently described the key areas of gene regulation research that investigators would pursue for the next several decades. One such area of interest involved a deeper understanding of how DNA was transcribed into messenger RNA (mRNA) and the relationship between the two nucleic acid polymers. A major surprise came in 1977 from researchers in the laboratories of Phillip Sharp and Richard Roberts. During studies visualizing adenoviral mRNAs hybridized to their complementary genomic DNA sequence by electron microscopy, it was observed that these viral mRNA sequences did not display a contiguous linear relationship with their corresponding DNA sequence (Berget et al., 1977; Chow et al., 1977). These results suggested that portions of a longer precursor mRNA (pre-mRNA) may be “spliced” together to generate the mature transcripts (Berget et al., 1977; Chow et al., 1977). Additional studies rapidly followed confirming that these initial findings were not isolated to viral transcripts; a number of eukaryotic genes were found to have protein-coding sequences that were intervened or split by non-coding segments of DNA (for example see Brack and Tonegawa, 1977; Jeffreys and Flavell, 1977). Shortly after their discovery, these intervening, non-coding sequences were defined as “introns”, and the corresponding coding sequences being spliced together as “exons” (Gilbert, 1978), terms that have persisted to this day. Since the initial discovery of pre-mRNA splicing, decades of research have shed light on the complex machinery regulating the precise cleavage and ligation of fragments of pre-mRNA, the near-ubiquity of splicing in eukaryotes, and how aberrant regulation of splicing can lead to various diseases. In the following sections, I briefly describe the mechanism of pre-mRNA splicing.
Mechanism of pre-mRNA splicing

Accuracy and fidelity of pre-mRNA splicing requires the dynamic assembly of a large ribonucleoprotein (RNP) complex known as the “spliceosome”, composed of hundreds of protein subunits and five U-rich small nuclear RNAs (U1, U2, U4, U5, and U6 snRNAs) (Nilsen, 2003). This dynamic RNP machinery undergoes significant compositional and structural rearrangements in order to catalyze the two sequential transesterification reactions required to excise introns and ligate neighboring exons (Wahl et al., 2009). These rearrangements are facilitated by the actions of protein kinases, phosphatases, and RNA helicases (Matlin and Moore, 2007). The core components of the spliceosome consist of the small nuclear ribonucleoprotein particles (snRNPs), composed of one or two (in the case of the pre-assembled U4/U6 di-snRNP) of the five snRNAs and associated protein subunits (Maniatis and Reed, 1987). Additionally, some metazoans and plants encode distinct snRNPs, called U11, U12, and U4atac/U6atac, that function analogously to the U1, U2, and U4/U6 snRNPs, respectively (Patel and Steitz, 2003). These snRNPs collectively assemble with the U5 snRNP to function in what is known as the “minor” spliceosome that recognizes splice sites distinct from those found in the canonical or “major” class introns (Sharp and Burge, 1997). A large body of work over the past few decades has suggested that both the snRNAs and their snRNP protein complement play a direct role in splicing catalysis (see Matlin and Moore, 2007; Wahl et al., 2009 for reviews). For instance, studies have shed light on similarities between snRNA:pre-mRNA interactions and group II self-splicing introns (Toor et al., 2008), and others have shown that protein-free extracts containing the snRNAs are sufficient to carry out splicing-like catalysis in vitro, albeit with low efficiency (Valadkhan and Manley, 2001; Valadkhan et al., 2009; Valadkhan et al., 2007). These results suggest that the spliceosome may function at least in part as a ribozyme, and that its current mode of action may have evolved from an RNA world ancestor, similar to the ribosome (Steitz, 2008). However, at present the precise contribution of protein subunits and snRNAs to splicing catalysis still remain unclear.

Studies with cell free extracts capable of in vitro splicing in addition to genetic approaches in yeast have provided important mechanistic insights into how pre-mRNA splicing occurs (Green, 1991; Ruby and Abelson, 1991). For instance, a large body of evidence supports a model where the spliceosome assembles on a pre-mRNA substrate in a step-wise manner (Matlin and Moore, 2007). The intermediates of this step-wise assembly have been identified and
characterized mainly by their differential migration and sedimentation during native gel electrophoresis, glycerol gradient centrifugation, and affinity purification of these complexes during *in vitro* splicing catalysis (Bindereif and Green, 1987; Grabowski and Sharp, 1986; Konarska and Sharp, 1986; Lamond et al., 1988). More recently, a fluorescence-based strategy analyzing single pre-mRNAs being spliced *in vitro*, and chromatin immunoprecipitation experiments using antibodies against snRNP components engaged on nascent pre-mRNA substrates, have provided additional support of this step-wise assembly model (Gornemann et al., 2005; Hoskins et al., 2011; Lacadie and Rosbash, 2005). However, there is also evidence, though not as extensive, indicating that the spliceosome can engage pre-mRNA in a pre-assembled state (Stevens et al., 2002). It is currently unknown whether this latter mode of spliceosomal assembly occurs less frequently than the step-wise mechanism, or is simply the result of differences in purification and assay conditions used. Below, I briefly describe key aspects of our current understanding of step-wise spliceosomal assembly and catalysis.

### 1.1.2 E and A complexes

The earliest step in spliceosome assembly, known as E complex formation, occurs in an ATP-independent manner and involves the association of U1 snRNP and several additional protein factors with all four pre-mRNA core splicing signals: the 5´ splice site, the branchpoint sequence (BPS), the poly-pyrimidine tract (PPT), and the 3´ splice site (Bennett et al., 1992; Michaud and Reed, 1991, 1993; Figure 1.1). The interaction between U1 snRNP and the 5´ splice site is mediated by base-pairing interactions between U1 snRNA and the pre-mRNA (Kramer et al., 1984; Mount et al., 1983; Seraphin and Rosbash, 1989). Splicing factor 1/branchpoint binding protein (SF1/BBP) binds the BPS, and the 65 kDa and 35 kDa subunits of the U2 auxiliary factor (U2AF65 and U2AF35) are associated with the PPT and 3´ splice site, respectively at this stage of assembly (Berglund et al., 1998; Kramer and Utans, 1991; Wu et al., 1999; Zamore et al., 1992). All of these interactions are assisted and strengthened by additional supporting proteins, including those with domains composed of alternating arginine and serine di-peptide repeats (RS domains; described in more detail below; Lin and Fu, 2007). The interactions formed in E complex on the pre-mRNA create the first contacts necessary to recognize exon-intron boundaries, and serve as scaffolds for additional interactions with other snRNPs and auxiliary proteins that will occur in subsequent steps.
After the assembly of E complex, U2 snRNP associates with the pre-mRNA in a U1 snRNP and ATP-dependent manner forming base pairing interactions between U2 snRNA and the BPS to establish the A complex (Barabino et al., 1990; Chiara et al., 1996; MacMillan et al., 1994; Query et al., 1994). The association between U2 snRNA and the pre-mRNA coincides with the dissociation of SF1 (Liu et al., 2001; Rutz and Seraphin, 1999) and is further stabilized by protein components in U2 snRNP interacting with the branchpoint adenosine and U2AF65 (Shen et al., 2004; Valcarcel et al., 1996). The interaction between U2 snRNP and the pre-mRNA is thought to cause the branchpoint adenosine to protrude or “bulge”, which will be important for the first catalytic step of splicing.
Figure 1.1: Step-wise assembly of the spliceosome.
A simplified version of current knowledge of spliceosome assembly, focusing primarily on ordered loading of snRNPs and a few other key factors. Sequences in the pre-mRNA are presented in their approximate locations relative to splice sites. Abbreviations and symbols are as follows: R, purine nucleotide; Y, pyrimidine nucleotide; PPT, poly-pyrimidine tract; U1, U1 snRNP; U2, U2 snRNP; U4, U4 snRNP; U5, U5 snRNP; U6, U6 snRNP; SF1, splicing factor 1 (branchpoint binding protein); U2AF, U2 auxiliary factor.

1.1.3 B, B*, and C complexes

After A complex assembly, a pre-associated tri-snRNP complex composed of U4/U6 and U5 interacts with the pre-mRNA along with other proteins to form B complex (Bindereif and Green, 1987; Cheng and Abelson, 1987; Konarska and Sharp, 1987; Figure 1.1). Approximately 50 additional proteins can be detected in B complex, and at this point all five snRNPs are associated with the pre-mRNA (Deckert et al., 2006). However, in B complex the spliceosome is not configured in a catalytically active form and must undergo several conformational
rearrangements. Key rearrangements that must occur include: 1) the unwinding of the base-paired U4 and U6 snRNAs and dissociation of U4 snRNP; 2) replacement of U1 snRNA contacts at the 5´ splice site with U6 snRNA contacts, and dissociation of the U1 snRNP; and 3) base-pair formation between U2 and U6 snRNAs. Several helicases are implicated in these rearrangements that generate what is known as B* complex (reviewed in Matlin and Moore, 2007). The base-paired U2 and U6 snRNAs along with the U5 snRNP protein Prp8 form what is believed to be the catalytic core of the spliceosome. At this stage, the spliceosome is ready to carry out the first transesterification reaction where the sugar backbone phosphodiester bond at the 5´ splice site is attacked by the 2´ hydroxyl group of the bulged branchpoint adenosine. This reaction generates a free 3´ hydroxyl group at the end of the upstream exon and the formation of a branched lariat structure in the intron.

The completion of the first catalytic step of splicing marks the formation of C complex (Figure 1.1). At this stage, the spliceosome is poised to complete the second catalytic step, where the 3´ hydroxyl group of the upstream exon attacks the phosphodiester bond at the 3´ splice site, ligating the two exons. However, once again catalysis cannot occur without additional conformational rearrangements facilitated by helicases and other factors. In particular, the branched lariat intermediate must be displaced from the active site allowing the 3´ splice site to be positioned in close proximity to the 3´ hydroxyl group of the upstream exon (Rhode et al., 2006; Umen and Guthrie, 1995). Once the second transesterification reaction is completed, a multi-protein assembly known as the exon junction complex is deposited on the transcript in the vicinity of the site where splicing has occurred (Tange et al., 2004). Interactions between the U2, U5, and U6 snRNPs and the pre-mRNA are destabilized at this point, allowing the snRNPs to be recycled and assembled in conformations amenable for another round of splicing. The excised intron lariat is debranched and digested to free nucleotides that are also recycled for future use.

The step-wise nature of spliceosome assembly and the use of combinatorial weak interactions to drive the precise biochemical reactions it catalyzes provide flexibility in regulatory potential. This flexibility allows splicing to be differentially regulated at multiple points, and has likely contributed to the evolution of alternative splicing.
1.2 Alternative splicing

Shortly after the discovery of pre-mRNA splicing, researchers speculated why protein-coding genes contained non-coding intervening segments. In a clairvoyant news and views article to Nature in 1978 entitled “Why gene in pieces”, Walter Gilbert hypothesized that the presence of introns could facilitate the rapid evolution of novel gene functions (Gilbert, 1978). For example, a point mutation leading to the complete exclusion of an exon from an mRNA transcript could potentially cause a large change in the function of the corresponding translated protein. Even more interesting, some mutations could lead to inefficient splicing such that multiple alternative splice variants could be produced. Each variant would be subject to selection pressures and have the potential to evolve diverse functions and be integrated into unique molecular pathways. Importantly, Gilbert’s predictions of alternative splicing (AS) suggested a departure from the one gene-one polypeptide hypothesis. He further speculated that AS could be differentially regulated in tissues and during development, indicating that it would serve an important role in differentiation. Experimental evidence for AS would soon follow from work in viruses (Berk and Sharp, 1978a, b for two examples). Shortly afterwards, in 1980, the labs of David Baltimore and Leroy Hood discovered that isotype switching of immunoglobulin proteins was controlled at the level of mRNA through alternative splice site selection, providing the first cellular examples of AS (Alt et al., 1980; Early et al., 1980; Rogers et al., 1980).

More than 30 years later, countless studies supporting many of Gilbert’s claims have documented the prevalence and importance of AS in metazoans. Extreme cases of AS have been uncovered, such as the thousands of isoforms generated by the Down syndrome cell adhesion molecule (Dscam) and neurexin gene loci in Drosophila and mammals, respectively (Missler and Sudhof, 1998; Park and Graveley, 2007). These examples highlight the power of AS as a mechanism for generating substantial proteomic diversity from a limited repertoire of genes. Many studies have also demonstrated that splicing can be controlled in a tissue- and developmental stage-specific manner (Calarco et al., 2011; Chen and Manley, 2009; Li et al., 2007). Additionally, as with other important gene regulatory processes, aberrant control of AS can lead to numerous diseases (Cooper et al., 2009). A significant amount of work has advanced our understanding of AS regulatory mechanisms and, in the past decade, genome-wide methodologies have rapidly led to further insights. In the sections to follow, I will highlight these key technologies and our current knowledge of AS regulation.
1.3 Genome-wide approaches facilitating the study of AS

Several microarray platforms developed in the past decade have enabled genome-scale analyses of pre-mRNA splicing. These platforms typically make use of combinations of short oligonucleotide probes that recognize exon and/or exon junction sequences (Ben-Dov et al., 2008; Calarco et al., 2007a; Moore and Silver, 2008; Figure 1.2A). These microarrays have been instrumental in the large-scale profiling of splicing efficiency in wild type and mutant strains of budding yeast (Clark et al., 2002; Pleiss et al., 2007), and in providing measurements of AS across tissues, during development, or upon disruption of splicing factor expression in metazoans (Blanchette et al., 2005; Boutz et al., 2007b; Calarco et al., 2009; Fagnani et al., 2007a; Hartmann et al., 2009; Johnson et al., 2003; Pan et al., 2004; Stolc et al., 2004; Sugnet et al., 2006; Ule et al., 2005b). A number of general principles of AS regulation have been deduced from these studies, including evolutionary features associated with classes of commonly regulated alternative exons, and comparisons of the dynamics and organization of AS and transcription networks (Blencowe, 2006; Calarco et al., 2007a).

More recently, high throughput sequencing (HTS) technologies have been employed to analyze polyadenylated transcripts at a remarkable degree of precision and sensitivity. In these approaches, cDNA fragments derived from mRNA populations are sequenced in a massively parallel manner (RNA-Seq), and aligned to reference transcriptomes (Pepke et al., 2009; Wang et al., 2009; Figure 1.2B). Recent RNA-Seq analyses profiling human cell lines and tissues reveal that greater than 95% of multi-exon genes have more than one splice variant, and these isoforms are frequently tissue-specific (Pan et al., 2008; Wang et al., 2008). RNA-Seq has the advantage over microarray-based approaches in that novel isoforms and splice junctions can be detected without any a priori knowledge, and cross-hybridization artifacts are not an issue. Currently, both microarray and RNA-Seq based approaches are actively used for genome-wide analysis of AS, but RNA-Seq is quickly becoming the method of choice (Blencowe et al., 2009; Wang et al., 2009).

Complementary to the approaches described above, AS regulatory networks can also be uncovered by identifying transcripts physically associated with specific splicing factors and ribonucleoprotein (RNP) complexes. In these approaches, an RNA binding protein and its associated RNP complex is immunoprecipitated from a cell lysate, followed by purification and
detection of any bound RNA transcripts. RNP complexes can either be purified under native conditions (Tenenbaum et al., 2000), or under more stringent conditions if a cross-linking method is employed (Niranjanakumari et al., 2002; Ule et al., 2003). The former approach has the advantage that all transcripts in an RNP complex can be identified. However, this approach runs the risk of loss of low-affinity but specific in vivo interactions and also transcript “reassociation”, in which transcripts that do not normally associate with an RNP complex bind during or following cell extraction (Mili and Steitz, 2004). Methods involving cross-linking circumvent this problem since stringent purification conditions can be used, but depending on the cross-linking reagent employed, may only afford detection of transcripts that are directly bound by the factor. In the most recent incarnations of the latter approaches, known as crosslinking and immunoprecipitation (CLIP), the associated RNA is partially digested during the protocol, leaving short associated fragments that can provide positional information of the physical association (Jensen and Darnell, 2008). Additional modifications to this protocol have further increased the resolution of this method (Hafner et al., 2010; Konig et al., 2010). Recently, these approaches have been coupled with microarray profiling or HTS to obtain genome-wide snapshots of the repertoire of transcripts and binding sites associated with specific splicing factors and other RNA binding proteins (Keene et al., 2006; Licatalosi and Darnell, 2010 for examples).

Computational and bioinformatic methods play an important role in the analysis of the large compendia of splicing events identified by the experimental approaches described above. Such analyses have facilitated the discovery of splicing enhancer and silencer elements associated with constitutively spliced and alternative exons (Fairbrother et al., 2002; Yeo et al., 2007; Zhang and Chasin, 2004). Computational approaches have also been successful in identifying cis-regulatory elements enriched in tissue-specific AS events (Fagnani et al., 2007a; Sugnet et al., 2006). More sophisticated methods have also begun to identify relationships between groups of cis-elements, other factors influencing regulated splicing, and insights into the evolutionary dynamics of AS (Barash et al., 2010; Friedman et al., 2008; Ke and Chasin, 2010; Keren et al., 2010; Xiao et al., 2007; Xing and Lee, 2006; Yeo et al., 2004b; Zhang et al., 2010). More recently, the utility of computational approaches has been further exemplified in the analysis of genome-wide chromatin immunoprecipitation (ChIP) datasets, which have revealed interesting connections between histone modifications and splicing (Huff et al., 2010;
Kolasinska-Zwierz et al., 2009; Schwartz et al., 2009b; Spies et al., 2009; Tilgner et al., 2009 and see below).

Collectively, the above techniques have become an indispensable set of tools in the analysis of AS. In the following section, I discuss how these and other experimental approaches have been utilized to investigate cis- and trans-acting factors regulating AS.
Figure 1.2: Technologies for the genome-wide analysis of alternative splicing

A. Microarray platform design used to measure exon inclusion levels of cassette type alternative splicing events. Six oligonucleotide probes (small black bars) are designed to monitor each cassette event. Three of these probes anneal to either the upstream constitutive exon (labeled C1), the downstream constitutive exon (labeled C2), and the alternative exon (labeled A). The other three probes anneal to exon junctions corresponding to the alternative exon-included isoform (represented by the C1:A and A:C2 probes) or the alternative exon-skipped isoform (represented by the C1:C2 probe). These probes are then printed on a microarray (image on right shows a hybridized microarray slide with arrows pointing to two spots containing specific probes.

B. Profiling alternative splicing by RNA-Seq. Massively parallel sequencing of complementary DNA (cDNA) libraries generated from fragmented poly A+ enriched mRNA populations leads to the production of short nucleotide sequence ‘reads’ (small black bars) that are then mapped to known gene models or de novo. For analysis of alternative splicing, reads mapping to exon sequences (depicted over the red and yellow exons) or exon junction sequences (displayed over dashed lines) can be used to identify and measure constitutive (when yellow exons are joined) and alternative splicing events (when red exons are included or skipped).
1.4 Cis and trans determinants of alternative splicing

1.4.1 Core splicing signals

Each of the four core splicing signals described above (5′ splice site, BPS, PPT, and 3′ splice site) play an important role in determining how efficiently the spliceosome will differentiate exon from intron sequences. In the budding yeast *Saccharomyces cerevisiae*, where virtually no AS occurs, these four signals rarely deviate from largely invariant consensus sequences that ensure optimal interactions with the snRNAs or proteins (Wahl et al., 2009). However, in metazoans and especially mammals, these signals are far more degenerate (Schwartz et al., 2008). The 5′ splice site can contain a number of mis-matches to the complementary U1 snRNA sequence and still be recognized accurately. The BPS consensus is much more loosely defined and its position relative to the 3′ splice site can be quite variable. The PPT sequence can also vary in its composition and length. The evolution of greater degeneracy in the core splicing signals in metazoans is speculated to have facilitated the emergence of AS (Ram and Ast, 2007). Significant deviations from preferred consensus sequences can lead to a competitive disadvantage for a weaker splice site to be recognized over a stronger splice site in a neighboring exon. In support of this model, alternatively spliced exons tend to have splice sites and PPTs that deviate more from consensus sequences than constitutively spliced exons, although there are exceptions to these trends (Itoh et al., 2004; Zheng et al., 2005). Furthermore, constitutively spliced exons with weak splice sites or PPTs often require additional cis-acting splicing enhancer elements (described in next section) to promote recognition by the spliceosome (Xiao et al., 2007). Thus, although these core signals play a fundamental role in all splicing events, their sequences can be modulated to varying degrees to influence regulated splicing decisions.

1.4.2 Exonic and intronic splicing enhancers and silencers

The specificity achieved by the spliceosome cannot be accounted solely through its recognition of the core splicing signals. *Bona fide* splice sites must be identified amidst an order of magnitude greater number of cryptic or pseudo-splice sites that match consensus sequences but are never recognized by the spliceosome (Chasin, 2007). Not surprisingly, additional sequence elements that facilitate the discrimination of real exons or suppress the utilization of the vast array of pseudo-splice sites have been discovered. Four classes of cis-elements, known as
exonic and intronic splicing enhancers and silencers (ESEs, ISEs, ESSs, and ISSs), have been identified both computationally and experimentally, and named on the basis of their activity and location in pre-mRNA transcripts (Brudno et al., 2001; Fairbrother et al., 2002; Goren et al., 2006; Liu et al., 1998; Mardon et al., 1987; Nagoshi and Baker, 1990; Reed and Maniatis, 1986; Wang et al., 2004b; Yeo et al., 2005; Yeo et al., 2007; Zhang and Chasin, 2004; Zhang et al., 2003 for examples; Figure 1.3). In general, these elements are thought to recruit additional RNA binding proteins (RBPs) that can either potentiate or prevent spliceosome assembly through either positive or negative protein interactions, or, by sterically hindering access to splice sites (for more detailed discussion of these mechanisms, see below). A recent study has also identified several ESSs and ISSs that negatively influence the ability of U1 snRNP to interact with the 5´ splice site, creating a competitive disadvantage with neighboring splice sites (Yu et al., 2008).

Several themes have emerged from focused and genome-wide studies of these splicing regulatory elements. First, location and context are critical determinants of their activity. For instance, several studies have shown that the same regulatory element can function as both an enhancer or silencer of splicing depending on whether it is located in an exon or its flanking introns (reviewed in Wang and Burge, 2008). Even when present in the same relative location, a regulatory element can function differently depending on the presence of additional cis-elements nearby. Second, most splicing enhancer and silencer elements are short (typically 5-8 nucleotides in length) and degenerate in nature, possibly allowing for flexibility in their recognition by RBPs, which can often bind a wide range of sequences. In light of these properties, current models suggest that numerous sequence elements function combinatorially to create the specificity required for splice site recognition. Ultimately, it is the balance between these enhancer and silencer elements, and the complement of trans-acting factors that associate with them in the nucleus, that will determine whether candidate splice sites are constitutively recognized, utilized only in specific tissues or developmental contexts, or never recognized at all.
Figure 1.3: Cis- and trans-determinants of alternative splicing
Numerous cis-element signals and factors acting in trans contribute to the extent to which an alternative exon (red box) is included in a transcript. In addition to the core splicing signals, several enhancer (turquoise bars) and silencer (yellow bars) elements can be found in exon and intron sequences. Enhancer elements are bound by SR proteins (in green) or other RNA binding proteins expressed generally or in specific tissues and cell types. These interactions can stimulate or strengthen the interactions of several essential steps in spliceosomal assembly, such as U1 snRNP interacting with the 5’ splice site, U2AF binding to the PPT and 3’ splice site, and recognition of the BPS. Additionally, SR family and SR-related proteins are believed to assist in bridging interactions between splicing signals across an exon (a process known as exon definition), or across an intron (known as intron definition). Silencer elements are bound by hnRNPs (purple) and additional tissue-specific or generally expressed RNA binding proteins (light yellow). This latter group of proteins can function to antagonize the stimulatory effects of SR proteins, or directly inhibit the binding of other essential splicing factors. It is the balance and combination of these cis- and trans-determinants that influence splice site utilization. Abbreviations: U1, U1 snRNP; U2, U2 snRNP; U2AF, U2 auxiliary factor subunits; SF1, Splicing factor 1, or Branchpoint binding protein; RBP, RNA binding protein; hnRNP, heterogeneous ribonucleoprotein; SR, SR protein; SR-related, SR-related protein.

1.4.3 Secondary structure

The single-stranded nature of pre-mRNA allows intramolecular base-pairing that can creating secondary structures such as stems and loops. Evolution of these secondary structural elements has contributed to the regulation of several aspects of RNA metabolism, including splicing. Several studies have demonstrated that secondary structure can influence splice site
choice in a number of ways (reviewed in Buratti and Baralle, 2004). First, some splicing regulators, such as the Nova proteins, can in some cases recognize binding sequences as loop elements in a surrounding structured RNA (Buckanovich and Darnell, 1997). Secondary structure can also influence the accessibility of core splicing signals as well as enhancer and silencer elements (Li et al., 2010). For example, it was found that alternatively spliced exons tend to have their splice sites overlapping with regions of secondary structure (Shepard and Hertel, 2008). In another study, it was observed that known splicing regulatory elements have a higher probability of being located in regions lacking secondary structure than randomly selected sequences (Hiller et al., 2007). These studies also indicated that the sequence space surrounding these elements is also under selective pressure, and mechanisms that alter accessible and inaccessible states have the potential to regulate AS.

Longer-range base-pairing interactions can be used to bring splice sites into close proximity while excluding other sites through unfavorable conformations. A classic example of this mode of regulation was identified in the mutually exclusive splicing of a pair of exons in beta-tropomyosin transcripts (Clouet d'Orval et al., 1991). In these transcripts, exon 6B is preferentially included in skeletal muscle, while exon 6A is included in most other tissues. It was found that a portion of the intron upstream of exon 6B base-pairs with a part of the exon and downstream intron in a non-skeletal muscle cell type. Importantly, disruption of this secondary structure allowed inclusion of this exon to occur, and repression could be restored by compensatory mutations that once again allowed base-pairing (Clouet d'Orval et al., 1991). These observations suggest that in skeletal muscle cells the inclusion of this exon occurs by overcoming the repressive effects of the structural element. Additional cases involving this type of regulation have been covered in greater detail elsewhere (Buratti and Baralle, 2004). A complex example of long-range secondary structures influencing AS has been found in the mutually exclusive splicing of the exon 6 cluster in the Dscam gene in Drosophila. Two classes of conserved elements, the docking site downstream of exon 5 and one of 48 selector sequences upstream of each variable exon 6, are capable of forming competing RNA secondary structures that play an important role in juxtaposing only one pair of splice sites (Graveley, 2005). Such a mechanism involving competing RNA structures has recently been validated in other extreme cases of mutually exclusive splicing (Yang et al., 2011), and suggests that it has served as a general evolutionary strategy.
An elaborate example of how RNA secondary structure can regulate AS has been demonstrated in the fungus *Neurospora crassa*, where pre-mRNAs from several thiamine metabolism genes carry riboswitches in their introns (Cheah et al., 2007). These riboswitches undergo conformational changes via their interaction with thiamine pyrophosphate (TPP). In one transcript encoding the NMT1 protein, when TPP levels are low, the riboswitch adopts a conformation that allows selection of an upstream 5’ splice site, leading to the translation of NMT1. When TPP levels are high, the riboswitch adopts a different conformation that leads to splicing at a downstream 5’ splice site, resulting in the inclusion of upstream methionine codons and the translation of an upstream “decoy” open reading frame and not the NMT1 protein product. This mechanism has thus evolved to homeostatically control thiamine levels. It will be interesting to determine if such riboswitch-mediated AS regulation also occurs in animals.

1.4.4 Core spliceosomal proteins

In vitro studies have identified a number of core spliceosomal proteins that inhibit splicing when depleted from nuclear extracts. Many of these factors have been thought to be required for all pre-mRNA splicing events. However, several observations now suggest that altered expression or activity of these factors can lead to specific changes in AS patterns. For example, an RNAi screen testing the effect of depleting RNA binding proteins on AS of several *Drosophila* transcripts revealed that snRNP components, both subunits of U2AF, and several helicases can regulate specific AS events (Park et al., 2004). More recently, it was demonstrated that reduced levels of the survival of motor neurons protein SMN1, which is an important regulator of snRNP biogenesis, can lead to tissue-specific changes in snRNA levels and splicing patterns (Zhang et al., 2008b). Similarly, depletion of the SmB protein, a subunit of the heptameric Sm ring that binds the U1, U2, U4, and U5 snRNAs, was found to reduce the steady-state levels of several snRNAs, leading to increased skipping of specific alternative exons, including several found in RNA processing factor transcripts (Saltzman et al., 2011). Detailed analysis of one of these alternative exons, found in the SmB transcript itself, suggests that the strength of the 5’ splice site flanking an alternative exon may play a key role in determining its sensitivity to SmB levels. Additionally, as mentioned above, several silencer elements have been identified that are believed to influence the ability of U1 snRNP to engage the 5’ splice site, potentially allowing the corresponding exons to be sensitive to fluctuations in snRNP abundance and activity (Yu et al., 2008). Collectively, these results indicate that splicing can be modulated...
by a much larger set of factors than previously appreciated, including components once thought to act in a general manner. A better understanding of how these core splicing regulators can interact with the landscape of other tissue-specific and generally expressed auxiliary factors will be important.

1.4.5 SR and SR-related proteins

Members of the SR family and SR-related splicing factor proteins were first identified through genetic analyses of the sex-determination pathway in Drosophila, and subsequently by their ability to influence splicing activity in vitro in HeLa extracts (Amrein et al., 1988; Ge and Manley, 1990; Goralski et al., 1989; Krainer et al., 1990). Since their initial discovery, many additional members have been identified, and SR and SR-related proteins currently represent a major class of constitutive and alternative splicing regulators (Lin and Fu, 2007; Long and Caceres, 2009; Figure 1.3). More recently SR proteins have also been found to play a role in other aspects of gene regulation, including translation (Sanford et al., 2004). A hallmark of these splicing factors is a protein domain consisting of stretches of alternating arginine-serine dipeptide repeats (known as the RS domain), implicated in facilitating important protein-protein and protein-RNA interactions during spliceosome recruitment and catalysis (Shen et al., 2004; Wu et al., 1993). SR proteins also possess one or two RNA recognition motifs (RRMs), allowing them to interact with pre-mRNAs in a sequence-specific manner. SR-related proteins such as both U2AF subunits, the U1 snRNP component U1-70K, and coactivators SRm160 and SRm300, may or may not have RNA binding activity and are believed to function primarily through protein-protein interactions across splice sites by binding to components of the splicing machinery and with other RS domain containing proteins (Blencowe, 2000; Lin and Fu, 2007; Sharma et al., 2008). Nearly all SR family and SR-related proteins are expressed ubiquitously but at varying levels across tissues, and it has been proposed that the variation in expression of these factors contributes to tissue-specific splicing decisions (Chen and Manley, 2009; Hanamura et al., 1998; Zahler et al., 1992). Recent evidence from my own work indicates that a subset of RS domain proteins with highly restricted expression patterns exist and regulate networks of tissue-specific AS events (Calarco et al., 2009; Chapter 3).

Current evidence suggests that SR proteins generally promote exon inclusion by binding to enhancer elements in exons. This stimulatory effect on splicing is believed to happen through
two non-mutually exclusive mechanisms. First, through its RS domain, an ESE-bound SR protein can recruit or stabilize components bound to the adjacent 5′ and 3′ splice sites, such as U1 snRNP or the U2AF complex, and through direct interactions with the BPS and 5′ splice site (Graveley et al., 2001; Kohtz et al., 1994; Li and Blencowe, 1999; Wang et al., 1995; Zahler and Roth, 1995; Zuo and Maniatis, 1996; Zuo and Manley, 1994). These stimulatory interactions are particularly important in pairing the splice sites across an exon through a process known as “exon definition” (Robberson et al., 1990). Second, an ESE-bound SR protein can inhibit or sterically block inhibitory splicing factors such as hnRNP proteins or other RNA binding proteins (Kan and Green, 1999; Zhu et al., 2001). A recent study has also demonstrated that SR proteins can promote alternative exon skipping by stimulating competing splice sites at neighboring constitutive exons, suggesting that they can act as both positive and negative regulators of splicing (Han et al., 2011).

Given their importance in regulating both constitutive and alternative splicing, RS domain proteins are regulated by post-translational modifications as part of integrated signaling circuits. The best characterized modification is the phosphorylation of the RS domain. In vitro, RS domain phosphorylation is critical for spliceosome assembly and its subsequent dephosphorylation is important for catalysis to occur (Cao et al., 1997; Mermoud et al., 1992; Mermoud et al., 1994; Xiao and Manley, 1998). Several kinases that phosphorylate RS domains have been identified, including members of the SR protein kinase (SRPK) and Clk/Sty families (Colwill et al., 1996; Duncan et al., 1997; Gui et al., 1994; Wang et al., 1998). Manipulations in the levels or activity of these kinases have been shown to influence splicing and the subcellular localization of SR proteins (reviewed in Lin and Fu, 2007). These results suggest that modulation of the signaling pathways that these kinases are involved in may serve as a potential mechanism to regulate AS.

1.4.6 heterogeneous nuclear ribonucleoproteins

Heterogeneous nuclear ribonucleoproteins (hnRNPs) were initially identified on the basis of their co-purification with nuclear pre-mRNA as high molecular weight complexes (Dreyfuss et al., 1993). On account of this rather non-specific defining criterion, members of this group of proteins have somewhat divergent domain architecture and are involved in diverse aspects of mRNA metabolism (Martinez-Contreras et al., 2007). The first indication that hnRNPs could
Regulate splicing when it was demonstrated that hnRNP A1 antagonizes the effects of the SR protein SF2/ASF in splice site selection in vitro (Mayeda and Krainer, 1992). Since then, a number of hnRNPs have been shown to influence AS regulation. Contrary to SR proteins, hnRNPs are thought to generally act as splicing repressors, however several studies have demonstrated that they can also stimulate splicing (Figure 1.3; see below).

Regulation of AS by hnRNPs can occur through several mechanisms. First, the binding of these factors to silencer elements in exons or introns can block the assembly of the spliceosome. This repression can occur through different mechanisms depending on the location of silencer elements. For instance, when silencer elements are located close to the splice sites, steric occlusion preventing the binding of U1 snRNP or the U2AF subunits has been observed in the regulation of several alternative exons (Buratti et al., 2004; Eperon et al., 2000; Jacquenet et al., 2001). Alternatively, if the silencer elements overlap with or are in close proximity to enhancer elements recognized by SR proteins, hnRNPs can compete with these factors for binding or inhibit their stimulatory effects on spliceosome assembly and exon definition (Cartegni and Krainer, 2002; Kashima and Manley, 2003; Zahler et al., 2004a). In addition to local effects, hnRNP-mediated inhibition can occur over distances. Specifically, it has been demonstrated that a single hnRNP A1 bound to an ESS can nucleate the cooperative binding of additional hnRNP A1 molecules, coating the pre-mRNA in a 3’ to 5’ direction until inhibiting the recruitment of an SR protein to an upstream ESE (Zhu et al., 2001). Additionally, hnRNP A1 has been reported to interact with two intronic regulatory elements flanking an alternative exon in its own transcript. Both of the RNA-bound proteins can interact with each other, looping out the alternative exon and bringing the splice sites of the neighboring constitutive exons in close proximity (Blanchette and Chabot, 1999). Other hnRNPs, such as the polypyrimidine tract binding protein (PTB; see below), are also believed to utilize some variation of this “looping out” mechanism to regulate AS events (Oberstrass et al., 2005). Several focused studies of specific pre-mRNAs, including the studies just mentioned, have suggested that hnRNPs generally function to antagonize SR proteins. However, a recent genome-wide analysis of transcripts regulated by hnRNPs and SR proteins in Drosophila S2 cells revealed little overlap between targets of these factors (Blanchette et al., 2005). Further investigation of how these factors regulate AS is needed to resolve some of these discrepancies.
Additional work has suggested that when bound to RNA, the activity of hnRNPs on splicing can differ depending on neighboring splicing regulatory elements. Poly-guanine stretches, or “G-runs”, have been identified as intronic splicing enhancer elements that can associate with hnRNPs F and H. Interestingly, when in the vicinity of strong and weak 5´ splice sites, G-runs do not stimulate splicing as highly as when they flank intermediate strength splice sites (Xiao et al., 2009). These regulatory effects suggest that G-runs can act as buffers to allow mutations that weaken splice sites to accumulate without initially influencing splicing. Eventually however, this buffering effect could contribute to the accelerated evolution of novel AS patterns. In another recent study, it was found that hnRNP L binds an ESS in variable exon 5 of CD45 transcripts, and inhibits its inclusion by repressing the activity of a neighboring enhancer element (Motta-Mena et al., 2010). However, when the enhancer element is absent, hnRNP L bound to the silencer stimulates splicing of variable exon 5. Additional experiments indicated that the activity of hnRNP L depends on the strength of flanking splice sites; strong splice sites are inhibited but weak splice sites are stimulated by this factor. The above studies reveal an emerging theme in splicing regulation: trans-factors and the cis-elements they recognize can lead to distinct regulatory outcomes depending on the molecular and cellular context.

1.4.7 RNA binding proteins and non-coding RNA

In addition to SR proteins and hnRNPs, other RBPs have been discovered that regulate AS. Several of these factors are differentially expressed or post-translationally regulated in specific tissues or at specific developmental stages (Nilsen and Graveley, 2010). These proteins have diverse domain architecture, and can influence splicing through two general mechanisms. Many of these factors regulate splicing through the inhibition or recruitment of spliceosomal components (Chen and Manley, 2009). Alternatively, some RBPs form multi-component complexes with RS domain proteins or hnRNPs to influence AS patterns. The specific actions of many of these factors will not be covered here due to space limitations. In the following sections, I will focus primarily on factors that play a role in AS regulation in the nervous system.

Non-coding RNAs (ncRNAs) have also emerged as regulators of AS. Micro-RNAs (miRNAs) have been found to modulate the expression of several splicing factors to help establish tissue- and developmental-specific splicing regulatory programs (Boutz et al., 2007a;
Kalsotra et al., 2010; Makeyev et al., 2007). Another study identified small nucleolar RNAs (snoRNAs)—a class of ncRNAs normally involved in editing snRNAs, transfer RNAs, and ribosomal RNAs—that can stimulate the inclusion of an alternative exon in transcripts encoding the serotonin receptor 5-HT$_{2C}$R (Kishore and Stamm, 2006). For instance, a particular subset of snoRNAs, called HBII-52 in humans, are thought to interact with the 5-HT$_{2C}$R pre-mRNA through base-pairing and inhibit the activity of a silencer element in the alternative exon. Several additional HBII-52 regulated AS events have now been identified, and further analysis has indicated that these snoRNAs exist in an RNP complex, associating with a number of hnRNPs and other factors with roles in splicing regulation (Kishore et al., 2010). Intriguingly, loss of expression of the cluster of HBII-52 snoRNAs and several other genes has been associated with Prader-Willi Syndrome, a congenital disorder with several phenotypes including mental retardation. In agreement with this observation, several of the targets regulated by HBII-52 snoRNAs also show AS differences between normal individuals and patients with Prader-Willi Syndrome (Kishore et al., 2010).

Another recent study has demonstrated that long ncRNAs can also regulate AS. It was found that the nuclear-retained long ncRNA MALAT1 (metastasis associated lung adenocarcinoma associated transcript 1) localizes to nuclear speckles, foci in the nucleus that are believed to function as “storehouses” for splicing factors (Tripathi et al., 2010). This ncRNA was found to be required for the proper localization of several SR proteins, including SF2/ASF, to nuclear speckles. The absence of MALAT1 also led to the hypophosphorylation of SR proteins, and altered AS of a number of SF2/ASF-regulated alternative exons (Tripathi et al., 2010). While additional work is needed to determine the extent to which ncRNAs play a role in controlling AS, these discoveries reveal a previously under-appreciated layer of gene regulatory complexity.

1.5 AS regulatory networks and the nervous system

As mentioned above, AS events are frequently regulated in a tissue- or cell-type-specific manner. In particular, extensive tissue-specific AS has been observed in the mammalian nervous system, and is thought to contribute to both its molecular and cellular complexity (Lipscombe, 2005). It is thus not surprising that many studies of tissue-specific AS regulation have been directed at understanding splicing in the nervous system (Li et al., 2007). In conjunction with biochemical and molecular genetic methodologies, several genome-wide approaches have
proven valuable in our understanding of how neural-specific splicing factors regulate AS in the nervous system (Blencowe, 2006; Licatalosi and Darnell, 2010; Ule and Darnell, 2007). These approaches have also helped define some of the first AS regulatory networks, providing insights into their organizational principles and new insights regarding how specific groups of gene products function together in the developing and mature nervous system. The following sections focus on the progress in our understanding of neural splicing factors and their regulatory networks, and insights gained from these studies.

1.5.1 Nova-1 and Nova-2

The first neuro-oncological ventral antigen (Nova-1) protein was originally identified as a factor recognized by antibodies produced in patients with paraneoplastic opsoclonus myoclonus ataxia (POMA), a motor control disorder resulting from an autoimmune response (Buckanovich et al., 1993). The Nova-1 antigen displays restricted expression in the nervous system, specifically in the hypothalamus, ventral midbrain, hindbrain, and spinal cord. Nova-2, a second antigen highly related to Nova-1, was subsequently identified and found to be expressed specifically in brain sub-regions, but with a reciprocal pattern of expression to Nova-1 (Yang et al., 1998). Both Nova proteins share sequence homology with the RNA binding protein hnRNP K (Buckanovich et al., 1993), alluding to their role in RNA processing.

Initial biochemical, crystallographic, and molecular genetic-based approaches indicated that the Nova proteins bind YCAY repeat sequences, and a small number of target transcripts were identified based on the presence of these elements (Buckanovich and Darnell, 1997; Dredge and Darnell, 2003; Lewis et al., 2000). A significant breakthrough was brought forth by the development of a Nova-1 knockout mouse. These knockout mice die shortly after birth from motor deficiencies caused by apoptosis of brainstem and spinal cord neurons, and display a number of phenotypes similar to those found in POMA (Jensen et al., 2000). Intriguingly, the loss of Nova-1 altered the neuron-specific alternative splicing patterns of target transcripts containing YCAY repeat elements. These results established the Nova proteins as the first tissue-specific splicing regulators contributing to developmental programs, and implicated these factors as causative agents in neurological disorders.

The functional analyses of the Nova proteins were greatly facilitated by the development of CLIP (Ule et al., 2003). The initial application of CLIP identified thirty-four target mRNAs
associated with the Nova proteins. These target transcripts were enriched in genes with known roles at the synapse, in particular inhibitory synaptic transmission. Ule and colleagues further analyzed mRNA samples from brain and immune system tissues of wild type and Nova-2 knockout mice using AS microarrays (Ule et al., 2005b). More than 50 AS events from the neocortex were identified that required Nova-2 for proper regulation, corresponding to ~7% of all brain-specific AS events detected in their profiling experiments, suggesting that Nova-2 plays a significant role in regulating brain-specific AS. In agreement with the CLIP analysis, a significant proportion of genes with Nova-2-regulated AS events were associated with functions at the synapse, in particular synaptic development, synaptic transmission, cell-cell signaling, and cortical actin organization. Intriguingly, many of the regulated target genes with AS events interact with each other in protein complexes, suggesting that Nova-2 regulates a network of synaptic proteins and that it plays an important role in affecting physical interactions between these factors.

Using this set of AS events, a bioinformatics approach was used to examine the effects on exon inclusion resulting from the positional binding of the Nova proteins to YCAY repeats in regions surrounding these alternative exons (Ule et al., 2006). The resulting “RNA map” predicted that Nova could act as either an enhancer or repressor of exon inclusion depending on the locations of the YCAY elements in pre-mRNA transcripts. These predictions were subsequently validated experimentally, and led to testable hypotheses regarding the mechanism by which Nova regulates AS. Specifically, when bound to exonic YCAY elements, Nova blocks U1 snRNP association with the 5′ splice site. However, when bound in the intronic region downstream of an alternative exon, Nova stimulates splicing complex formation. These results indicated how genome-wide approaches can provide key regulatory insights into splicing mechanisms. As new technologies become available, our ability to uncover novel insights continue. For example, Licatalosi and colleagues subsequently coupled the CLIP protocol with high-throughput sequencing (HITS-CLIP; also referred to as CLIP-Seq), and also used a more comprehensive AS microarray platform to identify a more extensive Nova regulatory network (Licatalosi et al., 2008). In addition to confirming and further refining the bioinformatically predicted RNA map, the authors identified a new role for Nova in regulating alternative polyadenylation. This new regulatory role for Nova appears distinct from its role in AS, since
transcripts regulated at the level of poly A site selection generally do not have Nova regulated alternative exons.

1.5.2 PTB and PTBP2

The presence of a tissue-restricted counterpart of the ubiquitously expressed polypyrimidine tract binding protein (PTB, also known as PTBP1 or hnRNP I) was initially detected in rat brain extracts based on its different electrophoretic mobility but similar RNA binding activity to PTB (Ashiya and Grabowski, 1997). Two independent studies subsequently identified and cloned PTBP2 (also known as nPTB and brPTB), confirming that it is a tissue-restricted paralog of PTB (Markovtsov et al., 2000; Polydorides et al., 2000). Markovtsov and colleagues identified PTBP2 as a key factor playing a role in the neuronal specific splicing of the N1 alternative exon from the c-src transcript. Polydorides and colleagues identified PTBP2 as a protein that interacts with Nova-1/2, and further demonstrated its ability to antagonize the stimulatory effect of Nova proteins in the splicing of exon E3A in GlyRα2 transcripts. Both studies revealed an enriched expression pattern of PTBP2 in brain and testis, suggesting that this factor might act as a key modulator of alternative splicing patterns in these tissues.

Although PTBP2 has high sequence similarity to PTB, by virtue of its restricted expression pattern and other properties, it has the ability to establish tissue-specific splicing programs. Data from in vitro splicing assays indicates that PTBP2 can act either neutrally or as a weaker repressor than PTB in preventing c-src N1 exon inclusion (Markovtsov et al., 2000). Additionally, PTBP2 binds to an intronic region of the c-src pre-mRNA critical for tissue-specific splicing of exon N1, known as the distal control sequence, with higher affinity than PTB, and this binding is stimulated by generally expressed co-factors such as hnRNP H (Markovtsov et al., 2000). These results and others have suggested that a combination of differential intrinsic RNA binding affinity and interactions with co-regulators likely play a role in the distinct properties of PTB and PTBP2 (Charlet et al., 2002; Gromak et al., 2003). Further mechanistic studies on the repressive role of PTB in a non-neuronal context have revealed that intron-bound PTB can block prespliceosomal E complex formation by preventing 5′ splice site dependent recruitment of U2AF, and also blocking the transition from exon definition to intron definition (Sharma et al., 2005; Sharma et al., 2008). In light of this evidence, it is possible that
PTBP2 somehow acts distinctly or in a less repressive manner at these steps in spliceosome assembly and commitment.

Recent focused and genome-wide analyses of PTB and PTBP2 have also shed light on the mechanism by which these factors regulate AS. Both proteins display mutually exclusive expression patterns in the brain, with PTB being expressed in glial and non-neuronal cells, and PTBP2 in neurons (Boutz et al., 2007b). When PTB is present at sufficient levels, it represses splicing of exon 10 in PTBP2 transcripts, producing an isoform containing a premature termination codon (PTC) that is degraded by the nonsense mediated mRNA decay pathway (Boutz et al., 2007b; Spellman et al., 2007). PTB also appears to further inhibit PTBP2 expression or turnover at the protein level independent of its effect on splicing of exon 10. Thus, downregulation of PTB should lead to the expression of functional PTBP2. In neurons, the silencing of PTB expression is achieved by the neuron-specific micro-RNA miR-124, which effectively de-represses PTBP2 exon 10 splicing, although efficient inclusion of this exon in neurons also requires positive acting splicing regulators (Calarco et al., 2009; Makeyev et al., 2007; see below). This mutually exclusive expression pattern appears to be established after cells have committed to a neuronal fate, because neural progenitor cells appear to express both PTB and PTBP2 (Boutz et al., 2007b).

In order to assess functional consequences resulting from modulating PTB and PTBP2 levels in a neuronal context, Boutz and colleagues performed AS microarray profiling experiments comparing the transcriptomes of neuroblastoma cells depleted of each splicing factor alone or in combination (Boutz et al., 2007b). These experiments revealed AS events with altered splicing patterns that depended solely on PTB or PTBP2 expression, as well as events that required both factors for regulation. Notably, events showing altered AS patterns undergo both increased inclusion or increased skipping, suggesting that these factors can act both as repressors or activators of splicing. Additionally, a subset of the genes encoding transcripts with AS events regulated by PTB and PTBP2 have previously known roles in neuronal differentiation and function, providing a first glimpse into the splicing network regulated by these factors.

Two recent genome-wide analyses have provided further insights into the repertoire of transcripts controlled by PTB and PTBP2 and the mechanisms by which they operate. Xue and colleagues performed CLIP-Seq in HeLa cells to identify sites in the transcriptome where PTB
binds *in vivo* (Xue et al., 2009). These binding sites were often found in introns, and enriched in pyrimidine-rich sequence elements. Intriguingly, an examination of the effect of shRNA-mediated knockdown of PTB revealed that PTB can also act as an enhancer of exon inclusion when its intronic binding sites closely flank neighboring constitutive exons. In another study, Llorian and colleagues knocked down PTB and PTBP2 expression in HeLa cells, and used Affymetrix splicing microarrays to identify hundreds of AS events regulated by these factors (Llorian et al., 2010). Similar to the studies mentioned above, the authors find that PTB proteins can act as context-specific activators or repressors of AS. Analysis of these events using an AS microarray dataset profiling diverse human tissues (Castle et al., 2008) revealed that PTB/PTBP2 dependent AS events were frequently differentially regulated in nervous system and muscle tissues, consistent with observations that pyrimidine-rich motifs are often associated with brain-specific AS events in mouse (Fagnani et al., 2007a). Notably, these differentially regulated AS events are associated with genes enriched in G-protein signaling and regulation of the cytoskeleton and membrane traffic. An investigation of cis elements enriched in PTB-regulated transcripts identified pyrimidine rich motifs that exhibited positional bias depending on the regulatory effect. For instance, PTB was found to act as a repressor when bound in an intron upstream of an alternative exon or within the alternative exon itself, but can function as an activator when bound to an intronic element downstream of the alternative exon. Finally, additional motifs belonging to the Fox and MBNL family of splicing regulators were also significantly associated with PTB-regulated AS events, suggesting that these regulators more generally act in combination with PTB and PTBP2, in agreement with previous studies on model pre-mRNA transcripts (Jin et al., 2003; Ladd et al., 2005; Underwood et al., 2005). The proposed RNA maps from Xue et al. and Llorian et al. do not perfectly overlap, likely due to the different methods from which they were derived. However, they have each provided novel mechanistic insights into PTB/PTBP2 AS regulation that warrant further investigation. Collectively, these studies have also elaborated the repertoire of tissue-specific AS events controlled by these factors.

1.5.3 Fox splicing regulators

Unlike most cis-elements recognized by RNA binding proteins which are typically degenerate, several studies identified the conserved hexanucleotide sequence (U)GCAUG enriched in introns downstream of brain-specific alternative exons (Brudno et al., 2001;
It was subsequently discovered that a family of tissue-specific RNA binding proteins, homologous with the *C. elegans* feminizing on X (FOX-1) protein involved in dosage compensation, regulated tissue-specific AS events through interaction with this cis-element (Jin et al., 2003; Nakahata and Kawamoto, 2005; Ponthier et al., 2006; Underwood et al., 2005). Both mammalian Fox-1 (also known as A2BP1) and Fox-2 (also known as Fxh or RBM9) show enriched or highly restricted expression in brain, heart and skeletal muscle tissues. A third member of the family with similar RNA binding specificity, Fox-3, has recently been identified as the antigen recognized by the anti-NeuN (neuronal nuclei) antibody, which for nearly two decades has served as a specific marker of most classes of post-mitotic neurons (Kim et al., 2009). A number of molecular genetic and biochemical studies on a small number of premRNA substrates have indicated that the Fox family of splicing regulators can act as position-dependent splicing activator or repressor proteins (Jin et al., 2003; Nakahata and Kawamoto, 2005; Ponthier et al., 2006; Underwood et al., 2005). However, it remained unclear whether these observations could generally be applied to Fox-dependent splicing regulation. An understanding of the network of transcripts regulated by these splicing regulators would yield novel insights into mechanisms that govern the development of neurons and muscle cells. Recent genome-wide analyses have begun to shed light on some of these questions.

Zhang and colleagues mined 28 sequenced vertebrate genomes for Fox binding sites in exons and 200 nucleotides of flanking intron sequence (Zhang et al., 2008a). Thousands of putative targets were identified, hundreds of which are found in the vicinity of known alternative exons. AS patterns of candidate Fox-associated transcripts across tissues were further assessed by microarray profiling experiments, associating the location of Fox binding sites with potential tissue-specific regulation in brain and muscle tissues. The results suggest that Fox binding sites are more often enriched in intronic regions downstream of alternative exons, and when present allow Fox proteins to activate inclusion of alternative exons. However, when binding sites are present in upstream intronic regions, Fox generally acts as a splicing repressor. Intriguingly, complex regulatory patterns were also observed where candidate Fox-dependent AS events displayed increased inclusion or skipping in either brain or muscle tissues, but not in both, indicating that combinatorial interactions with other splicing regulators play a role in regulating these events. GO enrichment analysis of the Fox splicing network revealed that genes with Fox-regulated AS events have a role in neuromuscular function, consistent with the tissues where
these factors are enriched. Additionally, genes with Fox-dependent AS events were frequently associated with diseases, including several neurological, heart and muscular disorders. Fox-1 has been associated with several neurological disorders and heart disease (Bhalla et al., 2004; Kaynak et al., 2003; Szatmari et al., 2007), thus making the identified splicing regulatory network an excellent dataset to identify potentially interesting relationships between RNA targets and the etiology of some of these diseases.

In another study, Yeo and colleagues observed that Fox-2 is robustly expressed in human embryonic stem cells (hESCs) (Yeo et al., 2009). The authors performed CLIP-Seq experiments to identify Fox-2 binding sites in vivo. Fox-2 binding clusters were identified in nearly 2000 protein coding genes, suggesting a large network of targets potentially regulated by this factor in hESCs. Interestingly, further analysis of the CLIP tag clusters found enrichment of not only the UGCAUG consensus Fox-binding motif, but also other consensus elements, suggesting potential alternative modes of RNA recognition through their interactions with additional factors. Investigating the effects on splicing patterns of interacting transcripts, the authors found that downstream intronic binding sites lead to exon inclusion, while upstream intronic sites lead to repression by Fox-2, consistent with the results of Zhang and colleagues. Genes encoding transcripts bound by Fox-2 were enriched in splicing factors and kinases, suggesting that this protein may act to coordinate a network of splicing events controlled by multiple factors, and also signaling cascades important for maintenance of the ESC state. Supporting this notion, the authors observed that knockdown of Fox-2 leads to a specific and dose-dependent increase in cell death of hESCs. Thus, the network of targets regulated by Fox-2 will likely provide insights into stem cell biology.

Two recent studies profiling AS in breast and ovarian cancers have also found a connection between the Fox family of splicing regulators and cancer (Lapuk et al., 2010; Venables et al., 2009). These studies implicate the altered regulation of Fox-dependent splicing networks in various cancers, and future analyses of these networks should uncover how these isoforms contribute to the development and progression of the disease.

1.5.4 Sam68

Sam68 (Src-associated in mitosis, 68 kDa) was originally identified as a phosphorylation substrate of the tyrosine kinase Src during mitosis, and subsequently found to be related to a
family of RNA binding proteins now known as STAR (Signal Transduction and Activation of RNA) or GSG (GRP33, Sam68, GLD-1) domain proteins (Fumagalli et al., 1994; Taylor and Shalloway, 1994; Vernet and Artzt, 1997). Sam68 is proposed to play roles in numerous aspects of mRNA metabolism, including AS, mRNA export and localization, and translation regulation (Lukong and Richard, 2003). Several studies have also suggested links between this factor and various cancers (Rajan et al., 2008). Although expressed in multiple tissues, Sam68 expression is induced at sites of neurogenesis in mice (Lim et al., 2006). Mice lacking Sam68 display motor coordination deficits and male sterility, suggesting that this factor is involved in aspects of neuronal function and fertility (Lukong and Richard, 2008; Paronetto et al., 2009). Given these intriguing phenotypes, it has been an important goal to identify the network of transcripts regulated by Sam68. Several studies have identified mRNAs bound to Sam68 both in neurons and in spermatocytes (Grange et al., 2009; Paronetto et al., 2009), and a small number of Sam68-dependent AS events (Matter et al., 2002; Paronetto et al., 2007; Pedrotti et al., 2010).

To better define the network of AS events regulated by Sam68, a recent study performed AS microarray profiling in mouse neuroblastoma cells treated with shRNAs targeting Sam68 expression (Chawla et al., 2009). From over 1,000 profiled AS events, approximately 30 Sam68-dependent AS events were identified. Consistent with other studies of splicing regulators, Sam68 appears to function both as an enhancer or repressor of exon inclusion. A search for previously identified AU-rich motifs known to bind Sam68 revealed an enrichment of these cis-elements in intronic regions flanking Sam68-regulated alternative exons. Mutagenesis experiments studying a Sam68-regulated exon in epsilon sarcoglycan (Sgce) transcripts demonstrated that multiple Sam68 binding sites are important in order to achieve full regulation by the RNA binding protein.

Given that Sam68 is activated during neurogenesis, Chawla and colleagues used P19 cells, which can be differentiated into neuronal-like cells, to assess the role of Sam68 in neuronal differentiation (Chawla et al., 2009). Sam68 expression was also found to be induced during P19 cell differentiation, and AS events from the identified Sam68 regulatory network were found to undergo splicing changes during this process. Intriguingly, depletion of Sam68 inhibits P19 differentiation and alters the regulation of these differentiation-dependent AS events. Thus, this work has uncovered an important network of isoforms that contributes to aspects of neuronal differentiation. Future work will be required to identify which isoforms are specifically required
for this process, and whether any of these targets are causal in the motor coordination deficits in Sam68 null mice.

1.5.5 Additional neural-specific splicing regulators

Several additional neural-restricted RNA binding proteins have been found to regulate AS, but to date no genome-wide analyses of these factors have been conducted in tissues of the nervous system. Mammalian Hu proteins, homologs of the Drosophila splicing regulator ELAV, are represented by four family members, three of which (HuB, HuC, and HuD) that are expressed highly in nervous system tissues (Hinman and Lou, 2008). Similar to the Nova proteins, the Hu proteins were also identified as neuronal antigens recognized by antibodies produced in patients with paraneoplastic disorders (Szabo et al., 1991). Furthermore, HuD knockout mice display a number of phenotypes both during development and in adulthood, including impaired motor and reflex skills, and control of neural progenitor cell proliferation and differentiation (Akamatsu et al., 2005). Work on these factors has indicated that they recognize AU-rich sequence elements and are involved in many aspects of mRNA metabolism, including AS (Hinman and Lou, 2008).

Two additional classes of RNA binding proteins with family members expressed specifically in nervous system and/or muscle tissues are the muscleblind-like (MBNL), and CUGBP/ETR-like (CELF; also known as Bruno-like) factors (Barreau et al., 2006; Pascual et al., 2006). Both of these factors have been implicated in neurological disorders, and in CUG tri-nucleotide expansion diseases such as myotonic dystrophy (Cooper et al., 2009; Gallo and Spickett, 2010). These factors have also been implicated in a number of aspects of mRNA metabolism, including AS (Ho et al., 2004; Ladd et al., 2001). Several studies using model pre-mRNA substrates have demonstrated that MBNL and CELF bind CUG-repeat and UG-rich elements, respectively, and typically antagonize each other and other factors such as PTB in AS regulation (Charlet et al., 2002; Ho et al., 2004). Recent genome-wide analyses of regulated AS during heart development and myogenic differentiation have revealed that MBNL and CELF proteins can be dynamically and reciprocally regulated (Bland et al., 2010; Kalsotra et al., 2008). The altered expression of these factors helps coordinate a network of splicing transitions that take place between embryonic and adult heart tissue, or between undifferentiated and differentiated myoblasts. Work from Dembowski and Grabowski has further demonstrated that
CUGBP2 (also known as NAPOR, ETR-3, CELF2, and BRUNOL3) can regulate neural-specific skipping of the NI exon in NMDA R1 receptor transcripts by binding to UG-rich cis-elements flanking the branchpoint sequence upstream of the alternative exon (Dembowski and Grabowski, 2009). Subsequent bioinformatic analyses identified additional exons regulated according to this branchpoint “perimeter” binding model, suggesting that this mode of regulation by CELF factors may be a general mechanism by which they regulate exon skipping in the brain.

More than a decade ago, the *Quaking (qk)* gene was identified as the locus mutated in *Quaking* mice, which exhibit rapid tremors due to a deficiency in myelination of the nervous system (Ebersole et al., 1996). The *qk* locus encodes several alternative isoforms of QKI, which are related to the STAR family of RNA binding proteins. Several lines of evidence have connected QKI proteins with regulating AS of relevant transcripts associated with myelination. For instance, the nuclear isoform QKI-5 represses the inclusion of exon 12 in myelin associated glycoprotein (MAG) transcripts, which encode a transmembrane protein known to play a role in the initiation and maintenance of the myelin sheath (Wu et al., 2002). Additionally, QKI proteins can indirectly regulate AS through the modulation of other splicing regulators. Specifically, QKI-6, a cytoplasmic variant encoded by the *qk* locus, represses the translation of hnRNPA1, which in turn regulates MAG exon 12 AS (Zhao et al., 2010). Several other transcripts encoded by genes relevant to myelin formation and maintenance display altered splicing patterns in *Quaking* mice (Wu et al., 2002). It will be interesting to determine whether similar regulatory mechanisms also apply to other QKI targets. As aberrant regulation of QKI in humans has been linked to schizophrenia (Aberg et al., 2006a; Aberg et al., 2006b), an in depth analysis of the repertoire of transcripts regulated by these factors may provide helpful insights into the disease.

Taken together, an understanding of the network of splicing events regulated by these proteins, and a better characterization of the mechanisms by which they regulate AS will help elucidate the physiological roles of these RNA binding proteins in nervous system development and function.

### 1.6 Coordination of gene regulatory machineries

Extensive connections between AS and other gene regulatory processes have been identified. For instance, splicing and other RNA processing events occur while RNA polymerase II (pol II) actively transcribes pre-mRNA (Maniatis and Reed, 2002). The emergence of a co-
transcriptional splicing mechanism has led to the evolution of numerous interactions between the splicing and transcriptional machineries, mediated largely through the carboxy-terminal domain (CTD) of the largest subunit of pol II (de Almeida and Carmo-Fonseca, 2008; Pandit et al., 2008). The CTD of pol II consists of repeats of the heptapeptide sequence YSPTSPS, which can be post-translationally modified in a dynamic manner during transcription (Egloff and Murphy, 2008). Recent genome-wide analyses in yeast have demonstrated that certain CTD modifications can be regulated in a gene specific manner during transcription (Kim et al., 2010; Tietjen et al., 2010). It will be interesting to determine if similar principles extend to tissue-specific modifications of the CTD during transcription in metazoans, and whether such modifications could lead to differential recruitment of AS regulators at specific splice sites.

The CTD and other factors can also influence the elongation rate of pol II, which in turn can impact the time nascent splice sites are exposed to the splicing machinery, with the potential to influence AS (Munoz et al., 2010). An altered rate of pol II elongation could also affect the association of splicing regulators with pol II during co-transcriptional splicing (Munoz et al., 2010). Recent global analyses have demonstrated that changing the elongation rate of pol II can influence AS of transcripts that encode proteins involved in apoptosis and RNA metabolism (Ip et al., 2011; Munoz et al., 2009). However, it is currently not known whether pol II elongation rates can be influenced differentially across tissues, and whether any such differences can lead to tissue-specific AS regulation. Future studies will hopefully shed light on these interesting possibilities.

Numerous links between chromatin regulation and splicing have also emerged over recent years. For instance, a number of chromatin remodeling factors have been found to interact with components of the spliceosome and other regulators of splicing (Allemand et al., 2008). In one interesting study, it was demonstrated that the SWI/SNF catalytic subunit Brm facilitates the accumulation of pol II with modified CTD phosphorylation patterns over a regulated alternative exon in CD44 transcripts (Batsche et al., 2006). Brm also associates with components of the spliceosome and Sam68, supporting its role in integrating signals from chromatin, the transcriptional machinery, and the spliceosome. A series of recent studies have demonstrated that specific histone modifications in nucleosomes are preferentially enriched or depleted in exon and intron sequences (Huff et al., 2010; Kolasinska-Zwierz et al., 2009; Schwartz et al., 2009b; Spies et al., 2009; Tilgner et al., 2009). These results suggest that particular histone marks could be
instructive or permissive in directing splicing decisions, and raise interesting mechanistic questions. For instance, can different histone modifications in the vicinity of exons recruit splicing factors, or are they generated as a result of splicing? Are histone marks associated with cell- and tissue-specific alternative exons different between tissues? A recent paper has suggested that, for at least some regulated AS events, histone modifications can display biased patterns of enrichment over exons in different cell types, and factors interacting with these marks can potentially recruit splicing regulators (Luco et al., 2010). A better understanding of how these signals are interpreted by chromatin regulators and the spliceosome in order to generate tissue-specific AS patterns will be an important goal.

In addition to the direct links between AS and other gene regulatory processes described above, there is also significant indirect cross-regulation. The coupling of alternative splicing with the NMD pathway serves both as a mechanism for quality control and in the regulation of transcript levels (Lareau et al., 2007a). Many transcription factors are alternatively spliced (Taneri et al., 2004), and knockdown of splicing regulators leads to altered AS of transcription factor isoforms (Boutz et al., 2007b; Calarco et al., 2009). Most studies of gene regulatory networks have thus far focused only on individual processes. The extent of cross-talk between trans-acting proteins at various stages of gene regulation suggests that analyses of the effects of splicing regulators should be extended to other modes of regulation. Furthermore, a detailed understanding of the interplay between chromatin factors, transcription factors, splicing factors, non-coding RNAs, and other pleiotropic regulators in a given developmental paradigm will help identify functional relationships among groups of genes targeted by these factors.

1.7 Evolutionary features of alternative splicing

It is currently unclear as to when AS first evolved. A comparison of several eukaryotic genomes revealed that genes containing multiple introns could be found in species in the plant, animal, and fungal kingdoms (Irimia et al., 2007). Moreover, ‘ancient’ genes (conserved in all three kingdoms) were found to be equally likely to encode AS isoforms as ‘newer’ genes (found only in one or two kingdoms) (Irimia et al., 2007). In a separate study, a computational analysis of the evolution of the 5′ splice site, BPS, PPT, and 3′ splice site across 22 eukaryotes revealed a surprising degeneracy in these core splicing signals for many species (Schwartz et al., 2008). As mentioned above, the degeneracy of these signals is generally considered an important feature
that facilitates AS. Collectively, these results suggest the intriguing possibility that a rudimentary form of AS may have already evolved in the common ancestor to all eukaryotes. However, at present it cannot be ruled out that AS may have evolved independently in the diverse eukaryotic kingdoms through convergent evolution.

Several mechanisms have been proposed to describe how novel AS events can evolve. First, existing exons (and also their surrounding splice sites and regulatory elements) can be duplicated during DNA replication (Kondrashov and Koonin, 2001, 2003). This process is known as exon shuffling. These duplicated exons, depending on the extent of regulatory sequences incorporated during the duplication event, can either immediately compete for spliceosomal recognition or slowly accumulate additional mutations that will ultimately lead to their selective inclusion in transcripts. Second, sequences found in introns can slowly accumulate mutations that eliminate silencer elements, create enhancer elements, and strengthen core splicing signals, leading to their recognition by the spliceosome. This process is known as exonization. Several reports have indicated that transposable elements can accelerate exonization (Nekrutienko and Li, 2001; Sorek et al., 2002). For example, the primate-specific Alu transposable element contains sequences that resemble consensus splice sites and is enriched in exonic splicing enhancers (Schwartz et al., 2009a). It has been experimentally demonstrated that through only a few mutations, an intronic Alu element can become exonized (Lev-Maor et al., 2003; Sorek et al., 2004a). Given these observations and the fact that ~10% of the human genome is composed of these sequences, it was recently estimated that many Alu elements are close to becoming exonized. The third mechanism by which AS events can evolve is through a process known as transition, where ancestrally constitutive exons accumulate mutations that reduce their inclusion efficiency in certain species lineages (Koren et al., 2007; Lev-Maor et al., 2007). The most common mutations leading to transitions involve weakening the efficiency of U1 snRNP in recognizing the 5’ splice site, coupled with the fixation of additional mutations in regulatory elements that modulate the recognition of splice sites.

Alternative exons and splicing events are now known to possess distinct features depending on their evolutionary history. Expressed sequence tag (EST), microarray, and RNA-Seq profiling experiments have been instrumental in directing these observations, providing both a genome-wide and quantitative readout of the frequency of alternative exon and splice site usage. Initial microarray and EST studies profiling AS events in mouse and human tissues
revealed that alternative exons conserved between human and mouse are on average more highly included than alternative exons unique to either species (Modrek and Lee, 2003; Pan et al., 2004). Exons that are highly included in transcripts also have a higher tendency of disrupting or altering protein domains (Pan et al., 2005; Pan et al., 2004). Several studies have analyzed alternative exons conserved in mammals and their flanking intronic sequences, and found that in general they have higher sequence identity than analogous regions spanning constitutive exons (Sorek and Ast, 2003; Yeo et al., 2005). The degree of conservation is even higher for tissue-specific AS events (Fagnani et al., 2007b; Sugnet et al., 2006), indicating that alternative exons generally require more complex regulatory mechanisms than constitutive exons for proper splicing control.

Comparative genomic analyses have also enabled the investigation of selective pressures acting on different classes of exons. Standard metrics used to calculate protein sequence evolution involve comparing the rates of non-synonomous amino acid substitutions ($k_a$) to the rate of synonomous substitutions ($k_s$). The $k_s$ value normally provides the rate of background nucleotide substitutions, because it is assumed that synonomous substitutions have no effect on protein-coding sequence (unless codon bias is important). The $k_a/k_s$ ratio can thus provide regional measurements of selection pressures acting on amino acid sequences (Yang and Nielsen, 2000). However, synonomous mutations can influence splicing regulatory elements and other sequences acting at the level of RNA regulation. Thus, the absolute value of $k_s$ can also serve as a measure of RNA selection pressure (Xing and Lee, 2005). When looking at the values of $k_a$ and $k_s$ spanning different classes of exons and surrounding intron sequences, again the evolutionary history and inclusion level of the exon seems to correlate with selection pressure. Specifically, it was found that alternative exons conserved between human and mouse that are weakly included (“minor” class exons) have significantly reduced $k_s$ values relative to constitutive exons or alternative exons that are frequently included (“major” class exons) (Xing and Lee, 2005). Additionally, it was found that recently evolved minor class exons had much higher relative $k_s$ values when compared with more anciently conserved minor class exons. These recently evolved minor class exons also had significantly higher relative $k_a$ values compared to minor class exons conserved over larger periods of time (Xing and Lee, 2005).

These observations provide several insights into the evolution of functionality in alternative exons. As new exons evolve, they are likely included into transcripts relatively
infrequently, present at low levels, or degraded by NMD in order to avoid potential negative effects on the proteome. Due to their low abundance and neutral or limited influence on overall protein function, little purifying selection pressure would act on these sequences, allowing them to accumulate both non-synonomous and synonomous nucleotide substitutions at an accelerated rate compared to neighbouring constitutive exons. Over time, these rapidly evolving exons could confer some type of selective advantage to the organism, and acquire two fates. First, they could evolve to be included at a higher frequency in transcripts, at which point they will become evolutionarily constrained in a similar manner as major class exons or constitutive exons. Alternatively, they could remain as minor class exons or even acquire tissue-specific functions and regulatory patterns. These latter scenarios would require more complex cis-element based regulation that would now be under purifying selection in addition to selection occurring at the level of the amino acid sequence. In agreement with the argument that function seems to evolve and become constrained over evolutionary time, more distantly conserved alternative exons have a higher propensity to preserve reading frame than exons conserved only over shorter evolutionary distances (Xing and Lee, 2005).

These observations indicate that AS can facilitate the accelerated evolution of functional diversity in proteins. It is therefore not surprising that AS events are frequently regulated in a species-specific manner in mammals. In fact, it is estimated that only ~20% of AS events in human are also alternatively spliced in mouse (Pan et al., 2005; Yeo et al., 2005). Although these conserved events are likely functionally important and have been selected for over time, the remaining more recently evolved AS events may also have critical functions and contribute to important species-specific characteristics. Alternatively, these latter splice variants may not have a function at present but could represent the raw material for continued evolution of novel protein isoforms. An important goal is therefore to determine the proportion of AS events that are functional, and their cellular roles. It is also of great importance to identify recently evolved or evolving AS events that may have contributed to the unique aspects of human biology, ultimately distinguishing us from our closest primate relatives. This latter subject will be discussed in greater detail in Chapter 2.
1.8 Alternative splicing in *C. elegans*

Many studies assaying tissue-specific splicing regulation *in vivo* have utilized mammalian cell based systems. Several splicing regulators have been studied genetically in knockout mice with great success. However, due to the extensive amount of work required to generate these transgenic animals, some investigators have chosen invertebrate species as tractable model organisms to study AS regulation *in vivo* (Blanchette et al., 2005; Graveley, 2005; Kabat et al., 2006). The nematode species *Caenorhabditis elegans*, with its defined cell lineage, and multiple cell and tissue types (including a simple but extensively differentiated nervous system), provides an attractive yet currently under-explored model organism to study the regulation and *in vivo* functions of AS. Most of the key factors involved in regulating AS, including RS domain proteins and members of the heterogeneous ribonucleoprotein (hnRNP) family, are conserved in *C. elegans* (Barbosa-Morais et al., 2006). Moreover, orthologs of the mammalian tissue-specific splicing regulators such as the Fox and CELF family of proteins are also present in *C. elegans*, suggesting considerable evolutionary conservation of the mechanisms underlying cell type-specific AS.

The ease of forward and reverse genetics in *C. elegans* has led to the development of several screens to study constitutive and alternative splicing *in vivo*. A number of splicing regulators have been identified (both directly or inadvertently) through their ability to suppress or enhance the phenotypes of mutants harboring loss of function alleles resulting from mutations in splicing regulatory elements. For example, the *e936* allele of the gene *unc-73* (uncoordinated) contains a mutation in the 5´ splice site flanking exon 15, leading to the use of cryptic splice sites in the pre-mRNA and a loss of proper protein translation (Roller et al., 2000). Screens looking for suppression of the *unc* phenotype identified several alleles, including a mutated U1 snRNA that base-paired more efficiently with the mutant 5´ splice site, and several spliceosome-associated factors that are thought to act after U1 snRNP has assembled on the pre-mRNA (Dassah et al., 2009; Zahler et al., 2004b). Mutation of the RNA binding protein gene *mec-8*, which is required for proper function of mechanosensory and chemosensory neurons, also display a synthetic lethal phenotype with specific alleles of *unc-52* (Lundquist and Herman, 1994). This genetic interaction is caused by the altered splicing of *unc-52* transcripts in the absence of *mec-8* activity (Lundquist et al., 1996). Recently, mutations in the gene *uaf-1*, the *C. elegans* ortholog of U2AF35, were found to suppress an aberrant muscle contraction phenotype...
in gain-of-function mutants of the gene *unc-93* through a mechanism involving altered 3′ splice site choice (Ma and Horvitz, 2009).

In addition to these suppressor and enhancer screens, more direct screens to identify AS regulators have been made possible through the development of two color or bichromatic fluorescent reporters. These reporters utilize an AS event to switch between the expression of one of two fluorescent proteins (Kuroyanagi et al., 2006; Kuroyanagi et al., 2010). The relative intensity of each fluorescent protein is then analyzed by microscopy in live animals to provide a visual indicator of isoform usage. Mutant animals can then be screened for alterations in the ratio of the two fluorescent proteins. Bichromatic reporters have recently been used to identify several regulators of tissue- and developmental stage-specific AS events (Kuroyanagi et al., 2006; Kuroyanagi et al., 2007; Ohno et al., 2008), and reveal the promise of using *C. elegans* as a model system to gain mechanistic insights into splicing regulation *in vivo*.

The availability of genome sequences for several *Caenorhabditis* species in addition to the sequencing of EST and cDNA libraries have enabled several large scale analyses of AS. A recent comparison of exons in *C. elegans*, *C. briggsae*, and *C. remanei*, revealed high conservation of alternative exons between these species, much higher than what has been observed in mammals separated by similar evolutionary distances (Irimia et al., 2008). This observation suggests that most functionally important AS events were already present in the common ancestor of *Caenorhabditis* species, and since the divergence of these species AS has not been extensively used to generate species diversity. Consistent with these isoforms having conserved functions, the temporal regulatory patterns of many of these AS events are also conserved (Rukov et al., 2007). In a separate study, a comparative genomic analysis between *C. elegans* and *C. briggsae* was used to identify conserved intronic sequences flanking alternative exons that might act as splicing regulatory elements (Kabat et al., 2006). This analysis identified many candidate cis-elements, several of which are known to be functional in vertebrates.

The recent development of a splicing microarray monitoring 352 exon skipping events in *C. elegans* has provided a means of analyzing AS at a more global level under various environmental conditions and developmental states. This microarray was first used to profile AS during development (Barberan-Soler and Zahler, 2008). Interestingly, nearly 20% of events monitored by the microarray were found to undergo strong changes in splicing between any
developmental stages. Moreover, among the classes of genes enriched with developmentally-regulated transcript isoforms were splicing factor genes. Several of these developmentally-regulated splicing factor isoforms can also be targeted by the NMD pathway (Barberan-Soler et al., 2009; Barberan-Soler and Zahler, 2008). While some of these differentially spliced isoforms are directly targeted by the NMD machinery, others are likely caused by indirect effects resulting from the misregulation of splicing factors. These results suggest that cross-regulation between these two processes can play a key role in the developmental control of gene expression. More recently, this microarray platform was utilized to identify AS events modulated by several splicing factors (Barberan-Soler et al., 2010). Interestingly, many transcripts are regulated by multiple factors, indicating that similar to mammalian systems, AS in *C. elegans* is subject to the combinatorial influence of splicing regulators. A number of these principles have been confirmed and elaborated by the use of RNA-Seq analyses, which will be discussed in greater detail in chapter 4. Collectively, these studies have demonstrated that *C. elegans* represents an excellent model system for studies of AS in vivo.

1.9 Overview of thesis research

The overall goal of my research has been to apply genome-wide analyses, as well as focused biochemical and molecular genetic approaches, to investigate the role of AS during evolution and development. In particular, this work has significantly benefited from the development of a quantitative AS microarray in our group (Pan et al., 2004), and later through advances in high-throughput sequencing methodologies (Blencowe et al., 2009). In Chapter 2, I describe how I have utilized microarray profiling to identify AS events differentially regulated between humans and chimpanzees (Calarco et al., 2007b). At the start of this study, we hypothesized that differences in AS likely contributed to the evolutionary divergence between these two species, but the extent to which these differences accumulated was not known. Moreover, we also felt that the identification of AS differences between these two species would facilitate the discovery of some recent molecular changes that have shaped the evolution of our species. This work has provided original insight into how AS may have contributed to human evolution, and has provoked further studies in this area of research.

As mentioned above, it has been thought that most RS domain proteins, which are expressed ubiquitously, contribute to tissue-specific splicing regulation through changes in their
expression levels. However, it was not known whether any tissue-specific SR or SR-related proteins existed. In Chapter 3, I describe how we have utilized a computational approach to create a comprehensive database of mammalian RS domain protein genes. Mining this database, I was able to identify the first example of a nervous system-specific SR-related protein (Calarco et al., 2009). In the remainder of the chapter, I demonstrate how this splicing factor acts to regulate AS in the nervous system by uncovering the network of isoforms that it controls, and how it is required for proper nervous system development \textit{in vivo}. This study has led to new models for tissue-specific AS regulation and for how RS domain proteins have evolved to increase molecular diversity in distinct tissue types and to coordinate gene expression programs during development.

In Chapter 4, I present our results from profiling AS during \textit{C. elegans} development, which expands upon previous studies through the use of a more comprehensive microarray platform and RNA-Seq analysis (Ramani et al., 2011). We hypothesized that AS is more prevalent in \textit{C. elegans} than what had been suggested by EST/cDNA coverage, and that a significant fraction of these splicing events are regulated differently across development. This is in fact what was observed in our studies. Additionally, in the second part of the chapter I focus on new resources I have developed to investigate AS regulation at single cell resolution and also to explore the functional roles of specific isoforms in \textit{C. elegans}. The genome-wide dataset and these additional resources have created a powerful set of tools for studies of AS regulation \textit{in vivo}, and will make it possible to address several interesting questions in the near future.

Taken together, these studies have demonstrated how new technologies can drive biological discovery, providing the means to address questions that were not previously tractable. More importantly however, this work has also led to multi-faceted discoveries into how AS has been utilized throughout evolution to contribute to molecular and cellular diversity.
Chapter 2

Global Analysis of Alternative Splicing Differences Between Humans and Chimpanzees

This chapter is derived from the following published article: J.A. Calarco, Y. Xing, M. Caceres, J.P. Calarco, X. Xiao, Q. Pan, C. Lee, T.M. Preuss, and B.J. Blencowe. *Genes and Development* (2007) 21(22): 2963-2975. Global analysis of alternative splicing differences between humans and chimpanzees. M. Caceres and T.M. Preuss provided the human and chimpanzee RNA samples. I performed all experiments and data analysis described in Figures 2.1-2.7 and tables 2.2-2.3. Y. Xing performed the genome alignments and I performed the data analysis in table 2.1. Table A1.1 can be found in appendix 1, and additional supporting documents can be found on the accompanying CD.
2.1 Abstract

Alternative splicing is a powerful mechanism affording extensive proteomic and regulatory diversity from a limited repertoire of genes. However, the extent to which AS has contributed to the evolution of primate species-specific characteristics has not been assessed previously. Using comparative genomics and quantitative microarray profiling, we performed the first global analysis of AS differences between humans and chimpanzees. Surprisingly, 6%–8% of profiled orthologous exons display pronounced splicing level differences in the corresponding tissues from the two species. Little overlap is observed between the genes associated with AS differences and the genes that display steady-state transcript level differences, indicating that these layers of regulation have evolved rapidly to affect distinct subsets of genes in humans and chimpanzees. The AS differences we detected are predicted to affect diverse functions including gene expression, signal transduction, cell death, immune defense, and susceptibility to diseases. Differences in expression at the protein level of the major splice variant of Glutathione S-transferase omega-2 (GSTO2), which functions in the protection against oxidative stress and is associated with human aging-related diseases, suggests that this enzyme is less active in human cells compared with chimpanzee cells. The results of this study thus support an important role for AS in establishing differences between humans and chimpanzees.
2.2 Introduction

A major challenge of the post-genomic era is to identify the set of molecular characteristics that account for human-specific traits. Recent global comparisons of human and chimpanzee genomes and transcriptomes have begun to identify molecular differences that could underlie some of the unique attributes of these primate species (Bustamante et al., 2005; Consortium, 2005; Khaitovich et al., 2006; Preuss et al., 2004). These differences stem from sequence and regulatory diversity acting on both protein coding and noncoding transcripts (Pollard et al., 2006; Prabhakar et al., 2006), as well as from structural genomic alterations including segmental duplications, large-scale copy number variations, and deletions (Cheng et al., 2005; Newman et al., 2005; Perry et al., 2006). Despite the recent accumulation of data revealing molecular differences between humans and chimpanzees, when I started in the lab additional possible sources of diversity, such as differential AS, had not yet been investigated.

Alternative splicing is the process by which splice sites in precursor mRNA transcripts are differentially selected to result in the production of different mRNA and protein isoforms (Blencow, 2006; Matlin et al., 2005). Two-thirds or more genes in human and mouse contain at least one alternative exon, and it is known that mammalian organs comprising many specialized cell types such as the brain are associated with relatively complex AS patterns (Fagnani et al., 2007b; Johnson et al., 2003; Yeo et al., 2004a). Surprisingly, comparisons of human and mouse transcript sequences have revealed that <20% of AS events have been conserved during the ~80- to 90-million-year interval separating these species (Modrek and Lee, 2003; Nurtdinov et al., 2003; Pan et al., 2005; Sorek et al., 2004b; Yeo et al., 2005). While these and related observations have suggested an important role for AS in the evolution of mammalian species (Ast, 2004; Xing and Lee, 2006), no experimental study has yet addressed the extent to which splicing patterns may differ between more closely related mammalian species such as humans and chimpanzees. The identification of orthologous human and chimpanzee transcripts with different splicing patterns represents a critical first step toward understanding the role of AS in the evolution of specific traits in these species.

In this study, we used comparative genomics and a quantitative AS microarray platform to compare AS patterns in human and chimpanzee tissues. Despite an overall sequence identity between human and chimpanzee genomic coding regions of 98%–99%, we observe that 6%–8%
of surveyed alternative exons display pronounced splicing level differences. These differences affect a primarily nonoverlapping subset of genes compared with the subset of genes that display differences in steady-state transcript levels between humans and chimpanzees. Alternative splicing differences between humans and chimpanzees are consistently detected when comparing corresponding cell and tissue types from multiple individuals from each species, and the associated genes have diverse and important functional roles, including regulation of gene expression, signal transduction, apoptosis, immune defense, and disease. Detection of increased expression of the major, functional splice variant of Glutathione S-transferase omega-2 (GSTO2) at the protein level in chimpanzee cells suggests that this enzyme could play a more active role in protection against oxidative stress and diseases associated with loss of such protection in chimpanzees compared with humans. Our results thus provide evidence that AS changes have evolved rapidly and in parallel with changes in transcriptional regulation, affecting an additional set of genes and associated gene functions. The identification of AS differences between humans and chimpanzees provides a new basis for understanding the molecular mechanisms underlying the evolution of primate species-specific characteristics.
2.3 Materials and Methods

2.3.1 Comparative genomic analysis of human and chimpanzee exons/flanking introns

Human mRNA and EST sequences were aligned to the human genome sequence as described previously (Kim et al., 2007) using human UniGene EST data (January 2006) (ftp://ftp.ncbi.nih.gov/repository/UniGene) and human genome assembly version hg17 (http://hgdownload.cse.ucsc.edu/goldenPath/hg17). Internal exons were detected as genomic regions flanked by two consensus splice sites. Orthologous chimpanzee sequences were identified using the University of California at Santa Cruz human versus chimpanzee pairwise genome alignments (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/vsPanTro2). The percent nucleotide divergence rate (number of substituted nucleotides within each region divided by the total number of nucleotides within the region) was determined for regions including each orthologous exon and the upstream and downstream 150 bp of flanking intronic sequence. Regions associated with elevated substitution rates were manually inspected for sequencing errors, and to ensure correct alignment to the orthologous genes and not to more distantly related paralogs.

2.3.2 Tissue samples, cell lines, and cell culture

Tissue samples were obtained post-mortem from nine humans (Homo sapiens), six common chimpanzees (Pan troglodytes), and six rhesus macaques (Macaca mulatta). Information on the sex, age, and origin of the samples from each individual is provided in Supplementary Table 2. Tissue samples were immediately frozen in liquid nitrogen after dissection and kept at -80°C. Brain cortical tissue samples were taken from the frontal pole (FP) of the left hemisphere of each species. Mouse brain cortex and heart tissue samples were obtained from approximately five mice, dissected immediately after sacrificing and flash-frozen in liquid nitrogen. Primary fibroblast and lymphoblast cell lines from human and chimpanzee individuals were a kind gift from Stephen Scherer (The Hospital for Sick Children, Toronto, Ontario, Canada). The lymphoblast cell lines (human and chimpanzee) were grown in RPMI medium supplemented with 15% FCS, sodium pyruvate, L-glutamine, and antibiotics. Fibroblasts were grown in alpha minimal essential medium supplemented with 10% FCS and antibiotics. HeLa cells (American Type Culture Collection) were used for all experiments.
involving the GSTO2 overexpression constructs and were grown in DMEM supplemented with 10% FBS and antibiotics. All cells were incubated at 37°C and 5% CO2.

2.3.3 Microarray hybridization, data extraction, and analysis

Total RNA was extracted from 1–2 g of each tissue sample using the Trizol reagent (Invitrogen) as per the manufacturer’s recommendations. Portions of the total RNA samples from the same tissue from each of the individuals were pooled and poly(A)+ mRNA was purified from these samples using oligo-dT cellulose resin (New England Biolabs), as described previously (Pan et al., 2004). cDNA synthesized from the pooled poly(A)+ RNA samples was separately labeled with cyanine 3 and cyanine 5 fluorescent dyes (Amersham Pharmacia), and hybridized in dye swap experiments to a custom human oligonucleotide microarray manufactured by Agilent Technologies, Inc., as described previously (Pan et al., 2004). Microarrays were washed and scanned with a GenePix 4000A scanner (Axon Instruments), and images were processed and normalized as described previously (Pan et al., 2004). Processed intensity values from the microarray scans were input into the GenASAP algorithm (Pan et al., 2004; Shai et al., 2006) to obtain confidence-ranked percent inclusion level predictions for 5183 unique cassette AS events.

2.3.4 RT–PCR assays and quantification

Pooled poly(A)+ RNA samples were normalized for RNA concentration by comparing the amplified band intensities to a portion of the coding region of the human beta-actin transcript. In addition, small aliquots of total RNA from each of the human, chimpanzee, and macaque individuals were analyzed by reverse transcription-polymerase chain reaction (RT–PCR) assays (Fig. 5; Supplementary Fig. 1). In each RT–PCR reaction, 0.2 ng of poly(A)+ RNA or 20 ng of total RNA were used as input and cDNA synthesis and amplification was performed using the One-Step RT–PCR kit (Qiagen) as per the manufacturer’s recommendations, with the following changes: Reactions were performed in a 10-uL volume, and 0.3 uCi of α-32P-dCTP was added to the reaction. The number of amplification cycles was 22 for beta-actin and 30 for all other transcripts analyzed. All reaction products were resolved by using 6% denaturing polyacrylamide gels. The gels were subsequently dried and analyzed using a Typhoon Trio PhosphorImager and software (Amersham). Percent inclusion levels from RT–PCR reactions were calculated as the percent of the isoform including an alternative exon over the total abundance of the isoforms.
including and excluding the alternative exon. In the case of candidates from the comparative genomic sequence analysis, the presence of novel bands on gels were also considered and validated by sequencing.

2.3.5 Plasmids

The full-length GSTO2 ORF was amplified by PCR from a plasmid (a kind gift from Dr. Richard Weinshilboum, Mayo Clinic College of Medicine, Rochester, MN) using the following primers: 5’-CGGAAGCTTATGTCTGGGGATGCGACCAGG-3’ and 5’-CGGGGA-TCTCAGCACAGCCCAAAGTCAAAG-3’. The insert was subcloned into pCMV-Myc (Rosonina et al., 2005) using HindIII and BamHI sites to generate pCMV-Myc-GSTO2. The construct expressing the cDNA corresponding to the GSTO2 isoform lacking exon 4 was generated by PCR amplification using pCMV-Myc-GSTO2 and the following primers: 5’-ATTCTTGAGTATCAGAACACCACCTTCTT-3’ and 5’-CTTACAAAATAGCTCCAAATTACATCTTTTGG-3’. The amplified product was then ligated with T4 DNA ligase (Fermentas) to generate pCMV-Myc-GSTO2-Δ4.

2.3.6 Transfection and Western blotting

HeLa cells were transfected with 24 ug of either pCMV-Myc-GSTO2 or pCMV-Myc-GSTO2-Δ4 with Lipofectamine 2000 (Invitrogen) as recommended by the manufacturer. Cells were harvested 48 h post-transfection. Total RNA was prepared as described above from both the transfected HeLa cells and the human and chimpanzee lymphoblast and primary fibroblast lines. Protein lysates were prepared by the addition of 100–200 uL of RIPA buffer (150 mM NaCl, 50 mM Tris-HCl at pH 7.5, 500 uM EDTA, 100 uM EGTA, 0.1% SDS, 1% Triton X-100, 1% sodium deoxycholate) to cell pellets, followed by lysis. Thirty micrograms of protein lysate from each transfected sample or 100 ug of human and chimpanzee lymphoblast cell lysates were run on 12% SDS–polyacrylamide gels. Immunoblotting was performed using anti-c-myc (Sigma), anti-GSTO2 (a kind gift of Dr. Richard Weinshilboum), or anti-alpha-tubulin (Sigma) antibodies and chemiluminescence reagents and secondary antibodies at dilutions recommended by the manufacturer.
2.3.7 Microarray data analysis

Only GenASAP values from the top half of the ranked dataset were analyzed, since %in predictions from this portion of different datasets have been shown correlate well with %in measurements from independent RT-PCR assays (Pan et al., 2004; Shai et al., 2006). GenASAP values were excluded from the analysis if one or more probes in a set of six probes for monitoring AS contained three or more mismatches to the chimpanzee genome sequence. This filtering criterion was employed since it has been reported that up to three mismatches in a 38mer probe can be tolerated without appreciable loss of hybridization signal (Castle et al., 2003). After these filtering steps, 1702 and 1727 cassette AS events remained for comparison between human and chimpanzee frontal cortex and heart, respectively. In addition to the above dataset, we analyzed 217 conserved AS events between humans, chimpanzees and mice using a mouse-specific microarray employing probe sets designed using the same criteria as those for the human-specific microarray employed in the present study (Fagnani et al., 2007b).

Gene expression levels were quantified using the normalized values for probes targeted to the constitutive exons flanking each profiled alternative exon. These probe values were transformed to the hyperbolic arcsine (arcsinh) domain and averaged to result in a gene expression value on the arcsinh scale. Differences in the arcsinh values were then calculated to determine the relative fold change in transcript levels.

In order to compare the patterns of divergence between AS and transcription profiles for the 217 conserved human, chimpanzee and mouse AS events we used the Spearman correlation coefficient for each pairwise comparison as a measure of similarity between species. Transcript abundance measurements were calculated as a fold over median value in the log_2 domain, and standardized by dividing by the standard deviation of the distribution of values. For splicing comparisons, percent inclusion levels were converted to relative isoform ratios in the log_2 scale, and similarly standardized by dividing by the standard deviation of the distribution of values. Spearman correlation values were then calculated for each of the pairwise comparisons.

2.3.8 Estimate for the percent of multi-exon human and chimpanzee genes with pronounced splicing level differences.

Analyses of cDNA and EST sequences have established that ~10% of human exons are cassette alternative exons (Kim et al., 2007), and that at least 60% of human genes undergo AS
(Blencowe, 2006; Matlin et al., 2005). Therefore, there are ~1.67 cassette alternative exons per gene with AS, taking into account 10 internal exons per gene. We determined in our analysis that ~7% of cassette exons have pronounced differences in splicing levels between human and chimpanzee. Therefore 4.19% (7% / 1.67) of human and chimpanzee orthologous genes are expected to have differently spliced cassette alternative exons.
2.4 Results

2.4.1 Analysis of human and chimpanzee splicing patterns by comparative genomic and microarray profiling strategies

To compare global splicing patterns between humans and chimpanzees and identify individual AS events that differ between these species, we applied a comparative genomics strategy together with a previously described quantitative AS microarray profiling system (Fig. 2.1; Pan et al., 2004; Shai et al., 2006; refer to Materials and Methods). In these comparisons, we focused on splicing patterns involving orthologous exons. While it is possible that some splicing differences could involve exons that are unique to either genome, the lack of substantial transcript sequence data and of a polished chimpanzee genomic sequence makes such events difficult to detect in a reliable manner (Newman et al., 2005). Moreover, given the 98%–99% sequence identity between the coding regions of human and chimpanzee genomes (Consortium, 2005), few exons are expected to be unique to either genome. Accordingly, in the comparative genomics strategy, we aligned 120,951 orthologous human and chimpanzee exons and the 150 base pairs (bp) of intron sequence flanking these exons and screened these regions for high substitution rates. Previous studies have shown that short and degenerate motifs referred to as splicing enhancers and silencers, which specify splice site recognition and regulate AS, are enriched in exonic and intronic regions proximal to splice sites (Brudno et al., 2001; Yeo et al., 2004b; Yeo et al., 2007; Zhang et al., 2003; for reviews, see Blencowe, 2006; Chasin, 2007; Matlin et al., 2005). Consequently, we expected that frequent nucleotide changes in exons and the flanking intron regions would be predictive of splicing differences between humans and chimpanzees.

2.4.2 Role of nucleotide substitutions in generating differences between human and chimpanzee splicing patterns

High substitution rates were rarely found in the aligned orthologous exons and flanking intron sequences: Of these regions, 0.3% displayed substitution rates of >5%, and substitution rates >10% were found in <0.1% of the data set (Table 2.1). To assess whether regions with >5% substitution rates are predictive of splicing differences, semi-quantitative RT–PCR assays were performed using primers specific for neighboring exon sequences and poly(A)+ RNA samples from human or chimpanzee frontal cortex and heart tissues. These tissue RNA samples were
Figure 2.1: Strategies used to identify alternative splicing differences between humans and chimpanzees.

(Left panel) In the comparative genomics strategy, regions including orthologous human and chimpanzee exons (red boxes) and the flanking 150 nucleotides of intron sequence were aligned. Nucleotide substitutions (indicated by Xs) in these regions were scored, and events displaying >5% substitution rates were analyzed for alternative splicing differences by RT–PCR assays using poly(A)+ RNA from human and chimpanzee tissues (see main text for details). (Right panel) In the quantitative alternative splicing microarray profiling strategy, labeled cDNA from the same poly(A)+ RNA samples were hybridized to an alternative splicing microarray designed to monitor inclusion levels of cassette-type alternative exons. Each alternative splicing event is monitored by a set of six oligonucleotide probes (black horizontal lines) (Pan et al. 2004). Predictions for alternative splicing differences between the corresponding human and chimpanzee tissues were then validated by RT–PCR assays.
Table 2.1: Data from the comparative genomic analysis of orthologous exon regions between humans and chimpanzees.

Number of exons/flanking intron regions aligned between humans and chimpanzees, and the number these regions with >5% and >10% substitution rates are listed. Also, the number of sequence defined constitutive and alternative exons (multiple EST/cDNA evidence supporting alternative splicing) with >5% nucleotide substitution rates are listed. The percentage contribution of each subgroup to the total dataset is displayed in the third column.

<table>
<thead>
<tr>
<th>Exons/Flanking regions</th>
<th>Number</th>
<th>Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>120,951</td>
<td>100%</td>
</tr>
<tr>
<td>Substitution rate &gt;5%</td>
<td>394</td>
<td>0.33%</td>
</tr>
<tr>
<td>- Constitutive exons</td>
<td>348</td>
<td>0.29%</td>
</tr>
<tr>
<td>- AS exons</td>
<td>46</td>
<td>0.04%</td>
</tr>
<tr>
<td>Substitution rate &gt;10%</td>
<td>100</td>
<td>0.08%</td>
</tr>
</tbody>
</table>
pooled, in each case, from several adult individuals (Appendix 1, table A1.1). Of 31 regions with >5% substitution rates selected for analysis, 14 contained exons that displayed evidence of AS in humans from analysis of the corresponding EST/cDNA (complementary DNA) sequences. Five of the 31 regions displayed splicing level differences between humans and chimpanzees in at least one of the two tissues (see below). Interestingly, all five of these isoform ratio differences involved exons included among the 14 alternative exons identified in the test set. At this validation rate, <20 (~0.02%) of the 120,951 orthologous exons associated with high substitution rates would be expected to display splicing level differences between humans and chimpanzees.

Since substitution rates of <5% could also result in splicing differences, especially if the substitutions target important cis-acting elements, a splicing element scoring system (Stadler et al., 2006) was also employed to predict alterations in splicing levels between humans and chimpanzees. However, RT–PCR analysis of 20 exons (including several alternative exons) associated with high differential scores did not yield detectable splicing differences (data not shown). This observation could reflect a current lack of knowledge of the full set of cis-elements and mechanisms responsible for splice site selection in endogenous transcripts. It is possible that relevant changes involving defined cis-regulatory elements could occur outside of the regions we analyzed. Another possibility is that selection pressure acting to reduce the frequency of substitutions proximal to most splice sites in humans and chimpanzees (Consortium, 2005; Gazave et al., 2007) results in a limited role for cis-acting differences in altering splicing patterns between the two species. Moreover, consistent with the results described above, high substitution rates could have a greater impact on the splicing levels of alternative exons, since the regions surrounding these exons are, on average, more highly conserved than the regions surrounding constitutive exons (Sorek and Ast, 2003; Sugnet et al., 2006; Yeo et al., 2005), and therefore potentially more susceptible to rare mutations that disrupt important splicing regulatory elements.

2.4.3 Detection of alternative splicing differences between humans and chimpanzees by microarray profiling

Splicing differences between humans and chimpanzees could also occur as a consequence of alterations in one or more trans-acting splicing regulators. Accordingly, to assess the extent of AS differences between human and chimpanzee tissues that are not the consequence
of elevated rates of substitution in the alternative exons and flanking intron sequences, I
compared splicing patterns using quantitative AS microarray profiling (Fig. 2.1; refer to
Materials and Methods). An AS microarray capable of profiling ~5000 cassette-type AS events
mined from human ESTs and cDNA sequences (Fig. 2.1) was hybridized with dye-labeled
cDNAs synthesized from the frontal cortex and heart mRNA samples analyzed above. This
microarray contains sets of exon body and splice junction probes that are capable of monitoring
the levels of inclusion of both the EST/cDNA-mined human cassette alternative exons and the
orthologous exons identified in alignments of the corresponding chimpanzee genome sequences.
The processed microarray data was analyzed using the Generative model for Alternative Splicing
Array Platform (GenASAP) algorithm (Pan et al., 2004; Shai et al., 2006), which generates for
each profiled AS event a confidence-ranked estimate for the fraction of transcripts that include
an alternative exon, represented below as percent inclusion (%in) values. After filtering steps
(see Materials and Methods), the levels of inclusion of ~1700 orthologous exons were available
for comparison between the corresponding human and chimpanzee tissues. None of these
profiled AS events were associated with a substitution rate of >5%.

As expected, given the close evolutionary relationship between humans and chimpanzees,
I found that the majority of the profiled orthologous exons in the two species have similar
splicing levels. For example, when sorting genes based on the magnitude of the percent inclusion
difference between human and chimpanzee frontal cortex and heart, ~80% of the profiled exons
display percent inclusion differences of <15%, which is close to the limit of sensitivity of our AS
microarray profiling system (Fig. 2.2 A; (Pan et al., 2006)). However, in addition to exons with
similar splicing levels, a surprisingly large fraction of exons (6%–8%) show pronounced percent
inclusion differences (i.e., >25%in). RT–PCR experiments were performed on 37 representative
AS events displaying a range from no percent inclusion difference to a >25%in difference
between the human and chimpanzee frontal cortex and/or heart. Of these events, 30 (81%) were
validated as having the expected differential splicing patterns (Figs. 2.4, 2.5, below; additional
data not shown). Thus, despite the remarkable degree of conservation between the coding
regions of the human and chimpanzee genomes, a substantial number of alternative exons that
are not associated with high substitution rates in the alternative exons and flanking intron regions
display pronounced splicing level differences between the two species.
Figure 2.2: Microarray profiling reveals the extent of alternative splicing differences between human and chimpanzee orthologous exons, and the degree of divergence between splicing patterns over different evolutionary time periods.

(A) Color spectrum plots indicating the number and magnitude of alternative splicing differences between human and chimpanzee frontal cortex and heart tissues. The Y-axes indicate the number of alternative splicing events profiled; these are sorted according to the magnitude of the absolute value of the percent exon inclusion level (%in) difference between the human and chimpanzee tissue being compared. The magnitude of the percentage inclusion difference is indicated by the color scale on the right. (B) Cumulative distribution plot displaying the distribution of percentage inclusion differences when comparing microarray data for 217 conserved alternative splicing events between the following pairs of tissues: human and chimpanzee frontal cortex (blue line), human and chimpanzee heart (red line), human frontal cortex and mouse cortex (green line), and human and mouse heart (purple line). (C) Spearman correlation coefficients are shown for pairwise comparisons between alternative splicing levels (black numbers) and transcript levels (purple numbers) for the set of 217 orthologous genes analyzed in human, chimpanzee, and mouse tissues in B. Double arrows indicate the pairs of species compared. (Hs) Homo sapiens; (Pt) Pan troglodytes; (Mm) Mus musculus.
2.4.4 Evidence for stabilizing selection pressure acting to preserve the majority of splicing levels of orthologous human, chimpanzee, and mouse alternative exons

The similarity between the splicing levels for the majority of the profiled orthologous exons could reflect stabilizing selection pressure acting to conserve the inclusion levels of most human and chimpanzee orthologous exons. However, it is also possible that insufficient evolutionary time has accumulated to result in a higher incidence of pronounced inclusion level differences than observed above. To investigate these possibilities, I compared the extent of divergence between AS profiles representing a longer evolutionary time span, namely the 80- to 90-million-year period separating the common ancestor of human or chimpanzee and mouse. Percent inclusion values for 217 AS events conserved between human and mouse were obtained from RT–PCR-validated, quantitative AS microarray profiling data from mouse heart and brain cortex tissues (Fagnani et al., 2007b) and compared with percent inclusion values from the human and chimpanzee datasets described above (Fig. 2.2 B and C; see Materials and Methods).

Notably, although the splicing levels of the mouse exons are also very similar to the orthologous exons in the corresponding human and chimpanzee tissues, they have diverged to a significantly greater extent in both tissues (P < 1 x 10^-5, Mann-Whitney U-test) (Fig. 2.2 B, cf. blue and green lines, and red and purple lines). Importantly, these increased differences in the AS profiles were not due to any inherent percent inclusion level biases in the datasets, since “control” profiles generated by randomly pairing exons from the same datasets were not significantly different (data not shown). Moreover, transcript level profiles obtained from the same datasets (see the Supplemental Material) also correlate to a higher degree between human and chimpanzee than between either primate species and mouse (Fig. 2.2 C).

The analysis described above reveals that a significant proportion of AS patterns involving orthologous human, chimpanzee, and mouse alternative exons are conserved in the corresponding tissues, despite sharing an 80- to 90-million-year period of divergence. These results therefore support the conclusion that stabilizing selection pressure has acted to preserve the inclusion levels of the majority of orthologous exons between mammals. Nevertheless, 6%–8% of orthologous human and chimpanzee exons display pronounced deviations from this overall pattern of conservation. Such exons, in addition to those identified in the comparative
genomic analysis, most likely represent AS events that could underlie important functional and phenotypic differences between the two species.

2.4.5 Primarily nonoverlapping subsets of genes have evolved splicing and transcript level differences between humans and chimpanzees

The 6%–8% of exons with pronounced splicing level differences detected in the above analysis involve ~4% of the microarray profiled genes. A comparable proportion of orthologous genes has been found to display at least twofold transcript level differences between corresponding human and chimpanzee tissues (Khaitovich et al., 2005; Preuss et al., 2004). Remarkably, however, despite a parallel divergence in AS and transcriptional patterns during the evolution of humans and chimpanzees, primarily nonoverlapping subsets of genes are associated with these changes (Fig. 2.3). By sorting genes according to the magnitude of the microarray-detected percent inclusion AS level differences (Fig. 2.3, columns 1 and 3), it is evident that relatively few genes display coincident changes in transcriptional levels (Fig. 2.3 in cf. columns 1 and 2, and columns 3 and 4). Among the 160–240 splicing events with predicted percent inclusion differences >20% in frontal cortex and/or heart tissues, only 25% are predicted to also have twofold or greater steady-state transcript level changes. This observation extends previous results indicating that predominantly different subsets of genes are regulated in a cell-specific/tissue-specific or activity-dependent manner at the transcriptional and AS levels (Fagnani et al., 2007b; Ip et al., 2007; Le et al., 2004; Li et al., 2006; Pan et al., 2004). In particular, the present results show that these two layers of regulation have evolved rapidly to affect different subsets of genes, and indicate that AS has served as an additional mechanism for diversifying gene regulation during the 5–7 million years of evolution separating humans and chimpanzees.
Figure 2.3: Transcript and AS level differences between humans and chimpanzees involve largely distinct subsets of genes.

The color spectrum plots compare splicing level differences and transcript level differences for the same set of genes expressed in the frontal cortex and heart. For each tissue comparison, the left column shows splicing differences (measured as the magnitude of percent inclusion difference, columns 1 and 3) and the right column measures differences in gene transcript level (measured as the magnitude of the hyperbolic arcsine [arcsinh] difference, columns 2 and 4). An hyperbolic arcsine difference of ~0.4 corresponds to a 1.5-fold change in expression level, and a difference of ~0.7 corresponds to a twofold difference in expression level.
2.4.6 Alternative splicing differences between humans and chimpanzees affect transcripts from genes associated with diverse cellular functions

Genes with AS differences between humans and chimpanzees detected using the methods described above, and confirmed by the RT–PCR assays, are associated with diverse biological processes including gene expression, signaling pathways, immune defense, and disease (Table 2.2). In most cases the AS differences were observed in both frontal cortex and heart tissues (Fig. 2.4). An example of an AS difference detected using the comparative genomics approach involves transcripts encoding the TATA-box-binding protein-associated factor 6 (TAF6) (Fig. 2.4; see Table 2.2 and below for additional examples and information). Examples of AS differences detected in the microarray data affect transcripts encoding the herpesvirus entry mediator receptor (HVEM/TNFRSF14), the spectrin-binding protein gamma-adducin (ADD3), ADP-ribosylation factor-like 3 (ARL3), serine/arginine (SR)-repeat family protein splicing factor of 40 kDa (SRp40/SFRS5), and GSTO2 (Fig. 2.4; see Table 2.2 and below for additional examples and information). To assess whether these observed differences are specifically associated with the human or chimpanzee lineage, I performed RT–PCR assays using RNA extracted from frontal cortex and heart tissue from rhesus macaques, an Old World monkey, as an outgroup comparison. The results of this analysis are summarized below and in Table 2.3, and representative examples are shown in Figure 2.5 A.
<table>
<thead>
<tr>
<th>Functional category</th>
<th>Gene name</th>
<th>GO processes/known features</th>
<th>Alternative splicing-altered protein domain(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signaling</td>
<td>Inositol hexaphosphate kinase 2 (IHPK2)</td>
<td>Inositol trisphosphate-3 kinase activity</td>
<td>IHPK</td>
</tr>
<tr>
<td></td>
<td>Hypothetical protein LOC2020686</td>
<td>Phosphatidylinositol-4 kinase activity</td>
<td>PI4K</td>
</tr>
<tr>
<td></td>
<td>ADP-ribosylation factor-like 3 (ARL3)</td>
<td>G-protein-coupled receptor protein signaling pathway</td>
<td>GTPase</td>
</tr>
<tr>
<td>Gene regulation</td>
<td>TBP-associated factor 6 (TAF6)</td>
<td>TFIID complex, regulator of transcription</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>SRp40 (SFRS5)</td>
<td>Splicing regulation</td>
<td>RRM, RS</td>
</tr>
<tr>
<td></td>
<td>Poly(A)-binding protein cytoplasmic 4 (PABPC4)</td>
<td>RNA binding, poly(A) binding</td>
<td>ND</td>
</tr>
<tr>
<td>Immune response</td>
<td>Toll/Interleukin-1 receptor 8 (TIR8)</td>
<td>Immune response</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>Herpesvirus entry mediator (HIVEM)/TNFRSF14</td>
<td>Immune response</td>
<td>TM, cytoplasmic</td>
</tr>
<tr>
<td>Other</td>
<td>Glutathione S-transferase omega 2 (GSTO2)</td>
<td>Glutathione transferase activity, metabolism of xenobiotics</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>Roundabout 1 homolog (ROBO1)</td>
<td>Axon guidance, dyslexia candidate gene</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>FSHD region gene 1 (FRG1)</td>
<td>Facioscapulohumoral disease candidate</td>
<td>Actin cross-linking</td>
</tr>
<tr>
<td></td>
<td>γ-Adducin (ADD3)</td>
<td>Calmodulin binding, structural constituent of cytoskeleton</td>
<td>Spectrin binding</td>
</tr>
<tr>
<td></td>
<td>Zonula Occludens protein (ZO-1)</td>
<td>Tight junction protein</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>Bridging Integrator 1 (Bln1)</td>
<td>Regulation of endocytosis</td>
<td>BAR</td>
</tr>
<tr>
<td></td>
<td>Calbindin 2 (CALB2)</td>
<td>Calcium ion binding</td>
<td>EF hand</td>
</tr>
<tr>
<td></td>
<td>Chr. 14 ORF 153</td>
<td>Uncharacterized</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>Unc-84 homolog (UNC84B)</td>
<td>Microtubule binding, mitotic spindle organization, and biogenesis</td>
<td>ND</td>
</tr>
</tbody>
</table>

Functional annotations from the Gene Ontology database or from literature searches are listed. Protein domains predicted to be altered by differential alternative splicing are also listed. In most cases, differential alternative splicing was detected in both frontal cortex and heart; exceptions are in IHPK2, frontal cortex only; Bln1, frontal cortex only; CALB2, heart only; and UNC84B, heart only. (IHPK) Inositol hexaphosphate kinase domain; (PI4K) phosphatidylinositol-4 kinase domain; (GTPase) Ras-like GTPase domain; (RRM) RNA recognition motif (R); (BAR) BIN/amphiphrin/Rvs domain involved in endocytosis; (EF hand) calcium-binding domain; (ND) no known domain is affected by alternative splicing.

Table 2.2: Genes associated with different cellular functions with experimentally validated splicing level differences between humans and chimpanzees
Figure 2.4: Examples of alternative splicing differences between humans and chimpanzees confirmed by semiquantitative RT–PCR and sequencing.

RT–PCR assays were performed using primers specific for sequences in constitutive exons flanking the alternative exons predicted to be differentially spliced by the comparative genomic or alternative splicing microarray profiling analyses. Corresponding tissues from human (Hs) and chimpanzee (Pt) are indicated. Major splice isoforms that include and skip alternative exons (black boxes) are indicated on the right of each panel. Diagrams below each gel illustrate the predicted consequence of alternative splicing changes at the protein and/or transcript levels for TAF6, ADD3, the SR-repeat family protein splicing factor of 40 kDa (SFRS5/SRp40) and GSTO2 (refer to Table 1 and main text for details). Protein domains are labeled as follows: (H) head domain; (N) neck domain; (C) C-terminal tail region known to interact with spectrin; (RRM1) RNA recognition motif 1; (RRM2) RNA recognition motif 2; (RS) arginine/serine-rich domain; (GST) glutathione S-transferase domain. The stop sign indicates the insertion of a PTC. The asterisk indicates an additional TAF6 splice isoform detected in human but not chimpanzee tissue, as confirmed by sequencing.
**Table 2.3: Splicing differences found to be human- or chimpanzee-specific from RT-PCR assays of macaque tissue RNA as an outgroup comparison.**

The table lists the species-specific nature of the splicing differences identified between human and chimpanzee. A human-specific AS difference indicates that human splicing pattern for a particular gene is markedly different from chimpanzee and macaque patterns. Chimpanzee-specific splicing differences were defined as those that appeared visibly different from human and macaque splicing patterns.

<table>
<thead>
<tr>
<th>Gene Name and Symbol</th>
<th>Specificity of AS Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glutathione S-Transferase Omega 2 (GSTO2)</td>
<td>Human Specific</td>
</tr>
<tr>
<td>TBP-associated Factor 6 (TAF6)</td>
<td>Human Specific</td>
</tr>
<tr>
<td>Gamma Adducin (ADD3)</td>
<td>Chimpanzee Specific</td>
</tr>
<tr>
<td>Splicing Factor, Arginine and Serine Rich 5 (SFRS5/SRp40)</td>
<td>Human Specific</td>
</tr>
<tr>
<td>Herpesvirus Entry Mediator (HVEM/TNFRSF14)</td>
<td>Human Specific</td>
</tr>
<tr>
<td>ADP-ribosylation factor-like 3 (ARL3)</td>
<td>Chimpanzee Specific</td>
</tr>
</tbody>
</table>
Figure 2.5: Comparisons of AS levels in RNA samples from macaque, human, and chimpanzee individuals.

(A) RT–PCR assays were performed using rhesus macaque total RNA from frontal cortex and heart samples from a number of different individuals. RT–PCR primers were designed to anneal to the exons neighboring each alternative exon, resulting in the amplification of two products (the isoforms including and skipping the alternative exon, as indicated in each of the panels). Asterisks denote novel isoforms observed primarily in one species but not the other. The transcripts shown correspond to the ADD3 (top panel), GSTO2 (second panel), HVEM/TNFRS14 (third panel), and ARL3 (bottom panel) genes. The macaque is used as an outgroup to define the ancestral splicing pattern for each gene. Examples of chimpanzee lineage-specific splicing differences (ADD3 and ARL3 transcripts) and human lineage-specific splicing differences (GSTO2 and TNFRSF14) are shown (see also Supplementary Table 3 for additional examples). (B) RT–PCR experiments were performed using total RNA from each of the individual samples that were pooled for analysis in the comparative genomic and microarray experiments. The same alternative splicing events displayed for macaque experiments in A are also shown in B for comparison. The labels for each gel lane represent the macaque, human, and chimpanzee individuals listed in Appendix 1, Table A1.1.
2.4.7 Alternative splicing differences are consistent between human and chimpanzee individuals

Since the analyses described above employed samples of poly(A)+ mRNA pooled, in each case, from several human individuals and from several chimpanzee individuals (Appendix 1, Table A1.1), it was important to assess the extent to which the AS differences might be explained by variations between individuals from each species. Individual samples comprising the pools of tissue mRNAs were therefore analyzed separately for splicing level differences. In each case examined, similar AS differences were observed between the several human and chimpanzee individuals (Fig. 2.5 B; data not shown). These data indicate that it is highly unlikely that the overall differences in splicing levels measured using the pooled samples are attributed to differences associated with individual variation within each species. This conclusion was further supported by an analysis of AS level differences in primary fibroblasts and lymphoblastoid cells, also from multiple additional individuals from each species (Fig. 2.6). Again, in all cases examined, similar species-associated splicing level differences were detected between the different individuals as detected in the tissue samples (Fig. 2.6; see below). The results from this last experiment further indicate that the majority of the splicing level differences we observed are unlikely related to possible contributions from environmental or physiological differences, such as diet or stress, since the corresponding cell lines from both species were cultured in parallel for several weeks under identical growth conditions, prior to harvesting.
Figure 2.6: Alternative splicing patterns are similar in cell lines from individuals within a species but consistently differ between species. RT-PCR experiments were performed using total RNA samples prepared from primary fibroblast and lymphoblastoid cell lines from different human and chimpanzee individuals. Primer pairs annealing to exons neighboring alternative exons were used to amplify products including or skipping the alternative exon of interest (indicated by icons with or without black exon, respectively). Asterisks highlight isoforms identified only in one species. Labels above each gel lane indicate the species and individual being analyzed. Hs; Homo sapiens, Pt; Pan troglodytes.
2.4.8 Examples of validated human and chimpanzee alternative splicing differences

The AS difference detected in TAF6 transcripts is human lineage specific and is pronounced in both the frontal cortex and heart (Fig. 2.4). TAF6 is a subunit of the general transcription factor TFIID, which is involved in gene activation (Sauer et al., 1995). It has been reported that TAF6 isoforms are associated with the control of apoptosis and cell cycle arrest (Bell et al., 2001; Wang et al., 1997; Wang et al., 2004a). The AS difference found between humans and chimpanzees lies in the 5′ untranslated region (UTR), and could therefore affect TAF6 regulation at the post-transcriptional or translational levels. Consistent with this proposal are previous observations of AS events in 5′ UTRs that affect translational efficiency (Singh et al., 2005; Wang et al., 1999). Such a difference impacting TAF6 expression could, in turn, account for some of the differences in transcriptional profiles that have been observed between humans and chimpanzees (Khaitovich et al., 2006; Preuss et al., 2004).

The variation in AS of SRp40, which is also human lineage specific, is intriguing in light of the known roles of SR family members in both constitutive and regulated splicing (Graveley, 2000; Lin and Fu, 2007; Sanford et al., 2005). These proteins generally function in splicing by binding to exonic enhancer sequences via their RNA recognition motifs (RRMs) (Fig. 2.4). The AS difference in SRp40 transcripts results in the increased inclusion of a highly conserved premature termination codon (PTC)-containing exon (located between exons 4 and 5) in human transcripts, in both frontal cortex and heart tissue, and correspondingly reduced levels of the shorter transcript encoding the full-length protein, particularly in the heart. Several RNA-binding proteins including SR family members have been shown previously to regulate their own expression levels by activating the splicing of PTC-containing isoforms that are subsequently targeted by the process of nonsense-mediated mRNA decay (Lareau et al., 2007b; Ni et al., 2007 and references within). Individual SR family members are essential and tightly regulated proteins (Graveley, 2000; Lin and Fu, 2007; Sanford et al., 2005). It is therefore possible that the differential expression of the two SRp40 isoforms described above could be related to some of the splicing differences we observed between humans and chimpanzees, including those that are not associated with nucleotide substitutions.
The AS difference affecting GSTO2 transcripts results in pronounced skipping of exon 4 of this gene, specifically in human frontal cortex tissue (Fig. 2.4; Table 2.3). GSTO2 and its closely related paralog GSTO1 are important enzymes in the biotransformation of endogenous and exogenous compounds and in protection against oxidative stress (Schmuck et al., 2005), which plays an important role in defense against aging-associated diseases including cancers and neurodegeneration. In this regard, it is interesting to note that in some studies single nucleotide polymorphisms in GSTO2 and GSTO1 have been tentatively linked to certain human cancers and the age at onset of familial Alzheimer’s and Parkinson’s disease (Li et al., 2003; Pongstaporn et al., 2006; Whitbread et al., 2005; see Discussion).

An important question raised by the present and previous studies on AS is the extent to which differences in splice variant levels detected at the transcript level correspond to actual differences at the protein level. It should be borne in mind that comparisons of datasets from large-scale protein mass spectrometry analyses and from microarray profiling of mRNA from the same mammalian tissues have revealed good overall correlations between steady-state protein and transcript abundance (Kislinger et al., 2006). To investigate this question in the context of splice variants that differ between humans and chimpanzees, we next asked whether the AS level difference involving skipping of the frame-preserving exon 4 of GSTO2 transcripts could result in species-specific differences in the expression of GSTO2 variants at the protein level.

2.4.9 Expression of the functional splice variant of GSTO2 differs between humans and chimpanzees at the protein level

To initially assess the stability of the GSTO2 protein variants, I transiently expressed expression vectors containing c-myc epitope-tagged GSTO2 cDNAs, with and without exon 4, in HeLa cells. RT–PCR experiments using primers specific to constitutive exon sequences 5´ and 3´ to exon 4 indicated that comparable levels of the two splice variant transcripts are produced (Fig. 2.7 A). Surprisingly, however, western blotting of the protein lysates from the transfected cells with an anti-myc epitope antibody revealed that only the GSTO2 variant containing exon 4 results in significant protein expression (Fig. 2.7 B,C); minor levels of the GSTO2 variant lacking exon 4 could be detected, but only after prolonged exposure of the blot (Fig. 2.7 C; data not shown). These results indicate that GSTO2 splice variants lacking exon 4 are unstable at the protein level, and that levels of active GSTO2 are therefore likely determined by the relative expression of the splice variant containing exon 4. Also in support of this proposal are recent
Figure 2.7: Increased skipping of exon 4 in GSTO2 transcripts reduces overall levels of functional GSTO2 protein in humans compared with chimpanzees.

(A) RT–PCR experiments were performed to measure transcript abundance of individual myc-tagged GSTO2 isoforms transfected into HeLa cells. Primers were designed to anneal to exons neighboring the alternative exon in order to amplify isoforms including and skipping exon 4. Beta-Actin transcript levels were measured to normalize for input of total RNA between samples. (B) Western blotting experiments on the same transfected samples from A. Anti-c-myc antibodies were used to detect protein expression of individual GSTO2 splice variants. Alpha-Tubulin protein levels were measured to control for loading input. (C) Measure of the relative protein/mRNA abundance for each of the transfected samples in A and B. Measurements represent the average of three independent transfection experiments, and standard deviations are shown. (D) Western blotting using samples from human and chimpanzee lymphoblastoid cells and antibodies to detect endogenous GSTO2. Alpha-Tubulin protein levels were measured to control for loading input.
findings indicating that residues overlapping exon 4 of GSTO2 are important for both the stability and activity of this class of enzymes in vitro (Schmuck et al., 2005). From comparisons of the levels of the GSTO2 variants between different tissues and cell types it is apparent that increased skipping of exon 4 can result in reduced steady-state levels of exon 4-containing transcripts in humans compared with chimpanzees and other nonhuman primates, although in some cell and tissue sources differences in overall GSTO2 transcript levels may also account in part for increased levels of this variant in chimpanzees (Figs. 2.4, 2.5 and 2.6). Western blotting of lysates from lymphoblastoid cells from two human and two chimpanzee individuals confirms that the exon 4-containing splice variant is expressed at higher levels in the chimpanzee than human cells (Fig. 2.7 D). This result therefore suggests that GSTO2 is likely more active in chimpanzee cells compared with the corresponding human cells.
2.5 Discussion

This study represents the first large-scale comparative analysis of AS patterns in humans and chimpanzees, and indeed between any pair of species separated by a relatively short period of evolutionary history. The results indicate that at least 4% of genes in humans and chimpanzees have one or more cassette alternative exons that display pronounced splicing level differences between the two species (refer to Materials and Methods). This is comparable with the proportions (2%–8%) of orthologous human and chimpanzee protein coding genes reported to have significant transcript level differences in corresponding tissues (Khaitovich et al., 2005; Preuss et al., 2004). However, the majority of orthologous human and chimpanzee genes with splicing level differences do not overlap genes that display transcript level differences. These results indicate that AS has evolved rapidly to substantially increase the number of protein coding genes with altered patterns of regulation between humans and chimpanzees. Furthermore, the AS differences identified between humans and chimpanzees are predicted to affect a diverse range of critical functions, including the ability of cells to protect against oxidative damage.

My result revealing a difference in the level of expression of the active splice variant of GSTO2 between humans and chimpanzees has interesting implications for how some AS differences could impact the evolution of important physiological and phenotypic differences between two species. Although GSTO2 transcripts that include and skip exon 4 have similar stabilities, only minor levels of protein expression were detected from the exon 4-skipped splice variant. This indicates that skipping of this exon leads to the expression of an unstable protein. Thus, even though it is unlikely that the exon 4-skipped splice variant of GSTO2 has a significant functional role at the protein level, differential inclusion of exon 4 appears to impact, at least in part, the relative expression levels of GSTO2 enzyme between humans and chimpanzees. This finding raises an important question: What is the extent to which AS events result in the expression of stable and functionally distinct protein products? Recent structural modeling of several splice variants from genes in the 1% of the genome surveyed by the ENCODE Consortium has predicted that many AS events are unlikely to result in correct protein folding, since the alternatively spliced regions are expected to often disrupt core structural domains (Tress et al., 2007). Based on this observation, it was concluded that a large proportion
of splice variants detected in sequenced transcripts may not have functional roles. An alternative possibility, which is supported by the results in the present study, is that some AS events that do not lead to expression of functional or active protein could nevertheless impact overall levels of functional protein, either as a consequence of regulated splicing decisions within a species, or as a consequence of the evolution of differences in splicing levels between species.

In regard to the above, what are the possible functional consequences of altered splicing levels of GSTO2 exon 4? As mentioned earlier, GSTO2 functions in the biotransformation of toxic compounds and in the protection against oxidative stress, and SNPs in the GSTO2 gene, as well as in the paralogous GSTO1 gene, have been tentatively linked to human aging-associated diseases including cancer (Pongstaporn et al., 2006), and Alzheimer’s and Parkinson’s diseases (Li et al., 2003). GSTO2 appears to be more highly expressed in frontal cortex compared with heart, and it is also detected in other cells and tissues (Wang et al., 2005a; Whitbread et al., 2005). Moreover, the differences I detected between human and chimpanzee GSTO2 transcripts appear to be quite specific, since we did not detect differences between the splicing or expression patterns of GSTO1 (data not shown), which nevertheless is closely related to GSTO2, although it has distinct substrate specificities (Schmuck et al., 2005; Wang et al., 2005a). Reduced expression levels of GSTO2 in human cells would therefore be expected to specifically impact GSTO2-dependent functions. Intriguing in this regard are apparent major differences between humans and chimpanzees in relation to aging-associated neurodegeneration. Normal brain aging in humans is associated with the aberrant phosphorylation of the microtubule protein Tau, resulting in intraneuronal accumulations and formation of paired helical filaments and neurofibrillary tangles, a process that is greatly accelerated in Alzheimer’s disease (Hof and Morrison, 2004). Comparable normal and pathological changes occur very rarely or not at all in aged chimpanzees or other nonhuman primates (Rosen et al., 2008; Walker, 1999). It is therefore possible that differences in gene regulation that impact the levels of proteins that function in the protection against oxidative damage, including GSTO2, could be linked to aging-associated normal and disease-related differences between humans and chimpanzees.

A major challenge for the future will be to definitively establish which differences in the expression levels of splice variants are associated with functional and phenotypic attributes of species, including humans and chimpanzees. This challenge equally applies to the growing list of differences that have been detected recently at the levels of gene structure and composition.
(Bustamante et al., 2005; Cheng et al., 2005; Consortium, 2005; Newman et al., 2005; Perry et al., 2006), as well as at other levels of gene regulation (Pollard et al., 2006; Prabhakar et al., 2006). Another important question in relation to all of these studies is the degree to which evolutionary changes in gene structure or expression are required to establish major differences in morphological and phenotypic characteristics. While the accumulation of small-effect changes at multiple loci probably underlie many species-specific differences, differences in the expression of individual genes that are integrally involved in developmental processes can also result in the evolution of major morphological changes (McGregor et al., 2007). Likewise, the differences I detected in the levels of splicing involving 6%–8% of orthologous cassette alternative exons, which in general are highly consistent between individuals from each species, could have a significant impact on morphological and other phenotypic differences between humans and chimpanzees. The results of the present study provide the basis for future investigations directed at elucidating the functional and phenotypic consequences of AS differences between humans and chimpanzees.
Chapter 3

Regulation of vertebrate nervous system alternative splicing and development by nSR100

This chapter is derived from the following published article: J.A. Calarco, S. Superina, D. O’Hanlon, M. Gabut, B. Raj, Q. Pan, U. Skalska, L. Clarke, D. Gelinas, D. van der Kooy, M. Zhen, B. Ciruna, and B.J. Blencowe. Cell (2009) 138(5): 898-910. Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. U. Skalska performed the genome-wide search of RS domain protein genes and expression profiling data analysis described in Figure 3.1. B.J. Blencowe performed the multiple species alignment and made Figure 3.2. I performed all experiments and data analysis described in Figures 3.1-3.12. Some experiments in figures 3.1-3.12 were performed with assistance from D. O’Hanlon. M. Gabut performed the ES cell experiments described in Figure 3.3E, and I performed the adult stem cell experiments in Figure 3.3E with assistance from L. Clarke. Experiments from Figure 3.13 were performed by S. Superina and B. Ciruna, with the exception of experiments which I performed in Figure 3.13N. Additional figures can be found in Appendix 2, and supporting data files in the accompanying CD.
3.1 Abstract

Alternative splicing is a key process underlying the evolution of increased proteomic and functional complexity and is especially prevalent in the mammalian nervous system. However, the factors and mechanisms governing nervous system-specific AS are not well understood. Through a genome-wide computational and expression profiling strategy, we have identified a tissue- and vertebrate-restricted Ser/Arg (SR)-repeat splicing factor, the neural-specific SR-related protein of 100 kDa (nSR100). We show that nSR100 regulates an extensive network of brain-specific alternative exons enriched in genes that function in neural cell differentiation. nSR100 acts by increasing the levels of the neural/brain-enriched polypyrimidine tract binding protein and by interacting with its target transcripts. Disruption of nSR100 prevents neural cell differentiation in cell culture and in the developing zebrafish. Our results thus reveal a critical neural-specific AS regulator, the evolution of which has contributed to increased complexity in the vertebrate nervous system.
3.2 Introduction

Alternative splicing has played a major role in the evolutionary expansion of proteomic and functional complexity underlying many cellular processes and is especially prevalent in the vertebrate nervous system, with emerging key roles in synaptogenesis, neurite outgrowth and axon guidance, ion channel activity and long-term potentiation (Li et al., 2007; Ule and Darnell, 2006). However, mechanisms that control neural-specific AS and which underlie the evolution of increased nervous system complexity are poorly understood.

Regulated AS requires the interplay of cis- and trans-acting factors that repress or activate splice site selection (Blencowe, 2006; Wang and Burge, 2008). RNA binding domain (RBD)-containing proteins belonging to the heterogeneous nuclear ribonucleoprotein (hnRNP) family and a superfamily of alternating Arg/Ser (RS) domain-containing proteins, referred to as “SR family” and “SR-related” proteins, act widely to control AS. Members of these protein families can either repress or promote the formation of active splicing complexes (spliceosomes), often depending on the location of cognate binding sites within exon or intron sequences (Lin and Fu, 2007; Martinez-Contreras et al., 2007).

SR family proteins contain one or two N-terminal RNA recognition motifs and a C-terminal RS domain, whereas SR-related proteins, which include factors that are pivotal in the control of sex determination in Drosophila, contain RS domains but may or may not contain RBDs (Lin and Fu, 2007). Phosphorylated RS domains are thought to function in the formation of protein-protein and protein-RNA interactions required for spliceosome assembly (Shen et al., 2004; Wu and Maniatis, 1993). Approximately 40 mammalian RS domain proteins have been implicated in splicing (Lin and Fu, 2007). Although ubiquitously expressed, it has been proposed that variable levels of these proteins may contribute to cell and tissue-dependent AS patterns (Hanamura et al., 1998; Zahler et al., 1993).

To date, only RBD proteins that lack RS domains have been shown to possess tissue-restricted expression patterns and several such proteins play important roles in cell and tissue-specific AS (Li et al., 2007). These include the neuronal-specific Nova-1/2 and HuB proteins, the neuronal and muscle-expressed Fox-1/2, CELF/CUGBP and MBNL family proteins, and the neural, myoblast and testis-expressed nPTB protein (also referred to as brPTB and PTBP2).
nPTB is a paralog of the widely expressed PTB (also hnRNPI/PTBP1) splicing repressor protein (Ashiya and Grabowski, 1997; Markovtsov et al., 2000; Polydorides et al., 2000). A switch from PTB to nPTB expression during neuronal differentiation has been implicated in regulating neuronal AS (Boutz et al., 2007b; Makeyev et al., 2007; Spellman et al., 2007). Previous evidence has suggested that this switch results in the formation of “less repressive” nPTB-bound splicing complexes which are responsive to positive-acting factors in neurons (Markovtsov et al., 2000). However, the mechanisms controlling neuronal-specific exon inclusion patterns in conjunction with nPTB have remained unclear.

Several tissue-restricted AS factors are involved in controlling subsets of exons that are enriched in functionally-related genes (Boutz et al., 2007b; Kalsotra et al., 2008; Ule et al., 2005b; Yeo et al., 2009; Zhang et al., 2008a). For example, through binding to clusters of YCAY motifs distributed within specific exon and intron zones, Nova2 can positively or negatively regulate ~7% of neural-specific alternative exons in genes that function in the synapse and in axon guidance (Licatalosi et al., 2008; Ule et al., 2005b). How AS specificity is established for the vast majority of other neural-specific exons is not understood, however. Moreover, the specific biological processes controlled by the majority of neural-specific exons are also not well understood.

I have identified a tissue- and vertebrate- restricted RS domain protein, referred to as the neural-specific SR-related protein of 100 kDa (nSR100). Knockdown of nSR100 disrupts the inclusion of a large set of nervous system-specific alternative exons that are significantly enriched in genes with critical functions in neural cell differentiation. Consistent with this molecular programming function, nSR100 is required for neural cell differentiation in vivo. nSR100 forges neural-specificity in AS by activating nPTB expression and, in conjunction with nPTB, by binding directly to its regulated target transcripts. nSR100 thus acts in a multi-faceted manner in the tissue-specific regulation of a network of exons associated with neural cell differentiation and the evolution of vertebrate nervous system complexity.
3.3 Materials and Methods

3.3.1 Computational identification and characterization of RS domain proteins

Tissue restricted mammalian RS domain genes were identified computationally essentially as previously described (Boucher et al., 2001). Human and mouse cDNA sequences were compiled from the RefSeq, FANTOM, and MGC databases. (SR)$_{30}$ was used as a query to BLAST search all translated cDNA sequences. BLASTALL (version 2.2.9) was used with all default options, except low complexity filtering of the query sequence was not applied. This search resulted in 289 human and 424 mouse hits. These hits were searched for an exact match to at least one SRSR or RSRS tetrapeptide sequence (Boucher et al., 2001). Following this search there were a total of 248 human and 332 mouse protein sequences. These mouse sequences were searched against one another to identify redundant hits. As before, this search was performed with all default options for BLASTALL, but with filtering of the query sequences turned off. Protein sequences were flagged as duplicates when all of the following applied:

1. (gap length + identity count) / (alignment length) ≥ 0.99

2. percent identity ≥ 90

3. alignment length > 0.9*(query length) OR alignment length > 0.9*(subject length)

Applying these criteria to screen redundant entries allowed for possible sequence errors, gaps and/or polymorphisms in the alignment. Groups of sequences flagged as duplicates were combined into sets and a representative sequence was chosen for each; in each case the representative sequence was the longest protein, and if all protein sequences were of the same length the longest corresponding cDNA sequence was chosen. This resulted in the selection of 111 mouse protein sequences and 107 human proteins sequences. Inspection of this set revealed that the only known mouse SR-family protein missing was SRp30c, and this was added to the database manually, resulting in a total of 112 unique mouse proteins. Refer to Supporting data file SD 3.1 on CD for a full list and associated annotation.
### 3.3.2 Expression profiling of computationally-defined mouse RS domain genes

The computationally-defined mouse RS domain cDNA sequences were BLAST searched (default filtering of query sequences turned off) against oligonucleotide probe sequences represented on custom Agilent microarrays used to perform an analysis of mouse gene expression (Zhang et al., 2004). A probe was flagged as matching to a mouse cDNA sequence if: 

\[
\frac{\text{identity count}}{\text{probe length}} \geq 0.90
\]

69 mouse cDNA sequences matched 84 microarray probes by this criterion. 64 of the cDNAs matched a probe perfectly (identity count = probe length); 2 displayed one mismatch, and 3 displayed 2-4 mismatches. 12 mouse RS domain cDNA sequences matched duplicate transcripts represented by probes on the microarray, and 1 probe matched two different RS domain cDNA sequences. These sequences were found to map to different chromosome locations and therefore are from different genes. The mRNA expression level for each gene in each tissue is indicated as the expression level following subtraction of the median expression value of the gene in all profiled tissues. Negative expression values were set to zero. Median-subtracted arcsinh transformed values are shown in the clustergram in Figures 3.1 and A2.1 (Appendix 2).

### 3.3.3 Tissue samples, cell lines and cell culture

Mouse tissue samples were collected from ~5 mice and were dissected and flash frozen in liquid nitrogen immediately after sacrificing. Zebrafish heads were dissociated from ~50 WT and MOspl morphant animals and immediately flash frozen. HeLa, 293-T, Weri-Rb1, Neuro2a, NIH-3T3 and C2C12 cell lines were obtained from ATCC. Phoenix cells were obtained from Orbigen. Cell lines were grown in Dulbecco’s Modified Eagles Medium (DMEM) supplemented with 10% FBS and antibiotics with the following exceptions: Weri-Rb1 cells were grown in RPMI medium supplemented with 10% FBS and antibiotics; Neuro2a cells were grown in DMEM supplemented with 10% FBS, Sodium pyruvate, MEM non-essential amino acids, and antibiotics. CGR8 ES cells (obtained from Karl-Heinz Krause, Université de Genève) were cultured as described previously (Suter et al., 2006). Neurospheres were induced in vitro by seeding CGR8 cells as single cell suspensions at a density of 50 cell per microliter in serum free media for 10 days, as previously described (Tropepe et al., 2001). ES cells were maintained and differentiated into neurospheres and neurons and glia as described previously (Tropepe et al., 2001). Neuro2a cells were differentiated by culturing the cells in 2% FBS and 20 mM all-trans...
retinoic acid for 24-48 hours. Adult neural stem and progenitor cells were isolated and differentiated as described previously (Morshead et al., 1994; Reynolds and Weiss, 1992). All cell lines were cultured at 37°C with 5% CO₂.

### 3.3.4 Lentivirus production, infection, and generation of stable cell lines

Stable cell lines nSR100 and GFP-targeting shRNAs from lentiviral vectors were produced as previously described (Moffat et al., 2006). Lentiviruses used to generate cell lines stably overexpressing nSR100 and SF2/ASF were produced using procedures previously described (Karni et al., 2007). Cell lines co-expressing nSR100 or SF2/ASF and shRNAs were generated by first generating over-expression lines, followed by infection with shRNA lentiviruses and performing double selection with puromycin and hygromycin.

### 3.3.5 Plasmids

The full-length nSR100 ORF was amplified by RT-PCR from human brain polyA+ RNA using the following primers: 5´-CGGAAGCTTATGGCGAGCGTTCAGCAA-3´ and 5´-CGGGAATTCTTAGCGCTCGTGCTGGA-3´. The insert was then cloned into pCMV-Myc (Rosonina et al., 2005) using available HindIII and EcoRI sites to generate pCMV-Myc-nSR100. The full-length mouse nSR100 ORF was amplified by PCR from a cDNA clone obtained from the Mammalian Gene Collection with the following primers: 5´-CGGTCTAGAGCTAGCCTGCAGCAGGGC-3´ and 5´-CGGTGATCATTAGCGC-CTCGTGCTGGCATA-3´. This insert was cloned into the vector pCG-T7 (a gift from Javier Caceres, MRC Edinburgh) using available XbaI and BamHI restriction sites to create pCG-T7-nSR100. The coding sequence spanning the N-terminal T7-epitope and nSR100 ORF were then amplified from this plasmid using the primers 5´-CGGGAATTCACCAGCACCATGGCATGACAGGAGG-3´ and 5´-CGGGTCGACTTAGCGCCTCGTGCTGGCATA-3´. This amplified product was cloned into the pWZL-Hygro vector (a gift of Adrian Krainer, CSHL) using EcoRI and Sall restriction sites to generate pWZL-Hygro-nSR100. The pWZL Hygro-T7-SF2/ASF plasmid was a gift of Adrian Krainer. The plasmid containing the shRNA targeting mouse nSR100 (pLKO.1-sh-nSR100) was constructed by annealing the oligonucleotides 5´-CCGGGCTCACGATATATCTCTCTAAACTCGAGTTTAGGAGATATAC-TGTGAGCTTTTG-3´ and 5´-AATTCAAAAAGCTCACGATATATCTCTCTAAACTCGAGTTTAGGAGATATAC-TGTGAGCTTTTG-3´ and cloning the annealed DNA into the pLKO.1
vector (a gift from Jason Moffat, University of Toronto). A control plasmid containing an shRNA targeting the coding sequence of GFP was also a kind gift from Jason Moffat. The Daam1 minigene reporter was generated by amplifying a region containing the alternative exon and flanking intronic sequences from a BAC clone using the primers 5’-CGGCTCGAGCAAATGTCAACA-CACCCACAC-3’ and 5’-CGGGCGGCGCCGCCC- CGAATCTCAGCAAAACA-3’. This product was inserted into the PET01 exon trap vector (Mobitec) using XhoI and NotI restriction sites to create PET01-Daam1. This vector placed the Daam1 exon and surrounding intronic sequence in between two heterologous exons. All other truncations and mutant derivatives were created from this construct by standard molecular cloning approaches. Primers for these constructs are available upon request. Full length ZnSR100 coding sequence was cloned by PCR using the following two ESTs as substrates: 5’ sequence was PCR amplified from dbEST Id: 23333826, and 3’ sequence from dbEST Id: 39404816. Primers used are as follows Primers used are as follows (sequence complementary to ZnSR100 in bold; Invitrogen Gateway attB cloning elements underlined): 5’ for: GGGGACAAGTTTTGTACAAAAAAGCAGGCTTCGCCACCATGGCTGGCGTACTTTTGGGAGAAAAGC, 5’ rev: CCTCTGCGATTCTTTTCTTTTCTTTTCTTGTTTATGGC ACTTGGCCGGGGG, 3’ for: CCCCTCGGCCAGTGCTAATAACGAAAGAA- AAAGAAGAAATCGCAGAGG, 3’ rev: GGGGACCACCTTTGTACAAGAAGCTGGGTGGGCGTCGGGGCTGCTGTAGCTTCTGC. 5’ rev- and 3’ for-primers are complementary, and the amplified products contained significant sequence overlap. To assemble the full length ZnSR100 clone, 5’ and 3’ PCR products were allowed to hybridize and extend in a limited PCR reaction, and the full length product was further amplified using 5’ for and 3’ rev primers. The PCR product was cloned into pDONR221 using the Invitrogen BP Clonase system to generate a ZnSR100 middle entry vector. Endogenous and N-terminal EGFP tagged ZnSR100 expression vectors were constructed as described in (Kwan et al., 2007). To generate a plasmid for producing recombinant nSR100 in insect cells, the full length human nSR100 ORF was PCR amplified from pCMV-Myc-nSR100 using the following primers: 5’-CGGAGATCATGTGCGGCTGGCGTACTTTTCTGCAGCAAA–3’ and 5’–CGGAAGCTTTTAGCACCCTCGATTGCGTCT–3’. The DNA fragment was then cloned into the BamH1/HindIII sites of the recombinant transfer vector pBlueBacHisA (Invitrogen) to generate the construct pBlueBac-His-nSR100. A plasmid for producing recombinant GST-tagged PTB was a gift from Mariano Garcia-Blanco (Duke University). A plasmid to express recombinant GST-tagged nPTB was
generated as follows: the human nPTB ORF was amplified by PCR from the vector pQE30-nPTB (a gift from Chris Smith, University of Cambridge) using the following primers: 5´-CCGGAATTTCATGACGGAATCG-TCACTGAAGTT-3´ and 5´-CGGGAATTCTTTAATTGTGACTTGGAGAAG-3´. This insert was then cloned into the pGEX-2T plasmid (Amersham) using an EcoRI restriction site.

3.3.6 Genome-wide analyses of nSR100 alternative splicing regulation

We developed a new mouse AS microarray in a similar manner as described previously (Pan et al., 2004) capable of monitoring 6702 cassette splicing events. After GenASAP %ex predictions were obtained for AS events in each cell line (Neuro2a expressing control or nSR100 targeting shRNAs) and tissue (brain, liver, heart, skeletal muscle, kidney, spleen) were analyzed, we focused on identifying physiologically relevant targets regulated by nSR100. All oligonucleotide probes on the microarray were present at least in duplicate, allowing us to obtain two GenASAP predictions for a given AS event. Candidate events regulated by nSR100 were selected according to the following criteria:

Expression levels of each AS event (as determined by the average of the intensities of flanking constitutive exon body probes for each AS event) had to be above background levels in both control and nSR100 kd Neuro2a cell lines. Background levels were determined by taking the 90th percentile intensity of a set of negative control probes on the microarray.

We required a |%ex difference| > 15% between control and nSR100 shRNA expressing Neuro2a cell lines in one pair of GenASAP predictions, and a |%ex difference| > 10% in the second pair of GenASAP predictions for a given AS event.

We also required that the events that meet the criteria in 1 and 2, were also predicted to display differential splicing between brain and non-neural tissues (brain-regulated) or between brain, heart, skeletal muscle and non-neural tissues (brain and muscle regulated) according to any of the following criteria.

|%ex difference| > 10% between brain and all other tissues (brain-specific)

|%ex difference| > 10% when brain, heart, skeletal muscle were compared to liver, spleen, and kidney (brain and muscle specific)
expression above background levels (as determined in 1) only in brain and in none of the other tissues surveyed.

858 events met the criteria in this latter step. The events that met all of these criteria above (criteria 1, 2, and 3) represented the final list of 115 AS events predicted to be regulated by nSR100 in our microarray analysis. The estimate of 11% for alternative exons displaying differential regulation in brain that are also regulated by nSR100 was obtained as follows: Based on our RT-PCR validation rate for nSR100-dependent brain-included splicing events (21/26, or 81%), we reasoned that nSR100 regulates 0.81 x (81 nSR100-regulated events/858 brain-regulated events), or ~8% of brain regulated events. Accounting for our false negative rate of ~6% (see main text) gives an upper bound estimate of ~14%. Accordingly, we have used a number between these two values of ~11% in the results section.

For the GO enrichment analysis, we looked for overrepresented GO terms associated with genes containing brain-specific alternative exons that undergo increased skipping upon knockdown of nSR100. These genes were compared with all genes containing AS events surveyed by our microarray and expressed above background in Neuro2a cells. The GO enrichment analysis was performed using the software tool GOstat (Beissbarth and Speed, 2004) searching the MGI GO annotation database under default parameters. Over-represented GO terms listed in Table 3.1 have p values < 0.1 after FDR correction.

To identify nSR100-regulated exons that overlap with transcript regions encoding known protein domains, we applied an approach similar to that described previously (Pan et al., 2005). To assess the extent to which nSR100-regulated alternative exons are conserved between human and mouse and/or alternatively spliced in both human and mouse transcripts, relative to a set of control alternative exons. We searched for mRNA/EST evidence supporting each of these possibilities using stringent criteria as previously described (Pan et al., 2005; Pan et al., 2004). Data from these analyses is provided in Supporting data file SD 3.2 (CD).

In order to search for overrepresented motifs associated with nSR100 regulated alternative splicing events, we used a modified version of the SeedSearcher algorithm as described previously (Barash et al., 2001; Fagnani et al., 2007a).
3.3.7 RT-PCR assays

Non-radioactive RT-PCR reactions were performed using the One Step RT-PCR kit (Qiagen), essentially as per the manufacturers instructions. Reactions contained 30 ng of total RNA or 0.2 ng of polyA+ RNA. Radiolabeled reactions contained 0.3 mCi of $\alpha$-32P-dCTP per 10 uL reaction. The number of amplification cycles was 22 for beta-actin, gapdh, and minigene reporters, and 27-30 for all other transcripts analyzed. Reaction products were separated on 2% agarose or 6% denaturing polyacrylamide gels. Quantification of isoform abundance was performed as described previously (Calarco et al., 2007b).

3.3.8 Immunofluorescence, antibodies, and western blotting

For co-immunofluorescence experiments, mouse neural progenitor cells were differentiated and immunostained as described previously (Tropepe et al., 2001). Anti-βIII-tubulin (Sigma), anti-GFAP (Sigma), and affinity-purified anti-nSR100 antibodies were used at 1:500, 1:250, and 1:2000 dilutions, respectively.

For western blotting, anti-α-tubulin and anti-c-myc epitope antibodies were purchased from Sigma-Aldrich. Anti-T7 tag antibodies were obtained from Novagen. Anti-GFP antibodies were obtained from Roche. Pan-PTB (recognizing all PTB paralogs) and anti-nPTB antibodies were a gift from Christopher Smith (University of Cambridge). Protein lysates were prepared by addition of Laemmli sample buffer to whole cell splicing extracts generated from all cell lines used in this study. 40-150 ug of lysate was separated on 10% SDS-polyacrylamide gels and transferred to PVDF membrane (Pall Corporation). Secondary antibodies (GE Healthcare Life Sciences) and chemiluminescence reagents (Perkin Elmer) were used as per manufacturers recommendations.

3.3.9 Expression and purification of recombinant PTB, nPTB, SF2/ASF and nSR100

Plasmids for expressing GST-tagged PTB and nPTB were transformed into XL10 gold ultracompetent cells (Stratagene). Transformed bacteria were grown at 37°C until reaching an $OD_{600}$ of 0.6. The bacteria were then induced with 1mM IPTG for 2-3 hours followed by lysis and purification of the GST-tagged proteins with glutathione sepharose 4B beads (Amersham) as recommended by the manufacturer.
A plasmid for expression of T7-epitope tagged SF2/ASF in mammalian cells was provided by Javier Caceres (MRC, University of Edinburgh). This plasmid was transfected in 293T cells and purification of recombinant epitope tagged SF2/ASF was purified as described previously (Cazalla et al., 2005).

To produce recombinant baculovirus used to express nSR100, we followed the protocols outlined in the Bac-N-Blue baculovirus transfection and expression guide. Briefly, Sf9 (Spodoptera frugiperda) cells were transfected with pBlueBac-His-nSR100 and Bac-N-Blue DNA (Invitrogen) to produce recombinant virus. These recombinant stocks were plaque purified and blue, occ- plaque plugs were picked and used to produce P1 virus stock by re-infesting fresh Sf9 cells. A portion of the generated P1 stocks were used to purify viral DNA which was then used as a PCR template to determine which stocks contained recombinant virus. Pure recombinant virus was then used to generate subsequent high titre viral stocks as described as per manufacturer’s instructions. Recombinant virus was used to infect suspension cultures of Hi5 cells (Trichoplusia ni, Invitrogen) for nSR100 protein production. Log phase suspension adapted cultures of Hi5 cells were seeded at 0.8 x 10^6 cells/mL 24 h prior to infection with recombinant virus. Infected cultures were grown in serum-free media (HyQ SFX Insect media, Hyclone) at 27°C until maximal protein production. Cells were harvested by centrifugation, washed with ice-cold PBS and the pellets snap frozen at -80°C until used for protein purification. Sf9 cells were maintained at 27°C in Graces Insect Media (Invitrogen) plus 10% FBS and supplements. Hi5 cells were adapted to suspension culture and maintained in suspension at 27°C in serum-free media (HyQ-SFX Insect media, Hyclone).

To purify nSR100, ~10 g of pellets were resuspended in 20 mL lysis buffer (50 mM Tris-HCl pH 8.3, 400 mM KCl, 1% NP-40, 5 mM β-glycerophosphate, 5 mM KF, 0.04% (v/v) β-mercaptoethanol, 1.5 mM PMSF, protease inhibitors), sonicated with 3 x 15 s bursts followed by a 10 minute incubation on ice. The sample was centrifuged at 20 000 x g for 20 minutes, supernatant removed and pellet containing insoluble nSR100 kept on ice for further purification. The pellet was resuspended in 7.5 mL of 3M urea lysis buffer (3M urea, 10 mM HEPES-KOH pH 7.6, 100 mM KCl, 5 mM β-glycerophosphate, 5 mM KF, 0.04% (v/v) β-mercaptoethanol, 1.5 mM PMSF, protease inhibitors). The resuspended pellet was sonicated with 3 x 10 s bursts
followed by a 20 minute incubation on ice. The sample was then centrifuged at 20 000 x g for 20 minutes and supernatant collected. This supernatant containing His-tagged nSR100 was passed through a column containing Ni-NTA agarose beads (Novagen) three times, followed by the following series of washes:

1. 10 column volumes of 3M urea/0.5M KCl buffer
2. 10 column volumes of 3M urea/1.5M KCl buffer
3. 10 column volumes of 3M urea/1M KCl buffer
4. 10 column volumes of 3M urea/1.5M KCl buffer
5. 10 column volumes of 3M urea/0.5M KCl buffer
6. 10 column volumes of 3M urea/0.1M KCl buffer
7. 10 column volumes of 3M urea/0.1M KCl buffer
8. 10 column volumes of Buffer E
9. 10 column volumes of Buffer E

   The bound protein was eluted from the beads with elution buffer (20 mM HEPES-KOH pH 7.6, 100 mM KCl, 5% glycerol, 1 mM DTT, 0.2 mM EDTA, 250 mM imidazole). 300 uL fractions were collected and fractions containing highest concentrations of protein (determined by Bradford assay) were pooled, glycerol was added to 10%, and samples were snap frozen and stored at –80°C.

### 3.3.10 RNA gel mobility shift assays

Binding reactions, used for gel shift assays contained 12 mM HEPES-KCl pH 7.9, 60 mM KCl, 0.12 mM EDTA, 6% glycerol, 1 mM DTT, and 100 ng/uL of tRNA. Purified recombinant His-tagged nSR100, GST-tagged PTB and GST-tagged nPTB were pre-incubated under binding conditions for 8 minutes at 30°C. Approximately 10 fmol of radiolabeled RNA was then added and each sample was incubated for 15 minutes at 30°C and then placed on ice for an additional 5 minutes. Gel loading dye was added and the samples were run for 5 hours on a
4% native polyacrylamide gel in 0.5X TBE. Gels were dried and subsequently exposed on X-ray film (Kodak).

3.3.11 Preparation of whole cell in vitro splicing extracts and in vitro splicing

Harvested cells were washed twice with ice cold PBS, resuspended in four packed cell volumes of ice cold high salt Buffer E (20 mM HEPES-KCl pH 7.9, 600 mM KCl, 0.2 mM EDTA, 10% glycerol, 1 mM DTT) and sonicated with a Branson sonifier at power setting 3.5, duty cycle 35% with three seven burst intervals and 30 s incubation periods on ice between intervals. The cell lysate was then centrifuged at 20 000 x g for 20 min at 4°C and the supernatant was dialyzed against Buffer E. After dialysis, the samples were centrifuged at full speed in a microfuge for 10 minutes and the supernatant was retained as functional whole cell extract that was snap-frozen in liquid nitrogen and stored at –80°C.

For in vitro splicing assays, a splicing substrate recapitulating Daam1 neural-specific exon inclusion was constructed. The region spanning Daam1 alternative exon 16 and its flanking 280 and 145 nucleotides of upstream and downstream intron sequence, respectively, was inserted between the 5’ and 3’ splice sites of the adenovirus derived pMINX splicing substrate at position 44 in the intron (Zillmann et al., 1988) by using an overlapping PCR strategy. A T7 RNA polymerase promoter sequence was incorporated into the final PCR product, which was used as a template for in vitro transcription using the T7 Megascript IVT kit (Ambion).

In vitro splicing assays performed in 20ul contained 1.5 mM ATP, 5 mM creatine phosphate, 5 mM DTT, 3 mM MgCl2, 2.6% PVA, RiboLock RNase inhibitor (Fermentas), 10 ng of splicing substrate, 40-60ug of Weri or HeLa splicing extract and up to 12 uL Buffer E. Reactions were incubated at 30°C for one hour prior to RNA isolation. Spliced products were amplified by RT-PCR assays using primers specific for MINX exons. RT-PCR assays for monitoring splicing activity employed 30 cycles of amplification and 40% of the RNA isolated from each reaction. RT-PCR products were resolved on a 2% ethidium bromide-stained agarose gel stained and visualized using the Gel Doc imaging system (BioRad).
3.3.12 In vitro and in vivo crosslinking and immunoprecipitation (CLIP)

For *in vitro* crosslinking and immunoprecipitation, Weri cell extract was pre-incubated under splicing conditions in the presence of 60 ng/μL tRNA for 8 minutes at 30°C. Radiolabeled RNA substrate was added followed by incubation for 15 minutes at 30°C. Samples were irradiated with UV light for 10 minutes on ice in a UV Stratalinker 1800 (Stratagene) at a distance of 1.5 cm from the light source, followed by digestion with 0.1 μg of RNaseA for 15 minutes at 37°C. Samples were separated by SDS-PAGE and analyzed by autoradiography. For *in vivo* CLIP, a method similar to that described in (Ule et al., 2005a) was followed with a few modifications. Weri cells were harvested, washed two times with ice cold 1X PBS, and ~0.5 mL of pelleted cells were resuspended in 2 mL of 1X PBS prior to crosslinking. The resulting suspension was placed in a pre-wet and pre-chilled 10 cm plate placed on ice. UV crosslinking was performed by irradiating the cells in a UV Stratalinker 1800 at an energy setting of 400 J/cm², with cells placed 4.5 cm away from the light source. Samples were then pelleted and resuspended in 4 packed cell volumes of high salt buffer E, sonicated as described above for splicing extract preparation, and then centrifuged at 20 000 x g for 20 minutes at 4°C. This supernatant after centrifugation was then snap frozen in liquid nitrogen and stored for subsequent use in immunoprecipitation. Control, non-UV irradiated extracts were made by resuspending the same amount of cells in standard Buffer E.

Extracts used for immunoprecipitation were first pre-cleared by incubation with protein A sepharose beads coupled with control rabbit anti-mouse IgG antibodies for 1 hour at 4°C with rotation. Supernatants after this pre-clearing step were then incubated with Protein A sepharose beads (GE Healthcare Life Sciences) alone for an additional 30 minutes at 4°C with rotation to remove residual unbound IgG antibodies. This pre-cleared extract was then transferred to Protein A sepharose beads coupled with either rabbit anti-mouse or rabbit anti-nSR100 antibodies and incubated for 3 hours at 4°C with rotation. After binding, the pelleted beads were washed at 4°C as follows:

- 1x with 1 mL of IP wash buffer 150 (10 mM Tris-HCl pH 7.5, 150 mM KCl, 0.1% NP-40, 2.5 mM MgCl₂, 5 mM KF, 5 mM β-glycerophosphate) with 5 minutes rotation.

- 2x with 1 mL buffer A (1x PBS, 0.1% SDS, 0.5% deoxycholate, 0.5% NP-40).
-2x with 1 mL buffer B (5x PBS, 0.1% SDS, 0.5% deoxycholate, 0.5% NP-40).

-2x with 1 mL buffer C (50 mM Tris-HCl pH 7.4, 10 mM MgCl2, 0.5% NP-40)

The beads were then resuspended in 100 uL proteinase K buffer (10 mM Tris-HCl pH 7.6, 100 mM NaCl, 1 mM EDTA, 0.5% SDS), proteinase K was added to the mixture at a final concentration of 300 ng/uL, and samples were incubated at 55°C for 45 minutes. 400 uL of Tri reagent (Sigma) was added to the samples and RNA was isolated and resuspended in DEPC-treated water. The resulting samples were then digested with DNaseI (Ambion) for one hour at 37°C to remove any contaminating genomic DNA, and extracted once more in Tri reagent. The resulting immunoprecipitated RNA was assayed by RT-PCR using the OneStep RT-PCR kit (Qiagen) with primers designed to specifically amplify cDNAs corresponding to pre-mRNA regions spanning the alternative exon and ~200 nucleotides of upstream intronic sequence.

3.3.13 Zebrafish embryo maintenance, imaging, in situ hybridization and injections

Zebrafish embryos were staged as described previously (Kimmel et al., 1995). Analyses were performed in a TLxAB background. Embryos >24hpf were treated with 0.003% phenylthiourea (PTU) to reduce pigmentation. RNA in situ hybridizations were performed using nitroblue-tetrazolium chloride and 5-bromo-4-chloro-indolyl-phosphate detection (Roche). Embryo staging and analyses were performed using the TLxAB background. One cell stage embryos were injected with 6ng of antisense morpholino oligonucleotide and 25pg of capped ZnSR100 mRNA synthesized using the mMESSAGE system (Ambion) as per the manufacturers recommendation. Embryos were anaesthetized using 0.01% MESAB (Sigma) and mounted in 3% methylcellulose for routine imaging. Confocal analyses of Tg(sensory:GFP) expression were performed as described (Sagasti et al, 2005). Embryos were mounted in agarose in a sealed chamber and imaged on a Zeiss LSM710 confocal microscope using a 40x objective. Approximately 60-80 confocal sections, each ~1um thick, were compiled for 3D projections.

3.3.14 Generation of probes for whole-mount RNA in situ hybridization

ZnSR100 antisense RNA was transcribed from a NcoI-linearized pCS2 plasmid using T3 polymerase and islet-1 antisense RNAs were transcribed from a BamHI linearized pBS KS(+) plasmid using T3 polymerase.
3.3.15 Morpholino injections, RNA injections, and oligonucleotide sequences

Embryos were injected at the one-cell stage with 6ng of p53-MO and 3ng MOatg or MOspl, as described previously (Nasevicius and Ekker, 2000). MOs were resuspended in water, and concentrations were determined as per the manufacturer’s instructions (Gene Tools). Plasmid containing ZnSR100 cDNA was linearized, and sense strand-capped mRNA was synthesized using the mMESSAGE mMACHINE system (Ambion).

The following morpholino sequences were used in this study:

p53 MO: 5´-GCGCCATTGCTTTGCAAGAATTG-3´

MOspl: 5´- TCAATCACTACCTATGTCGCTTCCT-3´

MOatg: 5´-TTCTCCCCAAAAGTACGCCAGCCATG-3´

3.3.16 siRNA and minigene reporter transfections

Neuro2a cells were transfected with SMART-pool siRNAs (Dharmacon) using Dharmafect (Dharmacon), as recommended by the manufacturer. Daam1 minigene reporters and expression plasmids were transfected using Lipofectamine 2000 (Invitrogen), as recommended by the manufacturer. Total RNA was collected 48 hours post-transfection and isolated using RNeasy kits (Qiagen).
3.4 Results

3.4.1 A genome-wide screen identifies a vertebrate and neural-specific SR-related protein

To identify mammalian SR-related proteins with the potential to function as cell type- and/or tissue-specific splicing regulators, we used a computational procedure for the genome-wide surveying of RS domain protein-encoding genes (Figure 3.1 A; refer to Supplementary Information). Applying this procedure to a set of non-redundant mouse cDNAs resulted in the identification of 112 known or putative new RS domain genes (Supporting data file SD 3.1, CD).

Manual annotation revealed that 36% (40/112) of the RS domain genes have a function associated with splicing, and 15% (17/112) have functions associated with RNA polymerase II-dependent transcription and mRNA 3′-end processing, which are coupled to and can influence splicing (Figure 3.1 A). Remarkably, 32% (36/112) of the identified genes have no annotated function. Given that more than one-third of the annotated genes have known functions in splicing, many of the previously uncharacterized genes likely also participate in splicing. Moreover, since 96% (108/112) of the predicted mouse proteins encoded by these genes have a closely related human ortholog with a computationally-defined RS domain, the majority of these proteins likely have conserved functions.

Using microarray profiling data we analyzed expression patterns of the computationally-mined RS domain genes in 50 diverse mouse cell and tissue types (Figures 3.1 B and A2.1, Appendix 2). Many of these genes appear to be widely but variably expressed, consistent with previous observations (Hanamura et al., 1998; Zahler et al., 1993). Interestingly, hierarchical clustering analysis reveals that two prominent subsets of genes display distinct spatiotemporal regulation. One subset displays elevated expression in whole embryo and embryonic head samples from days 9.5 to 14.5, and the other displays elevated expression in adult nervous system tissues (Figure 3.1 B, boxed region; Figure A2.1, Appendix 2).

I focused my attention on an RS domain gene (NM_026886; human NM_194286/KIAA1853) displaying increased expression in the developing embryo and highly restricted expression in adult nervous system and sensory organ tissues (Figure 3.1 B). The NM_026886 ORF is predicted to encode a 608 amino acid protein of 68 kDa that contains
prominent runs of alternating SR/RS repeats (Figure 3.1 C). While the protein encoded by the NM_026886 gene has not been functionally characterized, it has been detected as a tumour antigen in pediatric patients with medulloblastomas (Behrends et al., 2003).

In addition to its neural-restricted expression pattern, a feature of the NM_026886 ORF that distinguishes it from previously characterized RS domain proteins is that it is highly conserved only in vertebrate species (Figure 3.2 and data not shown). The NM_026886 ORF and its orthologs lack a canonical RBD. The only region of significant similarity with other protein sequences is a sequence spanning amino acids 490 to 521 that is identical to amino acids 573 to 604 in the human SRm300 splicing coactivator subunit-like gene (SRm300-like/SRRM2L). These results, together with additional experiments described below, indicate that the NM_026886 gene encodes a neural-specific SR-related protein of 100kDa, which I refer to below as “nSR100”.

Figure 3.1: Identification and Expression Characteristics of nSR100
(A) (Left) Computational strategy for the genome-wide survey of mouse RS domain protein genes. (Right) Functional associations of computationally defined mouse RS domain protein genes (refer to Supporting data file SD 3.1 (CD) for a full list and description). (B) Clustergram showing relative mRNA expression levels of RS domain protein genes (y axis) across 50 mouse cell and tissue types (x axis). Color scale indicates fold expression relative to the median value across the data set. The enlarged portion of the clustergram shows RS domain protein genes with elevated expression in neural tissues with NM_026886/Riken cDNA 1500001A10 (nSR100) indicated. Refer to Figure A2.1 (Appendix 2) for an enlarged view of the entire clustergram with all gene names indicated. (C) Translated sequence of mouse nSR100 (from NM_026886), with RS/SR dipeptides highlighted. The underlined region indicates amino acids (1–82) used to generate an anti-nSR100 polyclonal antibody. The boxed region is identical to a sequence in the SRm300 splicing coactivator subunit-like protein (SRRM2L). (D) RT-PCR analysis of mouse nSR100 mRNA expression across 20 diverse tissues and several brain subregions, relative to Gapdh mRNA expression in the same samples. (E) Western blot of nonneuronal and neuronal-derived cell line lysates probed with an anti-nSR100 antibody. Arrowheads indicate detection of nSR100 and the asterisk indicates a cross-reacting species.
Figure 3.2: Multiple alignment of nSR100 ORF sequences from different vertebrate species.

The multiple alignment was created and visualized using the Multiple sequence comparison by log-expectation (Muscle) tool (http://www.drive5.com/muscle/) and ORF sequences from the following NCBI Accessions: NP_919262 (H. sapiens), XP_854166 (C. familiaris), NP_081162 (M. musculus), XP_573405 (R. norvegicus), XP_415281 (G. gallus), XP_001336923.1 (D. rerio).
3.4.2 Distribution of nSR100

RT-PCR assays confirmed that nSR100 mRNA is strikingly enriched in multiple brain regions and sensory organ tissues (Figure 3.1 D). A rabbit polyclonal antibody raised against a GST-fusion protein containing amino acids 1 to 82 of human nSR100 recognizes a band that migrates at ~100 kDa in neuronal-derived cell lines (mouse Neuro2a and human Weri-RB1), but which is not expressed in non-neural mouse or human cell lines (C2C12, NIH-3T3, HEK293-T and HeLa) (Figure 3.1 E). The higher than expected mobility of nSR100 is typical of SR proteins, since RS domains are generally highly phosphorylated. Confirming the phosphorylation status of nSR100, myc-epitope-tagged nSR100 protein expressed in Neuro2a cells was detected using the monoclonal antibody mAb104, which specifically recognizes a phosphoepitope shared among SR proteins (data not shown; (Zahler et al., 1992).

To establish whether nSR100 is expressed in specific neural cell types, adult mouse neural progenitor cells were differentiated in vitro into neurons and glia, the latter comprising astrocyte and oligodendrocyte populations. These neural cell populations were co-immunostained with anti-nSR100 antibody and anti-βIII tubulin (neuronal marker), or anti-glial fibrillary acidic protein (GFAP, glial marker) antibodies. nSR100 is highly enriched in the nuclei of the neuronal but not glial cell populations, since prominent nSR100 nuclear staining (marked by coincident Hoechst staining) is detected only in the anti-βIII tubulin positive cells (Figure 3.3A). Together, these experiments indicate that after neural progenitor cells differentiate, nSR100 primarily functions in the nuclei of neurons.
Figure 3.3: nSR100 Is Required for Neural Cell Differentiation

(A) Confocal microscopy images of differentiated neural progenitor cells coimmunostained with anti-β-III tubulin (red, top panels), anti-glial fibrillary acidic protein (anti-GFAP; red, bottom panels), and anti-nSR100 (green, top and bottom panels) antibodies. Nuclei are stained using Hoechst dye (blue, top and bottom panels). All β-III tubulin-positive neurons show nSR100 staining. Scale bar represents 25 um. (B) Western blot probed with anti-nSR100 antibody showing lysates from Neuro2a cells stably expressing control shRNA (lane 1) or nSR100-targeting shRNA (lane 2). α-tubulin levels were used as a loading control. (C) Quantification of...
percentage of shRNA-expressing Neuro2a cells extending long neurites (measured as greater than three cell body lengths) under normal culture conditions and differentiation conditions. Values represent averages from three independent experiments ± one standard deviation. Figure 3.5 shows representative brightfield images of Neuro2a cells from each treatment. (D) RT-PCR analysis of nSR100 mRNA levels in ES, neural progenitor (NP), and differentiated neurons and glia (lanes 1, 2, and 3, respectively). Gapdh mRNA levels are shown for comparison. (E) Quantification of relative percentage of primary neurospheres formed by differentiating ESCs expressing control or nSR100-targeting shRNAs or secondary neurospheres from adult neural stem cells expressing control or nSR100-targeting shRNAs. Values represent the averages from four independent experiments ± one standard deviation.

3.4.3 nSR100 is required for neural cell differentiation

We next asked whether nSR100 functions in the differentiation of cultured Neuro2a cells, murine embryonic stem (ES) cells and adult neural stem cells. Neuro2a cells produce long extensions known as neurites in the presence of retinoic acid (RA). Embryonic stem cells can be differentiated to produce neural stem and progenitor cells (Tropepe et al., 2001), and neural stem cells isolated from adult mouse brains can be differentiated into all major cell types of the nervous system in culture (Morshead et al., 1994). These three systems allowed us to investigate the role of nSR100 during different aspects of neural differentiation in vitro.

In Neuro2a cells nSR100 is expressed prior to induced differentiation and does not significantly increase at 48 hours post-induction (data not shown). To test whether nSR100 is required for neurite extension, lentivirus-delivered short hairpin RNAs (shRNAs) were used to generate undifferentiated Neuro2a cells with constitutively reduced (by 70-75%) levels of nSR100, relative to undifferentiated cells expressing a control shRNA targeted to GFP (Figure 3.3 B). The reduced levels of nSR100 had no significant impact on cell viability (Figure 3.4 and data not shown), but dramatically impaired neurite extension (Figures 3.3 C and 3.5). Within 24 hours post-induction with RA, >35% of control shRNA-treated cells extended long neurites, whereas only ~5% of the nSR100 shRNA-expressing cells extended processes of comparable length (Figure 3.3 C; Figure 3.5). In contrast, knockdown of a widely expressed SR-related splicing factor, SRm160, resulted in reduced viability of undifferentiated Neuro2A cells (Figure 3.4). These findings suggest that nSR100 is required for specific steps during neural differentiation, rather than for a more general role associated with cell viability.

nSR100 mRNA levels increase when ES cells are induced to differentiate to form neural progenitor cells (Figure 3.3 D). Cultured ES cells were induced to proliferate and aggregate into
Figure 3.4: Knockdown of SRm160 but not nSR100 dramatically affects Neuro2a cell numbers.

A. Brightfield images showing populations of Neuro2a cells expressing control shRNAs (left), shRNAs targeting nSR100 (middle), or shRNAs targeting SRm160 (right) after four days of selection with puromycin. B. RT-PCR assays using primers specific for monitoring nSR100 or SRm160 mRNA expression were performed to assess the degree of knockdown in the shRNA expressing Neuro2a cell lines shown in A. gapdh mRNA levels were assessed as a loading control. C. A comparison of the number of cells present in cells with reduced nSR100 or SRm160 levels relative to the control shRNA expressing cell line after four and six days of puromycin selection.
Figure 3.5 Knockdown of nSR100 impairs neurite outgrowth in Neuro2a cells
Brightfield images of Neuro2a cells under normal culturing conditions (left panel) or 24 hours after growth in 2% FBS and 20 uM retinoic acid to induce neurite outgrowth (middle and right panels). Cells expressing control shRNA (middle panel) or nSR100-targeting shRNA (right panel) are shown for comparison.
spherical cell populations known as ‘neurospheres’, which contain neural progenitor cells. When stably expressing control or nSR100-targeting shRNAs prior to induced differentiation, no discernable differences in morphology or proliferation were observed (data not shown). However, when induced, nSR100 knockdown ES cells formed 90% fewer neurospheres than did the control shRNA-treated cells (Figure 3.3 E). Similar results were observed upon knockdown of nSR100 in induced adult-derived neural stem cells (Figure 3.3 E). Two independent shRNAs targeting nSR100 led to similar results, indicating that the effects are specific and not due to off-target effects (data not shown; see also below and Figure 3.3 D). These results therefore indicate that nSR100 plays a critical role in neural stem and progenitor cell formation and/or proliferation, in addition to a role in neurite extension.

3.4.4 nSR100 regulates a network of brain-specific alternative exons

I next investigated whether the effects of nSR100 knockdown on neural cell differentiation are associated with the deregulation of specific AS events. PolyA+ RNA from control and nSR100 shRNA-expressing, undifferentiated Neuro2a cells was profiled using an AS microarray (Figure 3.6 A; refer to Supplementary Information). Alternative splicing patterns in adult mouse whole brain and five non-neural tissues were profiled in parallel (Figure 3.6 B). Confidence-ranked percent exon exclusion estimates (%ex; the percentage of transcripts skipping a profiled alternative exon) were determined (Pan et al., 2004). In order to enrich for physiologically relevant events, we focused our analysis on exons displaying both nSR100-dependent regulation in Neuro2a cells and differential splicing between brain and non-neural tissues (Figure 3.6 B and data not shown).

115 high-confidence AS events met the above criterion and displayed %ex differences of at least 15% upon nSR100 knockdown (Supporting data file SD 3.2, CD). 70% (81/115) displayed increased exon skipping upon knockdown of nSR100, and RT-PCR experiments validated 81% (21/26) of these events (Figures 3.6 C and A2.2, Appendix 2; data not shown). Only one of seven tested events displaying increased exon inclusion upon nSR100 knockdown was validated, although the change in %ex was relatively modest (data not shown). Moreover, only 6% (3/52) of predicted non-nSR100 targets were found to change upon nSR100
Figure 3.6: nSR100 Regulates a Network of Neural-Specific Alternative Splicing Events

(A) Strategy used for the global analysis of nSR100-regulated AS. Black lines represent exon body and exon junction probes used to monitor AS levels on a custom microarray. (B) Heatmap displaying microarray %ex values for a subset of nSR100-regulated alternative exons. Columns show microarray %ex predictions from profiled adult mouse tissues and Neuro2a cells expressing control or nSR100-targeting shRNAs, and rows represent %ex predictions for individual nSR100-regulated exons undergoing increased skipping upon nSR100 knockdown. The AS events are sorted in descending order according to the magnitude of the %ex values in the control shRNA-expressing Neuro2a cells. All AS events displayed have detectable expression in whole brain and at least two other tissues; white indicates expression was below the threshold level used to derive %ex estimates. (C) RT-PCR assays using primers specific to the constitutive exons flanking nSR100-regulated alternative exons in Daam1, Clasp2, Dock4, and Elmo2 transcripts. Bands corresponding in size to exon-included and exon-skipped isoforms
are indicated. (D) Neuro2a cell lines stably expressing control (lanes 1, 3, and 5) or nSR100 targeting (lanes 2, 4, and 6) shRNAs and pWZL-hygro empty vector (lanes 1 and 2), T7 epitope-tagged SF2/ASF (lanes 3 and 4), or nSR100 cDNA (lanes 5 and 6). Top three panels show western blots monitoring the expression of nSR100 (top panel), T7-tagged SF2/ASF (middle panel), and α-tubulin (bottom panel). Bottom three panels show RT-PCR assays using primers to specifically monitor AS in the Daam1, Zdhhc20, and Asph transcripts.

knockdown, indicating a low false-negative detection rate (data not shown). The reduced levels of exon inclusion observed upon knockdown of nSR100 persisted following RA-induced differentiation in all of five analyzed cases (Figure 3.7). These results indicate that nSR100 acts predominantly to promote alternative exon inclusion prior to the formation of neurites in Neuro2a cells.

To address whether the altered exon inclusion levels in the knockdown experiments were specifically caused by the shRNA-reduced expression of nSR100, an RNAi-resistant form of nSR100 mRNA was stably expressed in the nSR100 shRNA knockdown line and splicing levels of several alternative exons were monitored by RT-PCR assays. Expression of nSR100 in the knockdown line, which was confirmed by western blotting (Figure 3.6 D, top panel, lanes 5 and 6), fully restored the inclusion of the tested alternative exons (Figure 3.6 D, lower three panels, compare lanes 5 and 6 with 1-4). In contrast, stable expression of another SR protein, SF2/ASF, did not function in this manner (Figure 3.6 D, all panels, lanes 3 and 4). These results support the conclusion that nSR100 functions specifically to promote exon inclusion. The nSR100 regulated exons were estimated to represent ~11% of AS events identified as being differentially regulated in nervous system tissues by microarray profiling (see Supplementary Information).

A Gene Ontology (GO) annotation enrichment analysis revealed that genes containing nSR100-regulated exons are significantly enriched in functions associated with membrane dynamics and cytoskeleton remodeling (Table 3.1), processes that are critical for the differentiation of neurons and neurite extension. Also supporting functional significance, nSR100 regulated alternative exons are significantly more often frame preserving (66.7% vs. 45.4%; p < 0.005, Fisher’s exact test) and conserved in human transcripts (62.8% vs. 24.5%; p < 1 x 10⁻⁹, Fisher’s exact test), than are the other profiled exons (Supporting data file SD 3.2, CD). These results thus reveal that nSR100 regulates a network of mostly conserved AS events in
functionally related genes that play important roles in the formation and function of the nervous system.

Figure 3.7: nSR100-regulated splicing patterns are established prior to induction of neurite outgrowth in Neuro2a cells.

Semi-quantitative RT-PCR assays for five RT-PCR validated targets from the microarray analysis (in the Daam1, Asph, and Zdhhc20 genes) and two additional validated alternative splicing (in genes corresponding to Acbd5 and Riken cDNA 1810009A16). Total RNA was extracted from Neuro2a cells stably expressing control shRNAs or nSR100 shRNAs placed under normal conditions (10% FBS, lanes 1 and 2) and conditions promoting neurite outgrowth (2% FBS and 2% FBS + RA, lanes 3-6).
Table 3.1: Overrepresented Gene Ontology (GO) terms associated with genes containing nSR100 regulated alternative exons.

The genes with enriched GO terms fall under two main categories: membrane organization and/or endocytosis or the cytoskeleton. The GO IDs corresponding to the GO terms and the represented genes associated with each GO term are listed. Genes and GO terms were grouped together if a particular GO term and its associated genes were a subgroup of another GO term/gene group. Alternatively, groups of genes associated with different GO terms were also grouped together if only one gene was found different between the groups.

3.4.5 An nSR100 regulatory code

To gain insight into the mechanism by which nSR100 regulates neural-specific AS, I identified motifs of 5-10 nucleotides in length that are significantly enriched in the regulated alternative and adjacent constitutive exons, and in the 500 nucleotides of intron sequence that flank these exons (Figure 3.8 A; refer to Materials and Methods). The most enriched and prevalent motifs are C/U-rich, and are specifically located within the upstream and/or downstream intronic regions flanking nSR100 regulated exons (Figure 3.8 A, FDR-corrected p value < 0.01). Additional high scoring motifs were identified using the same procedure (Figure A2.3, Appendix 2; see Discussion).

As a parallel approach to defining cis-regulatory elements, I generated an AS minigene reporter “Daam1 FL” consisting of a 2.8 kb genomic fragment encompassing nSR100-regulated Daam1 exon 16 (Figure 3.8 B). Daam1 is required for neurite formation and outgrowth during development (Matusek et al., 2008), and skipping of exon 16 in non-neuronal cells is predicted to
Figure 3.8: Elements of an nSR100 Regulatory Code

(A) Pyrimidine-rich motifs identified as significantly enriched in the 500 nucleotides of upstream and/or downstream intron sequence flanking nSR100-regulated alternative exons. Motifs with FDR-corrected hypergeometric p value scores <0.01 are shown. (B) (Left) RT-PCR assays monitoring AS levels of Daam1 exon 16 minigene reporters transfected in NIH 3T3 cells and in Neuro2a cells expressing control or nSR100-targeting shRNAs. The lengths of upstream and downstream intron sequence included in each minigene reporter are indicated above the gel. (Right) The diagram indicates the length of *Daam1* genomic DNA sequence included in each minigene reporter (green bars) and the degree of AS repressor activity associated with these regions (indicated by pluses). (C) RT-PCR assays monitoring the AS levels of transcripts derived from additional Daam1 minigene reporters in Neuro2a cells expressing control shRNA or nSR100-targeting shRNAs. Gels on the left display AS patterns observed with reporters containing either 271 or 250 nucleotides of upstream intron sequence and 50 nucleotides of...
downstream intronic sequence. Gels on the right display AS patterns from reporters containing 483 and 146 nucleotides of upstream and downstream sequence, respectively, when region -271 to -250 and additional upstream pyrimidine nucleotides (up to position -280; see Figure A2.4, Appendix 2) are unchanged (-483 to 146) or substituted (-483 to 146 mut) with an unrelated sequence of equal length. Percent exon inclusion levels (%inc) are shown below gel images in (B) and (C).

disrupt a formin homology-like domain that likely mediates important interactions with signaling partners and components of the cytoskeleton.

The Daam1 FL minigene reporter recapitulated the AS pattern of endogenous Daam1 transcripts, with the alternative exon displaying more inclusion in Neuro2a cells expressing control shRNAs relative to cells expressing shRNAs targeting nSR100 (Figure 3.8 B). Moreover, the inclusion of Daam1 exon 16 appears to be neural-specific, since reporter transcripts expressed in NIH-3T3 and myoblast C2C12 cell lines undergo complete exon skipping (Figure 3.8 B and data not shown). These results indicate that cis-elements required for nSR100-dependent regulation of Daam1 exon 16 are contained within the 2.8 kb Daam1 FL reporter genomic fragment.

Deletion of the Daam1 reporter intron sequences followed by refined mutagenesis revealed a region 250 to 271 nucleotides upstream of the alternative exon that significantly affected nSR100-dependent exon inclusion (Figures 3.8 B and 3.8 C). Deletion of this region resulted in increased exon inclusion. This effect was most pronounced in the nSR100 knockdown cell line (Figure 3.8 C, compare lanes 1 and 2). Similar effects were observed when this region and neighboring nucleotides were substituted with unrelated sequence (Figure 3.8 C, compare lanes 3 and 4). These results indicate that this 21 nucleotide region contains a strong repressor element that results in the skipping of Daam1 exon 16 in the absence of nSR100. nSR100 thus appears to play a critical role in overcoming the repressive effects exerted by this element.

In agreement with our computational analysis, the -271 to -250 AS repressor region and its neighboring nucleotides are highly enriched in pyrimidine nucleotides (Figure A2.4, Appendix 2), and two of the computationally-identified motifs overlap this region. Therefore, both approaches to defining cis-regulatory elements strongly implicate pyrimidine-rich motifs in nSR100-dependent splicing regulation.
3.4.6 nSR100 promotes nPTB protein expression

Pyrimidine rich motifs have previously been implicated in neural-specific regulation of AS involving PTB (also hnRNPI/PTBP1), a widely acting splicing repressor, and its neural-enriched paralog nPTB (also brPTB/PTBP2) (see Introduction and below). The switch from PTB to nPTB during neuronal differentiation is regulated in part by a neuronal miRNA miR-124 (Makeyev et al., 2007). In non-neuronal cells, expression of PTB results in skipping of exon 10 in nPTB pre-mRNA, and this leads to the introduction of a premature termination codon that elicits nonsense-mediated mRNA decay (NMD) of nPTB transcripts (Boutz et al., 2007b; Makeyev et al., 2007; Spellman et al., 2007). In neuronal cells, miR-124-repression of PTB facilitates nPTB exon 10 inclusion and productive nPTB expression.

To address whether PTB and/or nPTB are involved in the regulation of nSR100-dependent AS events, we knocked down these factors individually or together using siRNAs (Figure 3.9 A). As expected, knockdown of PTB led to increased inclusion of exon 10 in nPTB pre-mRNA. RT-PCR assays were then used to monitor the AS levels of Daam1 exon 16 and nine randomly selected, RT-PCR-validated nSR100-regulated exons (Figures 3.9 A and 3.10).

In contrast to the effect of nSR100 knockdown, Daam1 exon 16 and eight of the nine other exons displayed increased levels of inclusion when PTB levels were reduced, and to a lesser extent when nPTB was reduced. Simultaneous depletion of nPTB and PTB resulted in levels of exon inclusion that were at least as high and in some cases higher than the levels observed when PTB was knocked down alone. These findings establish a widespread role for PTB/nPTB in the repression of nSR100-dependent AS events. More specifically, these proteins promote skipping of the same exons that require nSR100 for neural-specific inclusion. Consistent with a previous proposal (Markovtsov et al., 2000), the less repressive activity of nPTB relative to PTB suggests that it might facilitate neural-specific AS by establishing complexes that are more permissive to positive acting factors.

To further explore the functional relationship between nSR100 and PTB/nPTB proteins, we next asked: (1) does nSR100 promote neural-specific AS by regulating the relative levels of nPTB and PTB proteins, (2) do PTB/nPTB bind directly to C/U-rich sequences involved in mediating nSR100-dependent regulation, and (3) does nSR100 act in a dominant-positive manner to counteract the repressive activity of nPTB?
Our microarray profiling experiments (Figure 3.6) in fact predicted increased skipping of nPTB exon 10 in Neuro2a cells depleted of nSR100 (Supporting data file SD 3.2, CD). Confirming this prediction, knockdown of nSR100, while not altering PTB mRNA or protein levels, resulted in decreased inclusion of nPTB exon 10 and, consequently, reduced levels of nPTB mRNA and protein (Figure 3.9 B). Consistent with the repressive effects of PTB and nPTB on nSR100 regulated exons, both proteins bind directly and specifically to C/U-rich intron regions upstream of Daam1 exon 16 inclusion (Figure A2.4, Appendix 2; Figure 3.11). However, essentially no binding to these regions was observed when comparable levels of purified nSR100 (Figure 3.11 A) were added alone or in combination with PTB or nPTB (Figure 3.12 A). At much higher amounts of nSR100, a low level of binding was observed although this appeared to be relatively less specific (Figure 3.12 A).

Further supporting a critical role for nSR100 in promoting neural-specific exon inclusion of PTB/nPTB repressed exons, when nPTB is over-expressed in Neuro2a cells increased skipping of Daam1 exon 16 is observed (Figure 3.12 B). However, simultaneous over-expression of nSR100 and nPTB effectively alleviates the repressive activity of nPTB and results in ~100% inclusion of Daam1 exon 16 (Figure 3.12 B). This confirms that nSR100 is required to promote the inclusion of neural-specific exons in cells that also express nPTB.
Figure 3.9: nSR100 Functions by Promoting nPTB Expression and by Binding to its Regulated Target Transcripts

(A) (Left) RT-PCR assays monitoring steady-state mRNA expression levels and AS patterns of nSR100, PTB, nPTB, and Daam1 transcripts in Neuro2a cells transfected with control siRNAs (lane1) or siRNAs targeting PTB, nPTB, or both of these factors. Expression levels of Gapdh transcripts were monitored to control for recovery and loading.

(B) RT-PCR assays monitoring the expression levels and/or AS patterns of nSR100, PTB, nPTB, and Gapdh transcripts in Neuro2a cells expressing control or nSR100-targeting shRNAs (top four panels). Western blotting (bottom three panels) was performed using lysates from the cells assayed in (A) and antibodies capable of detecting PTB paralogs (α-pan-PTB) and nPTB specifically (α-nPTB). Tubulin levels were monitored to control for sample recovery and loading.
(C) Immunodepletion of nSR100 from splicing competent Weri whole cell extracts. Western blotting on extracts immunodepleted of nSR100 (ΔnSR100) using anti-nSR100 antiserum (lane 4) and mock-immunodepleted (mock) using rabbit anti-mouse antibody (lane 3). Two different amounts of total extract, corresponding to 100% and 10% of the amounts of extract loaded in lanes 3 and 4, are shown in lanes 1 and 2, respectively.

(D) Purified recombinant nSR100 protein specifically promotes the inclusion of Daam1 exon 16 in vitro. A T7 pre-mRNA transcript (see diagram) consisting of strong constitutively spliced 5’ and 3’ exons (MINX) and Daam1 alternative exon 16, flanked by its native upstream and downstream intron sequence, was incubated in the Weri extracts (C) supplemented with and without recombinant proteins (Figure 3.10 A) and splicing activity was monitored by RT-PCR assays and primers specific for the MINX exons. The ΔnSR100 extract was supplemented with 10, 25, 50, 100, and 200 ng of purified nSR100 protein (lanes 4–8, respectively) or with 5, 12.5, 25, 50, and 100 ng of SF2/ASF protein (lanes 9–13). Splicing of the Daam1 pre-mRNA in the mock and ΔnSR100 Weri extracts without added proteins is shown in lanes 2 and 3, respectively. Lane 1 shows input pre-mRNA.

(E) Recombinant nSR100 protein is not sufficient to promote Daam1 exon 16 inclusion in a nonneural origin (HeLa) cell extract. Weri and HeLa cell extracts were incubated without (lanes 2 and 6) or with 10, 25, and 50 ng nSR100 protein, and splicing activity was monitored as described in (D). Input pre-mRNA is shown in lane 1.

(F) Anti-nSR100 antiserum enriches a protein of the approximate size of nSR100 that is UV crosslinked to Daam1 intronic regulatory sequences. A radiolabeled T7 transcript consisting of the upstream intronic regulatory region (nucleotides -280 to -1) of Daam1 exon 16 was incubated under splicing conditions in Weri cell extracts, without (lane 1) or with (lane 2) UV exposure, prior to extensive treatment of the extracts with RNase. Proteins in these extracts were immunoprecipitated with anti-nSR100 (lane 4) or control (lane 3; rabbit anti-mouse antibodies). Total extract and immunoprecipitated proteins were separated on an SDS-PAGE gel, which was dried and exposed to film. The asterisk denotes the crosslinked protein species migrating at approximately the same location as nSR100.

(G) CLIP reveals that nSR100 interacts directly with endogenous pre-mRNA transcript regions overlapping nSR100 regulated exons. Neuro2a cells were exposed (lanes 1–5) or not exposed (lanes 6–10) to UV light. Immunoprecipitation was performed from lysates prepared from the cells using anti-nSR100 antibody (lanes 3, 5, 8, and 10) or a control (rabbit anti-mouse) antibody (lanes 2, 4, 7, and 9). Detection of immunoprecipitated pre-mRNA regions was performed using primer pairs designed to amplify exon-intron regions within the nSR100 target genes Daam1, Asph, Elmo2, and Dock 4 and as controls to the nontarget genes Myo5a (false positive from microarray predictions) and Pbrm1. RT was omitted from the PCR reactions in lanes 2, 3, 7, and 8 to control for possible genomic DNA contamination.
Figure 3.10. Additional nSR100-regulated splicing events are also regulated by PTB and nPTB.

RT-PCR assays comparing the splicing patterns of eight AS events in Neuro2a cells transfected with control siRNAs or siRNAs targeting PTB and/or nPTB. In all cases knockdown of PTB and in some cases knockdown of nPTB leads to more alternative exon inclusion in these target AS events.
Figure 3.11: PTB and nPTB associate with Daam1 upstream intronic RNA *in vitro*

(A) Increasing amounts of BSA, or purified recombinant GST-tagged PTB, GST-tagged nPTB, His-tagged nSR100, and T7-tagged SF2/ASF were loaded on a 10% SDS polyacrylamide gel and stained with coomassie blue. (B-D) RNA mobility shift assays testing the ability of GST-tagged PTB and nPTB to bind the normal Daam1 intronic region (nucleotides -280 to -1) or mutant variants, as indicated by the legend shown in Figure S8.
Figure 3.12: nSR100 alone or in combination with PTB or nPTB does not associate strongly with the Daam1 intronic regulatory sequence, but can overcome the repressive activity of nPTB in vivo.

(A) RNA gel mobility shift assays testing for binding of the Daam1 intronic fragment with purified nSR100 alone (lanes 2-5) or in combination with 10 nM GST-PTB or GST-nPTB (lanes 7-10 and 12-15, respectively). Final concentrations of nSR100 used in binding reactions were 25, 50, 100 and 200 nM. (B) Neuro2a cells were cotransfected with a Daam1 minigene reporter and indicated amounts of myc-nSR100 and/or GFP-nPTB (lanes 2-5). Top panel displays splicing patterns of the minigene reporter as determined by semi-quantitative RT-PCR. Lane 1 represents the splicing pattern of the minigene reporter transfected alone. The bottom three panels show western blots to detect transfected nSR100 (anti-myc, second panel), nPTB (anti-GFP, third panel) and alpha tubulin (anti-tubulin, bottom panel).

3.4.7 nSR100 promotes neural-specific exon inclusion in vitro

nSR100 appears to act together with nPTB and possibly other factors to promote neural-specific exon inclusion. We next asked whether nSR100 functions in this manner in vitro. Splicing-competent extracts were prepared from neuronal Weri-Rb1 cells and efficiently (>90%) immunodepleted of nSR100 using affinity-purified anti-nSR100 antibody (Figure 3.1E), or mock-depleted using rabbit anti-mouse antibody (Figure 3.9 C, compare lanes 1 and 2 with lanes 3 and 4). A T7-transcribed, three-exon pre-mRNA consisting of Daam1 exon 16 with 280 nucleotides of upstream and 145 nucleotides of downstream intronic sequence, inserted between strong constitutive exons derived from an adenovirus pre-mRNA substrate (MINX), was assayed for splicing activity in these extracts.
This substrate splices efficiently in the mock-depleted Weri extract, producing exon 16 included and skipped transcripts (Figure 3.9 D, lane 2). When nSR100 is immunodepleted, increased exon 16 skipping is detected (Figure 3.9 D, lane 3), whereas addition of increasing amounts of purified nSR100 protein results in ~100% inclusion of exon 16 (Figure 3.9 D, lanes 4-8). This activity is specific, since addition of comparable levels of the SR family protein SF2/ASF (Figure 3.11 A) does not result in a substantial increase in exon 16 inclusion (Figure 3.9 D, compare lanes 4-8 with 9-13). Moreover, addition of purified nSR100 protein to a HeLa splicing extract does not promote the inclusion of Daam1 exon 16 (Figure 3.9 E, compare lanes 2-5 with 6-9). These results demonstrate that nSR100 functions to promote the inclusion of Daam1 exon 16 in a neural cell extract. However, consistent with the results demonstrating that nSR100 functions in part by promoting the expression of nPTB, it is not sufficient to promote a neural-specific splicing pattern in a non-neural extract.

We next performed in vitro and in vivo UV cross-linking followed by immunoprecipitation (CLIP; Supplemental Information) to establish whether, in the context of other cellular factors, nSR100 interacts directly and specifically with its regulated target pre-mRNAs. Anti-nSR100 antibody specifically enriched a radiolabeled protein approximately the size of nSR100 from Weri extract incubated and cross-linked in the presence of a 32P-UTP-labeled Daam1 RNA consisting of the upstream intronic regulatory sequences (Figure 3.9 F, compare lanes 2 and 4). Using in vivo CLIP, anti-nSR100 immunoprecipitates were analyzed by RT-PCR assays to detect intron-exon regions overlapping five validated, nSR100-regulated AS events. In all cases, immunoprecipitation of these regions was observed (Figure 3.9 G; representative examples are shown) and was UV- and anti-nSR100 antibody-dependent (Figure 3.9 G, compare lanes 4-5 with 9-10 and lanes 9 and 10). Moreover, little to no anti-nSR100 immunoprecipitation was detected for three regions overlapping profiled AS events that are not regulated by nSR100 (Figure 3.9 G, lower two panels and data not shown).

The results described above provide evidence that nSR100 establishes neural-specific AS patterns by forming direct interactions with its regulated target transcripts, as well as by promoting increased expression of nPTB.
3.4.8 A role for nSR100 in nervous system development

The high degree of homology (Figure 3.2) between mammalian and zebrafish nSR100 (ZnSR100) indicates that its function is likely conserved throughout the vertebrate lineage. Accordingly, we next examined the distribution and function of ZnSR100 in developing zebrafish embryos.

ZnSR100 expression was examined by whole-mount RNA *in situ* hybridization and no expression was detected prior to gastrulation stages (data not shown). At 24 hours post fertilization (hpf), significant expression was observed in the developing brain, including the tectum (tec) and telencephalon (tel) (Figure 3.13 A/A’). Expression was also observed around the trigeminal ganglion (tg), at the mid-hindbrain boundary and throughout the hindbrain (hb). At 36 hpf, additional ZnSR100 expression was observed in the cerebellum (cb) and within the retina (r) (Figure 3.13 B/B’). At 48 hpf, stronger and more widespread ZnSR100 expression was detected throughout the developing eyes and brain (Figure 3.13 C/C’). These results are highly consistent with the expression and immunostaining data for mammalian nSR100 (Figures 3.1 and 3.3).

To knockdown ZnSR100 function we injected single cell-staged embryos with an antisense morpholino oligonucleotide (MOspl) targeted against the 5´ splice site of exon 5 of ZnSR100 pre-mRNA. Severe neural degeneration was observed throughout the brain and spinal cord of 24 hpf embryos (data not shown). This phenotype resembles a common off-target effect of MO injection caused by induction of the p53-dependent cell death pathway, and can be circumvented by knockdown of p53 activity (Robu et al., 2007). We therefore co-injected MOspl together with p53-MO. RT-PCR assays confirmed that the MOspl oligonucleotide caused nearly complete disruption of splicing of ZnSR100 exon 5 to exon 6 (Figure A2.5, Appendix 2). We therefore expected efficient knockdown of ZnSR100 function.

At 50hpf, control embryos injected with relatively high levels of p53-MO alone appeared phenotypically normal (Figure 3.13 D; data not shown). In contrast, over 90% of MOspl-injected “morphant” embryos exhibited consistent developmental abnormalities that included a curved body axis and expanded brain ventricles (n=200; Figures 3.13 E and A2.6, Appendix 2). To examine neuronal differentiation in nSR100 morphant embryos, we performed RNA *in situ* hybridization for *islet-1*, a LIM homeodomain-containing transcription factor that is an early
Figure 3.13: nSR100 Regulates Vertebrate Nervous System and Sensory Organ Development

(A–C) RNA in situ hybridization assays monitoring ZnSR100 expression in the developing zebrafish embryo.

(A and A’) expression at 24 hpf, lateral and frontal views, respectively.

(B and B’) nSR100 expression at 36 hpf, lateral and dorsal views, respectively.

(C and C’) nSR100 expression at 48 hpf, lateral and dorsal views, respectively.

(D) p53 morpholino antisense oligonucleotide (MO) control-injected embryo at 50 hpf.

(E) p53 + ZnSR100 MOspl-injected embryo at 50 hpf.

(F) p53 + ZnSR100 MOspl-injected embryo at 50 hpf, rescued through coinjection of ZnSR100 mRNA.
(G–I) RNA in situ hybridization assays monitoring islet-1 expression in 50 hpf embryos. p53-MO-injected control (G), nSR100spl morphant embryo (H), and nSR100spl morphant embryo rescued through coinjection of 20 pg ZnSR100 mRNA (I) are shown. (J–M) Confocal projections of trigeminal ganglion in Tg(sensory:GFP) transgenic embryos. Control embryo at 30 hpf (J), representative nSR100spl morphant embryo at 30 hpf (K), control embryo at 36 hpf (L), and representative nSR100spl morphant embryo at 36 hpf (M) are shown. (N) nSR100-regulated alternative exons conserved between zebrafish and mouse display altered splicing patterns in nSR100spl morphant embryos, relative to control embryos. Total RNA was isolated from the heads of wild-type and nSR100spl morphant embryos at 50 hpf, and semiquantitative RT-PCR was performed to analyze AS patterns. Percent exon inclusion values are shown below each lane.

marker for the differentiation of diverse neuronal populations. At 24 hpf, islet-1 expression in ZnSR100 morphants appeared identical to p53-MO-injected controls (Figure A2.7, Appendix 2). Strikingly, islet-1 expression in 50hpf nSR100 morphant embryos was markedly reduced, consistent with a severe disruption to neuronal differentiation (Figure 3.13 G,H; compare to Figure A2.7). These effects were not due to significant developmental delay, as general embryonic morphogenesis and patterning appeared normal in nSR100 morphant embryos (Figure A2.6, Appendix 2).

Similar phenotypes as observed in MOspl morphants were observed after injection of a translation-blocking morpholino targeted to the start codon of ZnSR100 (MOatg; data not shown). Moreover, the morphological and neuronal differentiation defects observed in ZnSR100 MOspl-injected embryos could be rescued by co-injecting in vitro transcribed, capped ZnSR100 mRNA (n=42/52; Figure 3.13 F,I). These results further indicate that the observed phenotypes are not due to off-target effects of MO injection, but rather are the specific result of reduced ZnSR100 expression.

We next examined the consequence of knocking-down ZnSR100 function on the axonal extension and branching (arborization) of trigeminal sensory neurons. Trigeminal neurons are among the first neurons to develop in vertebrates and begin to display marked arborization at ~16 hpf in zebrafish embryos. To visualize axon arbors in vivo, we utilized Tg(sensory:GFP) transgenic animals that drive GFP expression in Rohon-Beard and trigeminal sensory neurons (Sagasti et al., 2005). Confocal projections of trigeminal ganglia in control embryos at 30 hpf (n=7) and 36 hpf (n=6) revealed multiple axonal projections and an elaborate array of peripheral
sensory arbors (Figure 3.13 J,L). In contrast, visualization of trigeminal sensory neurons in ZnSR100 morphant embryos at 30hpf (n=7) and 36hpf (n=6) revealed obvious defects in the formation of peripheral sensory arbors (Figure 3.13 K,M). As before, these defects could be rescued by co-injection of ZnSR100 mRNA (Figure A2.8, Appendix 2). Together, the results described above indicate a critical role for nSR100 in neuronal differentiation in a whole animal context.

Finally, we analyzed the effects of MOspl disruption of ZnSR100 on the AS levels of exons that are conserved between zebrafish and mouse. Reduced inclusion levels were observed for 6/7 exons that require nSR100 for inclusion in Neuro2A cells, whereas two conserved alternative exons that are not regulated by nSR100 in Neuro2a cells were not affected (Figure 3.13 N and data not shown). These results, which additionally support the specificity of the effects of MOspl-targeting of ZnSR100 function, lead us to conclude that nSR100 regulates a program of AS events that is highly conserved between zebrafish and mammals.
3.5 Discussion

3.5.1 A neural- and vertebrate-specific activator of alternative splicing

The discovery and characterization of nSR100 reveals how a large set of brain-specific exons are positively regulated in genes that play numerous critical roles in vertebrate nervous system development. While the evolutionary origin of nSR100 is unclear, its short region of perfect identity with an SRm300 splicing factor-like gene and extensive SR/RS-repeat regions (Figure 3.1), suggests that a duplication event involving an ancestral SRm300-related gene may have led to the emergence of nSR100. Subsequently, nSR100 may have been co-adopted with nPTB and other factors to afford the evolution of increased neural AS complexity. Consistent with the evolution of extensive co-dependent functions, nPTB also appears to have evolved around the time of emergence of the vertebrate lineage (Barbosa-Morais et al., 2006), and both proteins are highly enriched in post-mitotic neurons relative to glial cells (Figure 3.3; Boutz et al., 2007b). These observations support the conclusion that the emergence of nSR100 played a critical role in the evolution of the expanded proteomic and functional complexity of the vertebrate nervous system.

3.5.2 Global and local roles for nSR100 in neural-specific alternative splicing

The global regulatory properties of nSR100 and its remarkably tight neural-restricted expression pattern indicate that it serves as the specificity factor for ~11% of mammalian nervous system-specific AS events (refer to Materials and methods). While functioning in conjunction with nPTB, it is possible that nSR100 also functions with other neural-enriched AS regulators such as members of the Fox, CUGBP/CELF and MBNL families of proteins. In fact, our motif analyses in the present study reveal putative binding sites for several of these factors in the intronic sequences flanking nSR100-regulated exons (Figure A2.3, Appendix 2). Moreover, co-immunoprecipitation experiments reveal that nSR100 associates with specific SR proteins recognized by the pan-SR-specific monoclonal antibody mAb104 (J.A.C, D.O, B.R. and B.J.B, unpublished observations).

Together, these findings support a model (Figure 3.14) in which nSR100 binds to target transcripts and functions with nPTB and possibly other factors to overcome the repressive forces
of PTB during neural differentiation. This model is consistent with emerging evidence for complex regulatory codes involving combinations of different cis-regulatory elements bound by corresponding trans-acting factors in the regulation of neural-specific AS events (Li et al., 2007). It is also consistent with the established properties of SR family and SR-related proteins, which interact with one another in different contexts to form positive-acting regulatory complexes on pre-mRNA (Lin and Fu, 2007). Indeed, the relatively large size of nSR100 and its extensive RS domains suggests that it could form important interactions with many splicing factors to promote exon inclusion in neural cells.
Figure 3.14: Model for nSR100-Dependent Regulation of Neural-Specific Alternative Splicing

nSR100 activates the inclusion of neural-specific alternative exons in a regulatory circuit that involves the switch between PTB and nPTB proteins. In one part of this circuit, nSR100 promotes the inclusion of exon 10 in nPTB pre-mRNA (green solid arrows), leading to the increased expression of nPTB relative to PTB, which facilitates but is not sufficient for neuronal-specific alternative exon inclusion. In the presence of nPTB and other factors, such as SR family and/or SR-related proteins, nSR100 forges positive-acting interactions by binding to its regulated target transcripts.

3.5.3 Functions of an nSR100-regulated neural exon network

Many nSR100-regulated exons are found in genes required for remodeling of the cytoskeleton, including several with neuronal GTPase activity. Rho family small GTPases are key regulators of the actin cytoskeleton during neuronal morphogenesis and require guanine nucleotide exchange factors (GEFs) for activation (Govek et al., 2005). Besides Daam1, which is
required for the activation of the GTPase RhoA (Habas et al., 2001), other nSR100-regulated exons are located in Dock4 (a GEF) and its binding partner Elmo2 (see Figure 3.6). A complex of Dock4, Elmo2 and an adapter protein CrkII, promotes dendritic growth and branching in hippocampal neurons through the activation of another small GTPase, Rac. These observations are consistent with our finding that GTP-based signaling represents one of the most highly enriched functions associated with genes that display nervous system-specific AS (Fagnani et al., 2007a). Detailed annotation of the genes with nSR100-regulated exons that have the potential to alter coding sequences reveals that approximately half of the corresponding proteins can be assembled into a network, based on prior published evidence for protein-protein interactions (J. Ellis, J.A.C and B.J.B, unpublished observations). More than half of the regulated exons in this model for an nSR100-regulated network are conserved between human and mouse. It is therefore tempting to speculate that nSR100 target exons function in an elaborate protein-protein interaction network that is critical for remodeling of the cytoskeleton during neuronal differentiation. Future studies will be directed at elucidating the functions of nSR100-regulated AS events.
Chapter 4

Resources for in vivo analysis of alternative splicing in Caenorhabditis elegans

Part of this chapter (data from Figures 4.1-4.4) is derived from the following published article: A.K. Ramani, J.A. Calarco, Q. Pan, S. Mavandadi, Y. Wang, A.C. Nelson, L.J. Lee, Q. Morris, B.J. Blencowe, M. Zhen, and A.G. Fraser. Genome Research (2011) 21(2): 342-8. Genome-wide analysis of alternative splicing in Caenorhabditis elegans. Q. Pan designed the microarray platform with assistance from M. Zhen and myself. S. Mavandadi and Q. Morris designed the PATA algorithm for splicing microarray analysis. A.K. Ramani performed the RNA-Seq analysis along with my assistance. I performed all RT-PCR validation experiments, and A.K. Ramani and myself analyzed the data in Figures 4.1-4.4. The remainder of this chapter (data from Figures 4.5-4.9) forms a portion of a manuscript in preparation, and I have performed all experiments and data analysis corresponding to this work. Additional figures can be found in Appendix 3, and supporting data files can be found in the accompanying CD.
4.1 Abstract

Alternative splicing (AS) plays a crucial role in the diversification of gene function and regulation. Consequently, the systematic identification and characterization of spatio-temporally regulated splice variants is of critical importance to understand animal development. New resources have been developed to facilitate such analyses in *Caenorhabditis elegans*. First, we have combined quantitative AS microarray profiling and high throughput RNA sequencing to discover novel splicing events and hundreds of developmentally-regulated AS events. Second, a new fluorescent reporter system has been developed to allow *in vivo* visualization of any AS event at single cell resolution; its application has revealed complex, cell type-specific patterns of splicing regulation in the nervous system. Finally, I have developed a fosmid-based, isoform-specific transgene expression rescue approach that can uncover the unique functions of individual splice variants. Collectively, these resources will enable researchers to rapidly extend large-scale exploration of splice variants to focused, high resolution *in vivo* functional assays.
4.2 Introduction

Alternative splicing is one of the key mechanisms that has evolved in metazoans to generate increased transcriptome complexity, and is often differentially regulated across tissues and during development, suggesting that individual isoforms may serve specific spatial or temporal roles (Hartmann et al., 2009; Licatalosi and Darnell; Nilsen and Graveley, 2010). This differential regulation has the potential to increase molecular and functional diversity cellullarly heterogeneous organs such as the brain, which undergoes extensive AS (Li et al., 2007; Lipscombe, 2005; Ule and Darnell, 2006). However, since most analyses of AS have been carried out in tissue samples that often contain diverse classes of cells, the extent to which AS is differentially regulated in individual cell types of the same organ or tissue is not known.

The importance of proper regulation of AS during development has been elegantly demonstrated through studies of the sex determination pathway in Drosophila. In this pathway, the female-specific expression of a splicing regulator transformer stimulates the inclusion of exons in transcripts of the doublesex and fruitless transcription factor genes (Forch and Valcarcel, 2003; Lopez, 1998). The female-specific isoforms of these transcription factors subsequently activate the expression of genes required for female development, while the male-specific variants induce a gene expression program important for male differentiation (Dulac, 2005; Shirangi and McKeown, 2007). Additional spatio-temporally regulated AS networks are likely to exist in metazoans. The discovery and characterization of these AS networks, and an understanding of how they are integrated with other layers of gene regulation, will be necessary for a more complete understanding of development (Blencowe, 2006).

The nematode C. elegans, with its amenability for genetic manipulations, defined cell lineage, simple but extensively differentiated nervous system, conserved splicing machinery, and established molecular and cell biological resources, provides an attractive yet relatively under-explored model organism with which to study the regulation and functions of AS in vivo. Tissue- or developmental stage-specific AS events have been identified in ~70 C. elegans genes which, like those identified in vertebrates, are controlled by combinatorial interactions between splicing factors (Barberan-Soler and Zahler, 2008; Kuroyanagi et al., 2007; Ohno et al., 2008). These
studies, while informative, were limited to the analysis of isoforms represented by EST/cDNA libraries, which currently do not provide full coverage of the transcriptome.

The advent of deep sequencing technologies has now enabled transcriptome analyses at unprecedented levels of sensitivity and precision, and these have recently been applied in *C. elegans* (Hillier et al., 2009; Mangone et al., 2010; Ramani et al., 2009). However, due in part to limited depth of coverage of existing datasets and the ongoing development of analytical software, it has remained a major challenge to systematically categorize and quantify the relative usage of isoforms across development for both known and novel AS events. Moreover, as genome-wide analyses in whole animal samples provide little information about the spatial regulation of detected AS events, additional strategies are required to determine cell type-specific splicing patterns and functional roles of these alternative isoforms.

In this study, a set of integrated resources has been established allowing the discovery and physiological characterization of isoforms in *C. elegans*. Using deep sequencing of mRNA (RNA-Seq) and an AS microarray platform, we have generated an extensive and quantitative profile of alternative splice junction usage during *C. elegans* development. We have identified thousands of new AS events in *C. elegans*, and hundreds that are differentially regulated during development. To facilitate access to this information, we have created the “*C. elegans* Splice Browser”, a web resource for splicing analysis. To examine the tissue-specificity of splicing patterns *in vivo*, I developed a new bichromatic splicing reporter system. Applying this reporter reveals a remarkable degree of splicing regulatory complexity among neuronal subtypes. Finally, to facilitate the functional analysis of splice variants of interest, I developed a fosmid recombineering-based isoform-specific expression approach. By applying this system, I have uncovered splice variant-specific functions of AS events identified from our genome-wide and bichromatic reporter analyses. Collectively, the resources described in this study provide an integrated experimental and bioinformatic pipeline that permits the rapid identification and characterization of AS events at single cell resolution *in vivo*, as well as the functional analysis of these AS variants during animal development.
4.3 Materials and Methods

4.3.1 Strains and maintenance

*C. elegans* strains were cultured and maintained on NGM agar plates seeded with OP50 *E. coli* according to standard protocols (Brenner, 1974). Strains with the following genotypes were used in this study: the N2 Bristol wild-type isolate, *lin-15(n765ts), juIs1, cle-1(cg120); juIs1*, and *mod-1(ok103)*.

4.3.2 RNA extraction, purification, and cDNA synthesis

For RNA-Seq analysis, total RNA was prepared from mixed stage or synchronized whole animals using Trizol solution (Invitrogen) according to the manufacturer’s protocol, and subsequently cleaned using RNeasy columns (QIAGEN) followed by treatment with 10 units of Dnase I (Roche) in 1x One-Phor-All buffer (Amersham) for 30 minutes. Poly A+ RNA was then purified from total RNA using Oligotex midi kits (QIAGEN) according to the manufacturers protocol. cDNA was then produced using SuperScript™ Double-Stranded cDNA Synthesis Kit (Invitrogen) and purified using a PCR Purification kit (QIAGEN). Sequence data for the resulting cDNA was then obtained using an Illumina Genome analyzer as described (Wilhelm et al., 2008) For microarray analysis, total RNA was prepared from synchronized larval (L1 through L4) and adult stage animals using Tri reagent (Sigma Aldrich) according to manufacturers recommendations. Poly A+ RNA was isolated and used to synthesize cDNA as described previously (Ramani et al., 2009).

4.3.3 Microarray design, hybridization and analysis

Alternative splicing events were identified from alignments of *C. elegans* mRNA/EST sequences (UniGene Build #26) to *C. elegans* genomic sequence, essentially as previously described (Pan et al., 2005; Pan et al., 2004). In total, 499 cassette type AS events were identified. For each AS event, 3 exon probes and 3 exon junction probes were designed to profile the AS event on the microarray, essentially as previously described (Pan et al., 2004). To facilitate the profiling of unannotated AS events in the *C. elegans* genome, an exon and splice junction tiling array was designed, in which each internal exon in a gene was treated as a potential alternative exon. To maximize the number of new AS events that could be profiled, we included probes for all known or predicted internal exons for ~50% of *C. elegans* genes,
focusing on genes that are preferentially expressed in the nervous system. Using the same exon body and junction probe design as for the 499 EST/cDNA sequence-supported events, we were able to include probes covering 55,759 potential events in 8,649 genes. *C. elegans* genes and exon coordinates were first downloaded from Wormbase, and exon sequences were then extracted and probes were designed as described above.

### 4.3.4 cDNA labeling, microarray hybridization and analysis

C. elegans cDNA samples were labeled with cy3 and cy5 fluors using a ULS labeling kit (Kreatech) according to manufacturer’s recommendations. Each cDNA sample was independently labeled with cy3 or cy5, and used in dye-swap replicate experiments during microarray hybridization. Microarray hybridizations were performed as described previously (Calarco et al., 2009) using an HS 4800 Pro hybridization station (Tecan). After hybridization and washing, microarray slides were scanned using an Agilent microarray scanner. The raw probe intensity values from the microarrays were preprocessed by Agilent Feature Extraction software (version 9.5) using a customized protocol, followed by variance stabilizing normalization (VSN) performed by the vsn package within Bioconductor (version 2.1). Normalized probe intensity values were then analyzed by the Probe Affinity Transcript Abundance (PATA) algorithm (Mavandadi et al. *manuscript in preparation*) to provide quantitative predictions of alternative exon usage across the developmental stages tested. Briefly, PATA takes into account sequence features in probes that contribute to cross-hybridization artifacts. After correcting for these potential biases, the algorithm then estimates the relative abundances of the isoforms including or skipping the alternative exon for each AS event profiled on the microarray. Percent inclusion (%In) values are then calculated as the abundance of the included isoform divided by the sum of the abundances of the included and skipped isoforms.

### 4.3.5 Mapping and analysis of RNA-Seq data

Sequencing reads were mapped to the WS200 version of the *C. elegans* genome using MAQ version 0.7.1 (Mapping and Assembly with Qualities; Li et al., 2008). Only uniquely mapped reads at a quality threshold of 30 or greater were retained. These were further filtered to remove reads that mapped with insertions or had ≥2 mismatches, and had a quality score less than or equal to 30 (i.e. 1 in 1000 mistake). The remaining ‘unmapped’ reads were then analyzed as described below.
4.3.5.1 Identifying reads mapping to splice sites

To map the remaining unmapped reads to splice junctions we created a non-redundant set of sequences 66 nucleotides long corresponding to all possible known and predicted splice junctions (annotated adjacent and non-adjacent exons based on WS200 (Harris et al., 2010) were used). This exon junction database (EJDB) was created by combining 33 nucleotides from the 3’ end of the upstream exon with 33 nucleotides from the 5’ end of the downstream exon. In total we have 548,374 junctions from 20,534 annotated multi-exon genes in Wormbase. The reads were aligned to the EJDB using BLAT (Kent, 2002) and reads were identified as mapping to a junction if there was at least a four nucleotide overlap over either exonic half of the corresponding junction. Reads that had multiple hits to different junctions were eliminated. Junction sequences formed by combining non-adjacent exons and having reads mapping to them uniquely were determined to be alternative splice sites.

We classified AS events as either ‘cassette-type’ events, alternate initiation events, alternate termination events and ‘ambiguous’ based on specific rules. An event is assigned to be a cassette event if there are reads mapping to all three possible junctions and the ratio of the number of reads mapping to the two adjacent junctions is within 30 fold of each other (based on 5% false positive from the ‘true-positive’ data). If one of the adjacent junction count is ‘0’ and there are at least 30 reads mapping to the other adjacent junction, it is classified as an alternate start (if the upstream junction is missing) or stop (if the downstream junction is missing). We also required in every case for the %In values to be at least 1% and less than 99%. If any of these criteria are not met, the event is classified as ‘ambiguous’.

4.3.5.2 Identifying novel splice sites

We first trained our algorithm on WS180 version of the genome, assuming this was the current version and the WS200 version of the genome was the future ‘true’ version. We identified the set of events in WS200 that were not annotated in WS180 and thus considered ‘novel’. Splice junctions involving unannotated splice sites were detected from this set of reads if they mapped in two perfect blocks of genomic sequence separated by at least 25 nucleotides (most introns were at least 25 nucleotides apart) and if they met additional criteria determined from analyzing known splice junctions. Genomic features in these events were used to derive a set of rules required to identify junctions: A. candidate junctions were flanked by canonical
intronic splice site dinucleotide sequences (GT and AG) B. junctions had on average at least 5 supporting mapped reads. C. the ratio of the median number of reads mapping to the adjacent intron to the median number of reads mapping to the exon was < 0.1 (this indicates a positional change from an intron to an exon).

Reads that did not map to the genome or to ‘EJDB’ were mapped back to the WS200 version of the genome using BLAT. Here we accept reads to have positively mapped to the genome if they map uniquely to the genome in exactly two blocks and each of the blocks have at least 5 nucleotides perfectly matching the region. We then applied the rules learnt above to identify reads that span two distinct regions of a transcript.

4.3.6 Defining a set of annotated transcript variants in *C. elegans*

To define a “true positive” set of alternative mRNA processing events in *C. elegans* we first obtained all transcripts with EST/cDNA evidence from WormMart (Harris et al., 2010). We then used an algorithm based on the methodology and rules outlined in the *Drosophila melanogaster* Exon Database (Lee et al., 2004) to identify all annotated splice variants and isoforms. Based on these criteria, we placed the annotated events into seven categories: Exon skipping (cassette type), alternative 5’ splice site usage, alternative 3’ splice site usage, intron retention, alternative transcript start site usage, alternative terminal exon usage and mutually exclusive exons.

4.3.7 Filtering array events for analysis

By including all possible exon junctions from the 55,759 array events in the ‘EJDB’, we were able to identify the set of array events that had reads corresponding to the alternative junction. We used the intensities of the control probes on the array to determine the background probe intensity distribution at each stage. By setting a 5% false positive probe inclusion rate, we filtered events where the exons or junction probe intensity was less than this. We determined the expression level as the average of the intensities of the two constitutive exons. Once again, events were eliminated from a stage if this expression intensity was less than the determined background cutoff. We finally filtered out events that did not have reads mapping to all three junctions. Since the array was designed specifically for identifying cassette type AS events, we had to eliminate potential alternative starts and stops from the analysis. Since alternative starts
and stops have an extreme skew in the ratio of the two adjacent junctions (since one of them is used while the other is not), we determined the ratios of the two adjacent junctions in the cassette exons in the true positive set and applied a cutoff based on 5% false positive rate. We also determined the distribution of %In values based on RNA-Seq for these true-positive cassette exons and eliminated events in the array that had %In values that were below the lower 5% or higher than 95% of the events. By applying these filters we identified 704 events for further analysis. To identify temporally regulated events from the array, we compared the %In values of the events that were present in more than one stage and found that 624 of the 704 events were in at least 2 stages. Additionally, from the complete set of junctions mapped, we identified the set of exons that were part of at least 2 different junctions i.e. exons that were multiple donor splice sites or multiple acceptor splice sites.

4.3.8 Computational analysis of alternative exon properties

4.3.8.1 RNA-Seq percent inclusion value calculation

Percent inclusion values were calculated as the ratio of the average number of reads across the two adjacent junctions over the average of the number of reads across the adjacent and alternative junction.

\[
\text{%In} = \frac{(c1 \cdot a + a \cdot c2)/2}{c1 \cdot c2 + (c1 \cdot a + a \cdot c2)/2}
\]

where, c1 and c2 refer to the two constitutive exons and a refers to the alternative exon, c1.a and a.c2 are the number of reads mapping to the two adjacent junction and c1.c2 is the number of reads mapping to the alternative junctions.

4.3.8.2 Frame preservation

Exon lengths were calculated and grouped in one of three categories (3n, 3n+1, or 3n+2) based on whether their lengths are perfect multiples of 3. Distribution of the three possible outcomes in each of the seven alternative splice categories determined above and in the novel alternative splice events was compared to the background evaluation of the three categories based on all the exons annotated in the WS200 version of the genome.
4.3.8.3 Sequence Conservation

Exon sequences were obtained from WormMart (WS200 version of the genome) and were aligned to the C. Briggsae genome (version WS200) using BLAT. For each exon, maximum percent identity in sequence aligned to the C. briggsae genome was calculated. This was compared against the average percent identity of the exons undergoing cassette type AS, either in the true-positive set or from the novel events.

4.3.9 Estimates of frequency of AS and developmentally-regulated AS

To estimate the frequency of AS in C. elegans, we first identified the number of genes possessing novel AS events involving annotated junctions (1354 genes with 2183 events) that belonged to either the cassette, alternative initiation, or alternative termination classes. From this subset we extrapolated to estimate all genes with AS (Total_AS_Genes) as:

\[
\text{Total}_{-}^{\text{AS}}_{-}\text{Genes} = \frac{\text{NG}/(\text{TP}_{\text{subsetID}}/\text{TP}_{\text{tot}})) + (\text{TP}_{\text{subset}})}{\text{TP}_{\text{subsetID}}/\text{TP}_{\text{tot}}}
\]

where, ‘NG’ is the 1,354 genes with 2,183 novel AS events (Fig 2A), ‘TP_{tot}’ is the total genes in the ‘true-positive’ set (2,065). While ‘TP_{subset}’ is the set of ‘true-positive’ genes having either a cassette, alternative initiation or alternative termination events (1,015), ‘TP_{subsetID}’ is the subset of these (815) that we positively identify with our RNA-Seq data. Similarly we calculated the total AS events in the genome, by replacing gene numbers with AS event numbers.

To estimate the frequency of AS events that are temporally regulated, we averaged the fraction of events identified to change by at least 20% from our microarray or have a pvalue of < 0.01 from our RNA-Seq data between any two stages. We find that 38% of the array events are temporally regulated (275/624 events corrected for false positive rate determined from RT-PCR). Of the 1,573 AS events with sufficient read depth we identify 22% (349) with a pvalue < 0.01.

4.3.10 Motif analysis

We used SeedSearcher (http://www.psi.toronto.edu/~weijun/snap/tools.php), a hypergeometric approach for identifying putative binding sites. The algorithm compares a set of input sequences against a set of background sequences to identify statistically enriched motifs of specified length (pentamers in our case). We used 50-nucleotide intron sequences upstream (or
downstream) of temporal-regulated exons and compared these with the set of 50-nucleotide upstream (or downstream) intron sequence from all identified AS exons. This analysis therefore identifies motifs that are enriched specifically in temporally-regulated events rather than generic AS events.

4.3.11 Database setup

The *C. elegans* Splice Browser database (http://splicebrowse.ccbr.utoronto.ca) was created using Gbrowse ((Donlin, 2009; Stein et al., 2002) http://gmod.org) with a mySQL backend. The database was specifically designed to be similar to Wormbase (the widely popular *C. elegans* web resource) and utilized data mirrored from Wormbase. Sequence and array data were converted to a format compatible with gbrowse and the gbrowse admin tutorial was followed extensively (http://gmod.svn.sourceforge.net/viewvc/gmod/Generic-Genome-Browser/trunk/htdocs/tutorial/tutorial.html).

4.3.12 RT-PCR assays

Non-radioactive RT-PCR assays were performed with the OneStep kit (QIAGEN) as recommended by the manufacturer. Radioactive RT-PCR assays were carried out essentially as described previously (Calarco et al., 2009).

4.3.13 Plasmid and fosmid constructs and microinjection

4.3.13.1 Bichromatic reporters

In order to generate a bichromatic splicing reporter for use in *C. elegans*, we adapted the previously described RG6 bichromatic reporter plasmid (a kind gift from Tom Cooper, Baylor College of Medicine; (Orengo et al., 2006)). First, XhoI and NotI restriction sites were inserted downstream of the sequences encoding the FLAG epitope and nuclear localization signal by a PCR-based approach. *C. elegans* genomic DNA was used as a template in PCR reactions to amplify the genomic region spanning exons 3, 4 and 5 and all intervening intronic sequences from the *unc-32* locus. A single nucleotide insertion in alternative exon 4B was introduced by an overlapping PCR based strategy. This ‘splicing cassette’ was then inserted into the plasmid using the XhoI and NotI restriction sites. The dsRED open reading frame was then replaced with the EGFP open reading frame that was inserted using NotI and SacII restriction sites. The mRFP open reading frame was then inserted downstream of the EGFP sequence using SacII and Apal
sites. The resulting plasmid was then used as a template to amplify a PCR product with the following primers: 5´-GGGGACAAGTTTGTACAAAAAAGCAGGCTACCATGGATTAC-AAGGATGACGATCAAGGGGTACCTG-3´ and 5´-GGGGACCACCCTTT-GTACAGGAAAGCTTGTTACTTGACAGCGTCCATGCGCGCCGCTGA-3´. This PCR product was then recombined into the Gateway pDONR221 plasmid using the Gateway BP clonase kit (Invitrogen) as recommended by the manufacturer. The resulting vector (pWASR1) served as a template for the creation of all other bichromatic reporters. For all other reporters, genomic regions of interest were amplified from genomic DNA by PCR and inserted into pWASR1 using the XhoI and NotI sites.

In order to drive expression of the bichromatic reporters in neuronal or muscle cell types, the reporter cassettes were recombined into destination vectors containing either a pan-neuronal promoter (Prgef-1: pJH906) or body wall muscle promoter (Pmyo-3: pJH940) and the unc-54 3´UTR using a Gateway LR clonase kit (Invitrogen) according to manufacturer’s recommendations.

4.3.13.2 Generation of isoform-specific recombineered fosmids

Isoform-specific rescuing fosmids were generated by adapting a previously described recombineering protocol (Tursun et al., 2009). Regions of fosmid DNA spanning alternative splicing events (encompassing alternative exons, flanking introns, and neighbouring constitutively spliced exons) were replaced with cDNA sequences derived from spliced mRNA transcripts that either include or exclude the alternative exon. In this approach, the 34 base pair ‘scar’ corresponding to the FRT site remaining after the recombineering protocol is contained within a constitutively spliced intron sequence. This strategy therefore allows for the seamless joining of neighbouring exons in all mature mRNA transcripts.

4.3.14 Microinjection and generation of integrated lines

Animals were transformed with reporter plasmids or fosmids using a standard microinjection approach (Kadandale et al., 2009). All fluorescent protein reporter plasmids and fosmids were co-injected with a plasmid expressing the LIN-15 gene product (each at 20-50 ng/uL) into lin-15(n765ts) mutant animals. Isoform-specific rescuing fosmids were co-injected with a plasmid expressing mRFP under a myo-2 promoter (pCFJ90; a kind gift from Dr. Eric
Jorgensen, University of Utah) into specific mutant strains. Integration of extrachromosomal bichromatic reporter arrays was achieved by irradiating animals with UV light and then selecting individuals that stably integrated the array.

4.3.15 Bichromatic reporter imaging

Transgenic animals were placed on slides with air-dried 2% agarose pads on their surface and spotted with a drop of M9 medium, and then fixed in place with a cover slip. Animals were then immediately imaged by using a Leica DMI6000B confocal microscope.

4.3.16 Synaptic development and behavioral assays

The morphology and quantity of dorsal nerve cord synapses were evaluated in L4 stage larval strains as described previously (Zhen and Jin, 1999). Serotonin-induced paralysis assays were also performed as described previously (Ranganathan et al., 2000). To assay egg laying, late stage L4 animals were transferred to fresh plates and allowed to develop for another 22-24 hours at 20 degrees Celsius. The resulting adult animals were then individually placed in a well of a 96 well plate containing 100 uL of low salt M9 buffer (5.8 g Na\textsubscript{2}HPO\textsubscript{4}, 3 g KH\textsubscript{2}PO\textsubscript{4}, 0.5 g NaCl, 1 g NH\textsubscript{4}Cl per litre of water) containing 5 mg/mL serotonin (Sigma Aldrich) or 5mg/mL serotonin and 6 mg/mL dopamine (Sigma Aldrich). The number of eggs laid after one hour of incubation with the different neurotransmitters was counted. All egg laying assays were performed at room temperature.
4.4 Results

4.4.1 Large-scale identification of splice variants in *C. elegans* by RNA-Seq and microarray profiling

We used two distinct but complementary approaches to identify and quantify splice variant usage in *C. elegans* (Figure 4.1). In the first approach, we performed RNA-Seq of polyA+ RNA and identified splice junctions using a computational approach that does not require prior knowledge of splice junction connectivity (Figure 4.1 A). In a parallel approach, we developed a custom microarray and associated computational tool in which combinations of exon body and splice junction probes were used to monitor single exon skipping or ‘cassette’ type splicing events (Figure 4.1 B; see Materials and Methods). Our overall strategy, described in more detail below, was to use RNA-Seq profiling to identify and measure as comprehensively as possible alternative splice site usage in the wild-type transcriptome across development, and then to cross-validate and augment this dataset with information from quantitative AS microarray profiling data.

Our RNA-Seq dataset was generated from polyA+ RNA isolated either from mixed populations of wild type hermaphrodite animals of various developmental stages, or from synchronized L4 or adult stage animals (Ramani et al., 2009). These samples yielded 130 million, 54 million, and 69 million reads respectively, over 80% of which could be aligned to the genome. To maximize coverage, we incorporated recently published RNA-Seq datasets from four developmental stages (Hillier et al., 2009; Figure 4.1 C). In total, we analyzed ~540 million reads, of which ~332 million alignable reads (62%) could be mapped to 19,809 of the 20,534 genes (96.5%) annotated in Wormbase (Harris et al., 2010). In contrast, less than 1.5% of the reads mapped to intergenic regions. The extent of coverage and specific enrichment of reads mapping to annotated genes demonstrates that the RNA-Seq datasets in this study afford a high degree of sensitivity and specificity in the detection of expressed genes.

To generate our AS microarray dataset, we hybridized cDNAs derived from polyA+ RNA isolated from synchronized wild type hermaphrodite animals harvested at five developmental stages (larval stages L1 to L4 and young adults). Alternative exon inclusion level measurements were estimated by using a new method capable of accurately modeling the
Figure 4.1: Outline of RNA-seq and microarray approaches for profiling AS in *C. elegans*.

(A) For RNA-seq analysis, short sequence reads from the transcriptomes of animals isolated at multiple or specific developmental stages are first aligned to the genome and those that uniquely map in their entirety are used to identify exonic regions of transcripts. The remaining unmapped reads are then mapped against an exon junction database containing all possible combinations of splice junctions between annotated exons to identify and quantitatively measure AS events. Finally, any remaining unmapped reads are aligned to novel splice junctions that are not annotated or predicted. (B) Our microarray platform uses exon body (labeled C1, A, and C2 in diagram) and exon junction (labeled C1:A, A:C2, and C1:C2) probes to monitor cassette-type AS events, where the internal exon in a triplet of exons can be skipped in spliced mRNAs. We have included probes to monitor a total of 55,759 exon triplets in 8649 genes (~50% of annotated genes), in addition to a set of 499 previously annotated cassette AS events. Probe sets corresponding to exon triplets with evidence of AS by our RNA-seq analysis are then combined with the PATA algorithm to generate quantitative predictions of relative isoform usage across the developmental stages analyzed. (C) Number of sequence reads uniquely mapping to the *C. elegans* genome by stage and sample. The values in parentheses indicate the total number of reads generated for each sample.
behavior of splice junction probes (see Materials and Methods; Mavandadi et al. *manuscript in preparation*). Of 8,649 microarray profiled genes 7,031 (81%) were detected above background in one or more of the developmental stages, suggesting that a large proportion of genes can also be profiled using this complementary method.

To identify splice junctions with our RNA-Seq data, we created an exon junction database (EJDB) by joining pair-wise combinations of predicted exons from all annotated multi-exon genes from Wormbase (version WS200). We also created a control exon junction database by joining randomly shuffled pairs of exons from the EJDB. We identified reads mapping to 104,246 of the 126,628 adjacent junctions (~82%) in 18,555 (90%) annotated genes (see Materials and Methods). In contrast, at all depths of coverage surveyed, only a small fraction of reads could be mapped to our control junction database (~2% relative to the annotated adjacent junctions detected; Figure A3.1, Appendix 3). These results indicate that our splice junction mapping procedure also achieves high specificity and sensitivity.

To assess the quality of our RNA-Seq data, we examined how well we could identify a set of previously annotated splice junctions. A total of 4,049 non-adjacent splice events are annotated in Wormbase and 2,842 (~71%) of these were detected with at least one supporting read and 2,257 (~56%) junctions were detected by five or more reads (Figure 4.2 A; see Materials and Methods). At read depths of greater than 200 million reads, the number of detected adjacent and non-adjacent junctions did not substantially increase (Figure A3.1, Appendix 3), suggesting that at our current read depth we have detected the vast majority of AS events in the samples we have analyzed. It is likely that many of the undetected AS events occur in other developmental states (such as males or dauers) or in response to other environmental conditions not profiled in this study. We note that ~20% of the known AS events that we detected have only a single sequence read spanning the alternatively spliced junctions. However, >90% of single read AS events can be confirmed as coverage is increased (Figure A3.1, Appendix 3). These results provide evidence that the majority of single read AS events reported here are true events. We therefore report these AS events, and all other single-read junction data in the sections to follow.

We next cross-validated our RNA-Seq data with data generated by AS microarray profiling. Our microarray analysis provides quantitative predictions of the degree to which an
alternative exon is included in a mature mRNA transcript, referred to as a percent inclusion (%In) value. %In values can also be generated from the RNA-Seq data using ratios of reads mapping to splice junctions formed by exon inclusion and skipping. Focusing on a set of 317 annotated cassette AS events from the true positive (cDNA/EST-validated) set that can be monitored both by RNA-Seq analysis and AS microarray profiling, we observe a strong correlation ($r^2 = 0.85$) between the %In values measured by these two technologies (Figure 4.2 B; see Methods).

Collectively, our results demonstrate that the majority of previously annotated AS events in the *C. elegans* transcriptome are present in our RNA-Seq data and that RNA-Seq yields accurate measures of relative splice site usage.
### Table

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Detected Events</th>
<th>Novel Events</th>
<th>Annotated</th>
<th>Detected</th>
<th>Novel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>&gt;5 reads</td>
<td>&gt;10 reads</td>
<td>All</td>
<td>&gt;5 reads</td>
</tr>
<tr>
<td>Cassette Exons</td>
<td>754 (57%)</td>
<td>448 (60%)</td>
<td>409 (55%)</td>
<td>1,759</td>
<td>479</td>
</tr>
<tr>
<td>Alternate Donor</td>
<td>682 (92%)</td>
<td>251 (37%)</td>
<td>190 (28%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alternate Acceptor</td>
<td>883 (92%)</td>
<td>482 (55%)</td>
<td>429 (49%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alternate Termination</td>
<td>226 (70%)</td>
<td>132 (59%)</td>
<td>114 (51%)</td>
<td>159</td>
<td>96</td>
</tr>
<tr>
<td>Alternate Initiation</td>
<td>730 (72%)</td>
<td>375 (52%)</td>
<td>291 (40%)</td>
<td>255</td>
<td>184</td>
</tr>
<tr>
<td>Intron Retention</td>
<td>719 (93%)</td>
<td>539 (75%)</td>
<td>468 (65%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mutually Exclusive</td>
<td>55 (95%)</td>
<td>49 (89%)</td>
<td>46 (84%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Graph B

- **Sequence inclusion (%)**
- **PATA inclusion**
- \( r^2 = 0.85 \)

### Graph C

- **erp-1 (5-6-7)**
- **rbf-1 (10-11-12)**
- **daf-4 (5-6-7)**
- **exc-7 (2-3-4)**
- **unc-10 (20-21-22)**
- **pdr-1 (3-4-5)**
Figure 4.2: Identification of known and novel AS events

(A) Table displaying the proportion of different classes of AS events annotated in WormBase that are identified by our RNA-seq analysis with varying degrees of junction count support. Also listed in the last three columns are the number of unannotated AS events identified by our analysis and the number of reads supporting these splice junctions. (B) We compared the %In values for a set of 317 annotated cassette AS events from the true positive set that were detected both by RNA-seq analysis and by our microarray platform. For these AS events, RNA-seq derived %In values profiled across four developmental stages (L2, L3, L4, and adults) were compared to the PATA derived %In values in the same stage. The plot shows a high correlation between the two platforms indicating that both methods are reliable estimates of relative isoform usage. (C) RT-PCR validation of a subset of unannotated AS events in the larval (L2) and adult developmental stages. Primers were designed to anneal to sequences corresponding to neighboring constitutively spliced exons, amplifying two products: one representing isoforms including the alternative exon in transcripts (top bands in gel images) and the other one skipping the alternative exon (bottom bands).

4.4.2 Identification of novel splicing events

We also identified novel junctions in our RNA-Seq analysis. These were derived from two classes of splicing events: (1) AS events involving annotated splice sites (novel-class A events), in which the exons and splice sites are correctly annotated, but the splicing event joining the exons has not been previously detected, and (2) splicing events involving unannotated splice sites (novel-class B events), in which new junctions are created from a predicted or novel exon, or from an unannotated internal splice site.

Our analysis detected 2,183 novel-class A events from 1,354 genes. Out of these events, 1,769 could be classified as putative cassette AS events, while the remaining 414 were grouped as potential alternative initiation or termination events (see Materials and Methods; Figure 4.2 A). We performed semi-quantitative RT-PCR validations for 13 of the RNA-Seq predicted novel AS events in L2 and adult RNA samples, and found that 12 (92%) events had detectable expression of both isoforms (Figures 4.2 C and Figure A3.2, Appendix 3). A large proportion of the previously unannotated AS events also share key characteristics of known AS events, further supporting the conclusion that they are truly expressed. For example, annotated cassette AS events are significantly enriched for frame-preserving exons relative to the total set of annotated exons (68% vs. 40%, respectively, p = 1x10^{-4}, \chi^2 test). Similarly, our set of novel cassette AS events are also enriched for frame preserving exons (54%, p = 4x10^{-3}, \chi^2 test). In addition, the level of conservation of both annotated and novel exons affected by AS is markedly higher between C. elegans and C. briggsae relative to the total set of annotated exons (64% for
annotated and 65% for novel AS events, compared with 55% for all exons, \( p < 0.05, \chi^2 \) test). These analyses, together with our RT-PCR validation experiments strongly suggest that the majority of the previously unannotated splicing events are *bona fide* AS events.

We further identified 6,468 putative novel-class B events by analyzing reads that had not been mapped to either the genome, annotated splice junctions, or to the novel-typeA events (Materials and Methods). Of the 6,468 junctions, ~17% connect two novel splice sites while the remaining events involved junctions combining both novel and annotated splice sites. Of 25 putative novel splicing events randomly selected, all were validated by RT-PCR assays (Supporting data file SD 4.1, CD). These results indicate that our method for detecting novel splicing events is specific. We further analyzed each of the 6,468 novel splice junctions for evidence of alternative usage of one of their two splice sites, and found that 256 of these novel junctions represent potentially novel alternative splicing events.

In summary, our RNA-Seq analysis has identified 8,651 putative novel splicing events in more than 2,682 genes. Based on our analyses of known and novel junctions, we estimate that transcripts from more than 5,000 genes (~25% of all protein coding genes) of the *C. elegans* genome undergo AS, with each gene on average generating 2.2 isoforms (see Materials and Methods). This estimate represents a 2.5 fold increase over what is suggested by EST/cDNA evidence, and a two-fold increase over previous estimates (Zahler, 2005).

### 4.4.3 Genome-wide analysis of temporally regulated alternative exon usage

We next determined whether any of the known or novel alternative splice junctions were differentially-regulated during development. Notably, from the RNA-Seq dataset, we detect 349 AS events that are predicted to have significant differences in junction usage between two or more developmental stages (\( p < 0.01, \) Fisher’s exact test). These developmentally-regulated junctions involved all classes of profiled AS events, and significant quantitative differences were detected between every pair of developmental stages compared (Supporting data file SD 4.2, CD). We performed semi-quantitative RT-PCR assays to validate a subset of the RNA-Seq derived predictions. Of seven events examined, all were experimentally validated (Figures 4.3 A and 4.4). Additionally, of 80 events that could also be monitored by our microarray, 61 (77%)
Figure 4.3: Identification of temporally regulated AS events.

(A) Semiquantitative RT-PCR validation for a subset of regulated AS events across larvae (L1, L2, L3, and L4) to adult (Ad) stages. Primers were designed to amplify both isoforms, and in each case, the two possible isoforms (top bands, included isoform; bottom bands, excluded isoform) show changes in relative intensities across developmental stages. (B) AS exons that were identified to undergo temporal regulation across development of at least 20% were median centered and hierarchically clustered and visualized using Cluster and TreeView (Eisen et al., 1998). Blue boxes indicate higher exon inclusion, while yellow boxes indicate higher exon exclusion. We find exons that are excluded in specific stages (projected panels to the left) while we also notice shared patterns of relative exon inclusion between multiple stages for other subsets of AS events (projected panels to the right). (C) SeedSearcher was used to identify pentamer motifs that are statistically enriched in either the 50 nt upstream of or downstream from splice junctions that are differentially regulated during development. Examples of these enriched motifs are shown either for motifs located upstream of the AS exon (left) or downstream from the AS exon (right). Note that four of these motifs correspond to the published motifs of HRP-2, SUP-12, FOX-1 and ASD-2, and these are indicated.
were found to display a %In value difference of 20% or greater between the corresponding stages predicted to undergo differential junction usage by our RNA-Seq analysis.

Although the data generated from our RNA-Seq analysis are reliable, a number of developmentally-regulated AS events are likely to be missed due to an insufficient depth of coverage. However, our microarray platform is sufficiently sensitive to detect some of these missing events. Indeed, 214 additional AS events that lacked sufficient read coverage for reliable quantification were predicted by our microarray to undergo %In value differences of 20% or higher between two or more developmental stages (Figure 4.3 B and Supporting data file SD 4.3, CD). Out of 21 candidates tested from this set of events, 18 (86%) were confirmed to exhibit significant changes by RT-PCR assays (Figures 4.3 A and 4.4; data not shown). These results suggest that our microarray predictions are accurate and can identify additional regulated AS events that were not detected in our RNA-Seq analysis. Intriguingly, the alternative exons and flanking intron sequences of the developmentally-regulated splicing events detected in our RNA-
Seq and microarray analysis are significantly more conserved between *C. elegans* and *C. briggsae* than are other AS events (68% p < 0.025 $\chi^2$ test), suggesting that the temporal regulation of this class of AS events is likely to be crucial for normal development.

In summary, the combined analysis of the RNA-Seq and splicing microarray datasets revealed 574 temporally regulated AS events in 394 genes. When considering the number of events that can be accurately quantified by our methods, and extrapolating to all AS events, we estimate that roughly 30% of AS events are differentially regulated between the *C. elegans* developmental stages we have profiled, consistent with previous estimates (Barberan-Soler and Zahler, 2008). To our knowledge, this study provides the most comprehensive dataset and profiling information for developmentally-regulated AS events to date.

Alternative splicing events are regulated by the interplay between trans-acting splicing regulators and cis-regulatory elements associated with alternative exons and flanking intron and exon sequences (Blencowe, 2006; Matlin et al., 2005; Wang and Burge, 2008). Identifying these cis-elements and the factors that associate with them is an important step towards better understanding the code governing AS regulation. As a first step towards this goal, we used the SeedSearcher algorithm (http://www.psi.toronto.edu/snap/) to identify motifs that are enriched in the regulated exons or intron sequences flanking these exons among the set of temporally-regulated AS events (Materials and Methods). This search identified 12 pentamer motifs enriched in the upstream intronic sequence and 19 pentamer motifs enriched in the downstream intronic region of the temporally-regulated AS events (FDR corrected p-values of <0.01; Figures 4.3 C; Figures A3.3 and A3.4, Appendix 3). Twelve of the 19 (65%) downstream motifs and 7 of the 12 (60%) upstream motifs were previously identified as being enriched in intronic regions flanking AS events by an independent approach (Kabat et al., 2006). The set of motifs we have identified include perfect matches to motifs previously shown to be regulated by the splicing factors ASD-1/FOX-1, ASD-2, SUP-12, or HRP-2 (Figure 4.3 C; (Kabat et al., 2009; Kuroyanagi et al., 2006; Kuroyanagi et al., 2007; Ohno et al., 2008)). Of these factors, only ASD-2 was previously demonstrated to play a role in modulating temporally-regulated AS (Ohno et al., 2008). Our results thus suggest that these factors (and as yet uncharacterized RNA binding proteins) and their cognate motifs identified here, play important roles in differential splicing during development.
4.4.4 The *C. elegans* Splice Browser

In order to make our results easily accessible to the research community, we have created the “*C. elegans* Splice Browser” (http://splicebrowse.ccbr.utoronto.ca). This web resource includes relevant information from the AS analysis generated in this study. For each of the gene predictions, read counts supporting all detected adjacent, alternative, and novel splice junctions are included. These read counts can be displayed either by developmental stage of interest or as a cumulative count from all stages. Where applicable, microarray-derived %In value measurements are represented as color-coded heat maps that indicate splicing changes throughout development. Users can also upload personal data and create customized tracks to compare their data with the current dataset. In this regard, the Splice Browser not only serves as a reference database but can also act as a repository for future datasets as they become available. With these features, the Splice Browser should provide biologists with easily accessible, quantitative information regarding alternative splice site usage during *C. elegans* development.

The data described above form a catalog of AS events which can be accessed easily via the Splice Browser. Importantly, this catalog of events can also serve as a platform for driving detailed high-resolution experiments to address the precise spatio-temporal regulation of individual splice events and to test the functional significance of individual isoforms. In the remainder of this chapter, I present experimental approaches to explore these latter two research areas and illustrate these methods with examples.

4.4.5 Versatile bichromatic reporters reveal cell type-specific AS patterns

Analyses of our microarray and RNA-Seq datasets have identified developmentally-regulated AS events. However, it is currently impossible to predict cell type-specific splicing events from these datasets alone. The recent development of bichromatic reporters to monitor AS in *C. elegans* and other species has enabled the *in vivo* detection of condition-dependent, cell type and developmental stage-specific splicing events (Kuroyanagi et al., 2006; Kuroyanagi et al., 2007; Ohno et al., 2008; Orengo et al., 2006; Stoilov et al., 2008). These reporters drive expression of green or red fluorescent proteins depending on the reading frame retained after inclusion or exclusion of an upstream alternative exon. The relative fluorescence intensity in
single cells is then used as an index for the relative usage of the alternative exon in each spliced mRNA.

To explore splicing regulatory patterns of AS events identified in our genome-wide analyses in vivo, I have adapted a versatile bichromatic reporter that can monitor all classes of AS events (Orengo et al., 2006; Figure 4.5 A). This reporter was engineered such that the inclusion or exclusion of an alternative exon switches between the usage of two tandemly positioned open reading frames encoding either enhanced GFP (EGFP) or monomeric RFP (mRFP). As a proof of principle experiment, I examined the splicing patterns of transcripts derived from the unc-32 gene, which encodes the A subunit of a vacuolar ATPase (Pujol et al., 2001). Transcripts encoded by unc-32 contain one of three mutually exclusive fourth exons generated by AS, and exon 4B containing transcripts are specifically expressed and functionally required in neurons (Pujol et al., 2001). I monitored unc-32 exon 4 splicing by inserting the genomic DNA fragment spanning exons three to five of unc-32 into the splicing reporter, such that the inclusion of exon 4B would lead to translation of mRFP, while the inclusion of either exons 4A or 4C would lead to production of EGFP. As predicted, when the splicing reporter was expressed pan-neuronally using the P_{region-1} promoter, a strong mRFP signal was detected in all cells, while only a few cells also displayed weak EGFP (Figure 4.5 B). When the reporter was expressed in muscle cells using the P_{myo-3} promoter, the opposite pattern was observed, with strong EGFP, and little to no detectable mRFP fluorescence. To confirm that the observed fluorescent signals reflect cell type-specific AS differences, total RNA was harvested from these transgenic animals, and reporter-derived transcripts were monitored by RT-PCR. Indeed, exon 4B was predominantly included in transcripts expressed pan-neuronally, whereas exon 4A and 4C containing transcripts represented the major isoforms in muscle cells (Figure 4.5 C). Therefore, the bichromatic reporter system accurately recapitulates previously characterized cell type-specific AS events.

I next utilized the bichromatic reporter to examine whether extensive differential AS can be observed between cells of a given tissue type. I focused on the nervous system because I speculated that differential splicing might contribute to increased molecular and functional diversity between diverse classes of neurons (Li et al., 2007; Lipscombe, 2005). Bichromatic reporters were generated for thirteen cassette-type AS events that were selected because they were found in our genome-wide analyses to be developmentally-regulated, or if their
Figure 4.5: Bichromatic reporters reveal cell type-specific AS

(A) Schematic of the bichromatic reporter system. The DNA sequence spanning the EGFP coding sequence can be translated in two alternative reading frames: one encoding EGFP and the other encoding a non-fluorescent amino acid sequence. The mRFP coding sequence is fused in frame with this alternative non-EGFP reading frame. An alternative splicing event of interest containing the alternative exon (labeled A), flanking exons (labeled C1 and C2) and introns is inserted between the translation start site (AUG in diagram) and the fluorescent protein reading frames such that exclusion of the alternative exon leads to translation of EGFP, while inclusion of the exon alters the reading frame and leads to mRFP translation. The reporter also contains an N-terminal nuclear localization signal (NLS) and FLAG epitope to facilitate in vivo and in vitro detection.

(B) Fluorescent images of animals expressing a bichromatic reporter monitoring unc-32 exon 4B inclusion (mRFP; red signal, top panels) and exon 4A or 4C inclusion (EGFP; green signal, middle panels) in neurons (left panels) or muscle cells (right panels). Merged images (with DIC) are displayed in the bottom panels. Arrows are pointing to neuronal nuclei with prominent RFP signal (left top panel) and muscle nuclei with prominent EGFP signal (right middle panel). Scale bar: 25 um.

(C) RT-PCR assays detecting relative usage of unc-32 exon 4 variants from transgenic animals expressing bichromatic reporter-driven transcripts in neurons or muscle cells.

(D) Fluorescent images of animals expressing a bichromatic reporter monitoring unc-16 exon 14 inclusion (mRFP signal) and exclusion (EGFP signal) in neurons. Arrows are pointing to two neuronal nuclei expressing both included and skipped isoforms. Scale bar: 25 um.
corresponding genes are annotated in Wormbase to be expressed in the nervous system (Supporting data file SD 4.4, CD). Additional features such as whether or not the alternative exon was frame-preserving and conserved among nematode species were also considered in our selection of events. The $P_{rgef-1}$ pan-neuronal promoter was used to drive the expression of these reporters to analyze the splicing pattern of each event in diverse classes of neurons.

Six of the 13 AS events tested displayed striking differences in relative EGFP to mRFP fluorescence intensities between specific neuronal subtypes (Figures 4.5 D and 4.6). The neuronal subtype-specific AS pattern is unique for each of the six reporters, indicating that these differences cannot be simply attributed to the exogenous promoter used to drive their expression (Figure 4.6). Moreover, the observed regulatory pattern was consistent among transgenic animals expressing the same reporter, indicating that the differential AS patterns are not due to stochastic changes, but rather determined by consistent regulatory effects associated with specific neuronal subtypes. For example, skipping of exon 14 from the $unc-16$ splicing reporters was reproducibly detected in GABAergic motor neurons (Figure 4.6 D and 4.7). Cell type-specific AS regulation of several other genes was also frequently observed in the central nervous system (CNS), where diverse classes of sensory, inter- and motor neurons reside (Figure 4.6). A potential caveat to these experiments is that the use of the $P_{rgef-1}$ promoter produces ectopic expression of the reporters in neurons that may not normally express some of these transcripts and do not reflect truly endogenous splicing patterns. However, expressing the reporters pan-neuronally has offered the advantage of identifying differences in the splicing regulatory machinery between individual cells of the nervous system. I view these reporters as a starting point towards more detailed experiments aimed at identifying endogenous regulatory patterns and functions of the AS events of interest. Collectively, our results support the conclusion that the splicing machinery is differentially controlled at the level of single neuronal subtypes, revealing a high degree of AS regulatory complexity in the nervous system.
Figure 4.6: Bichromatic reporters identify complex AS patterns in the nervous system

Fluorescence microscopy images of animals expressing bichromatic reporters in the nervous system monitoring AS events in the *unc-16, dyn-1, cle-1,* and *mod-1* genes. In all cases examined, inclusion of the alternative exon leads to mRFP translation (red signal) and skipping of the exon leads to EGFP translation (green signal). Examples of differential splicing between individual CNS neurons and ventral cord neurons can be clearly observed. Scale bar represents 25 um for all images.
Figure 4.7: *unc-16* exon 14 is excluded in GABAergic motor neurons

(A) Fluorescent images of an animal expressing a bichromatic reporter monitoring *unc-16* exon 14 inclusion (mRFP; red signal, top panel) and exclusion (EGFP; green signal, middle panel) in neurons (left panels). Merged images are displayed in the bottom panel. Arrows are pointing to neuronal nuclei expressing both included and skipped isoforms. Scale bar: 25 um.

(B) Fluorescent images of an animal expressing both the bichromatic reporter monitoring *unc-16* exon 14 AS as in A, and a reporter expressing GFP in GABAergic motor neurons (*P*_{unc-25::GFP}, green signal, middle panel). Merged images are displayed in the bottom panel. Arrows are pointing to cell bodies of GABAergic motor neurons and overlap completely with nuclei expressing both included and skipped isoforms in (A). Scale bar: 25 um.
4.4.6 Functional investigations of splice variants using isoform-specific transgenes

Once splice variants of interest have been confirmed, it is important to determine whether they have overlapping or unique functions. To this end, I developed a fosmid-based isoform-specific transgene rescue approach. Using recombineering (Tursun et al., 2009), fosmids containing a gene with an AS event of interest are engineered such that only one splice variant can be expressed (Materials and Methods). In this strategy, regions of the fosmid overlapping the AS event are replaced with cDNA sequences from spliced mRNAs that either include or skip an alternative exon of interest. These isoform-specific fosmids are then transformed either individually or in combination into animals harboring strong loss of function mutations for the corresponding gene of interest. As a proof of principle, I applied this system to initially analyze two validated AS events that also exhibited neuron-specific splicing patterns as determined by our bichromatic reporter experiments.

First, I performed an isoform-specific analysis of the cle-1 gene, which encodes multiple isoforms of a type XVIII collagen protein that is a component of basement membranes and enriched at neuromuscular junctions (Ackley et al., 2001; Ackley et al., 2003). Loss of function mutations in this gene lead to a number of phenotypes, including defects in morphology and number of GABAergic neuromuscular junctions that were visualized and quantified by julsl, a GFP-tagged synaptobrevin (SNB-1) vesicle reporter expressed in GABAergic motor neurons ((Ackley et al., 2003); Figure 4.8). I identified that splicing of alternative exon 18 was differentially regulated both during development and in specific neuronal subtypes (Figures 4.6 A and 4.8). The selective expression of either isoform introduced into transgenic animals was confirmed by RT-PCR assays (Figure 4.8 B). Intriguingly, expression of either isoform alone improved the morphology of GABAergic synapses, but only led to a partial rescue of the number of synaptic puncta (Figure 4.8 C and D). Only when both isoforms were co-expressed, both the morphology and number of GABAergic synapses were fully restored. This more complete rescue was not the result of a dose-dependent increase in overall cle-1 transcript levels, because a strain expressing endogenously regulated cle-1 transcripts from the unmodified fosmid (Ex[unmodified]) at comparable or even lower levels than the strains expressing either isoform alone also completely rescued the phenotype (Figure 4.8 C and D). These results suggest that
Figure 4.8: Isoform-specific analysis of the cle-1 gene

(A) RT-PCR assay detecting cle-1 exon 18 isoforms in wild-type and cle-1(cg120) mutant animals (top panels). Note that the cg120 deletion spans a region downstream of the alternative splicing event, and could still be detected for both transcripts by RT-PCR. gpd-2 transcripts were used as a loading control (bottom left panel). The top right panel shows that selective expression of exon 18 included (inc), excluded (exc), or both (inc+exc) isoforms were re-introduced into cle-1(cg120) animals from isoform-specific fosmids. Since these isoforms are over-expressed, only relatively trace amounts of endogenous transcript were detected. To control for differences in transmission rates of extrachromosomal arrays, transcripts from the Pmyo-2::mRFP injection marker used in our transformation experiments were measured (bottom right panel).

(B) Morphologies of GABAergic synapses along dorsal nerve cord visualized by the vesicle marker juIs1 in various genotypes as described in A. Scale bar: 10 um.

(C) Total number of GABAergic synapses along the dorsal nerve cord in genotypes analyzed in panels A and B. Error bars represent +/- one standard deviation from the mean of the number of individuals (n) scored. Asterisks indicate p-values less than 0.05 (*) or 0.01 (**), Kruskal-Wallis test.
both cle-1 isoforms are required together for proper synaptic development in GABAergic neurons.

Second, I performed an isoform-specific analysis of the mod-1 gene which encodes a subunit of a serotonin-gated chloride channel that modulates several animal behaviors, including the enhanced-slowing response, sensitivity to exogenous serotonin (5-HT), egg-laying, and aversive learning in response to pathogenic bacteria (Dempsey et al., 2005; Ranganathan et al., 2000; Zhang et al., 2005). My bichromatic reporter analysis identified neuron subtype-specific differential splicing of exon 9 (Figures 4.6 and 4.9 A). Remarkably, different behaviors were found to depend on the expression of specific isoforms (Figure 4.9 B, C, and D). For example, locomotion in wild-type animals is strongly inhibited within 20 minutes of exposure to exogenous 5-HT, whereas mod-1(ok103) null animals are not severely affected (Ranganathan et al. 2000; Figure 4.9 C). Serotonin-dependent immobilization was restored in transgenic animals transformed with each of the mod-1 exon 9 isoform-specific foslids, either alone or in combination (Figure 4.9 C). Therefore each isoform functions redundantly to mediate the sensitivity to 5-HT for locomotion. However, these isoforms exhibit non-redundant functions in egg laying. In wild-type animals, egg-laying is stimulated by exogenous exposure to 5-HT, and this stimulation is inhibited in the presence of dopamine (Figure 4.9 D). In mod-1 mutants, this dopamine-mediated inhibitory effect is eliminated (Dempsey et al., 2005; Figure 4.9 D). The exon 9-included mod-1 isoform was specifically required to restore the dopamine mediated inhibitory response (Figure 4.9 D), indicating a specific role of this isoform in transducing the inhibitory dopamine signal.

The inclusion of exon 9 in mod-1 transcripts leads to the insertion of a 14 amino acid polypeptide in the cytoplasmic linker region of MOD-1. In vertebrate ligand-gated chloride channels, this linker region is implicated in modulating the biophysical properties of the channel, as well as serving as a scaffold for interactions with downstream signaling molecules (Lynch, 2004). Thus, our integrated approach of combining global profiling data and physiological characterization of splice variants provide a valuable resource for understanding extensive yet under-explored roles of AS regulation during C. elegans development.
Figure 4.9: Isoform-specific analysis of the mod-1 gene

(A) Merged fluorescent and DIC images of CNS neurons of an adult animal expressing a bichromatic reporter for mod-1 exon 9 inclusion (mRFP; red signal) and exclusion (EGFP; green signal) events. Scale bar: 25 um.

(B) RT-PCR assay validation of mod-1 exon 9 isoforms in wild-type and mod-1(ok103) mutant animals (top panels). gpd-2 transcript levels were used as a loading control (bottom left panel). The top right panel shows selective expression of exon 9 included (inc), excluded (exc), or both (inc+exc) isoforms re-introduced into mod-1(ok103) animals. To control for differences in transmission rates of transgenic lines, transcripts from the Pmyo-2::mRFP injection marker used in our transformation experiments were measured (bottom right panel).

(C) The percentage of animals immobilized after 20 minutes exposure to 33 mM 5-HT. Error bars represent the mean of three experiments +/- one standard deviation (n=20 individuals per experiment). Asterisks indicate a p < 1x10^-4 (student’s T-test). Genotypes for each strain are as described in B.

(D) Animals were either exposed to 5 mg/mL 5-HT alone, or to 5 mg/mL 5-HT and 6 mg/mL dopamine and the number of eggs laid in each case was normalized to the number of eggs laid during exposure to 5-HT alone. Genotypes examined are as described in B. Error bars represent mean +/- one S.E.M values. Asterisks indicate p < 1x10^-4 (student’s T-test).
4.5 Discussion

Development depends on highly regulated changes in gene expression profiles. In *C. elegans*, significant progress has been made to globally map these gene expression changes through the use of expression microarrays and through the generation of transgenic reporter lines. However, regulated AS events also contribute to development, yet the spatio-temporal expression patterns as well as the functional specificity of individual isoforms remain poorly understood. We have combined the resources developed in this study into a pipeline to facilitate the systematic analysis of AS during development.

Using complementary genome-wide approaches, we have verified the majority of annotated AS events, have identified thousands of novel candidate splice variants, and estimate that AS in *C. elegans* is far more prevalent than previously believed. Our dataset has also revealed hundreds of splicing events that are differentially regulated during development, providing the most comprehensive catalog of such events to date. Results from this study complement and extend recent analyses of the *C. elegans* transcriptome by RNA-Seq and microarray profiling (Barberan-Soler and Zahler, 2008; Hillier et al., 2009; Mangone et al., 2010; Ramani et al., 2009). For instance, our RNA-Seq analysis, which combined recently generated datasets (Hillier et al., 2009; Ramani et al., 2009), maximized the quantity of reads available for interrogating splice site usage. Our set of temporally regulated AS events are enriched in cis-regulatory elements bound by known splicing regulators not previously believed to play a role in differential splicing during development. Furthermore, additional motifs recognized by as of yet unidentified trans-factors have been identified, shedding light on potentially novel mechanisms of AS. Future studies will be directed at identifying these novel RNA binding proteins and determining how they act either alone or in combination to regulate AS during development.

The incorporation of our data in the *C. elegans* Splice Browser ([http://splicebrowser.ccbr.utoronto.ca](http://splicebrowse.ccbr.utoronto.ca)), will allow researchers to assess the relative usage of individual AS events in genes of interest, and to facilitate future functional analyses. As additional RNA-Seq and microarray data generated by either the modENCODE consortium or individual researchers become available, they will be integrated into the Splice Browser, further improving the quality and coverage of *C. elegans* AS events. This increased quality and coverage will improve our capacity to identify additional cis-regulatory motifs in all AS events as well as
temporally-regulated events. As these data accumulate it will become possible to build splicing regulatory networks and better understand mechanisms governing AS.

The bichromatic reporter system facilitates the *in vivo* monitoring and detection of splicing variation at single-cell resolution. Recently, a bichromatic reporter similar in nature to what I have presented here has been described for use in *C. elegans* (Kuroyanagi et al., 2010). Collectively, these next-generation reporters provide a versatile system to afford the interrogation of all classes of AS events. My reporter system is also compatible with fosmid recombineering and Gateway recombination systems (Tursun et al., 2009; Walhout et al., 2000), such that expression under the control of either endogenous or heterologous promoters can be achieved with simple manipulations (Materials and Methods). My reporter analyses have revealed unexpected insights: the observation of complex neuron-specific splicing patterns. Some of the detected AS events occur in genes encoding proteins thought to be expressed and function ubiquitously throughout the nervous system, such as UNC-16 (Byrd et al., 2001). It is intriguing that these genes exhibit intricate AS patterns with neuronal subtype specificity and it will be important to address the functional consequences of these isoforms on nervous system development and function with future genetic and biochemical analyses.

My observations in *C. elegans* suggest that neuron-specific splicing may be frequently used as a mechanism to generate cellular diversity in more complex nervous systems. Collectively, these results advocate the necessity to examine AS regulation in specific cell populations. Genetic screens using bichromatic AS reporters in *C. elegans* have been effective in identifying splicing regulators involved in tissue and development-specific splicing decisions (Kuroyanagi et al., 2006; Ohno et al., 2008). Utilizing the reporters described in our study, these approaches will undoubtedly yield novel insights into how AS contributes to cellular complexity in the nervous system.

I have also demonstrated in this study that isoform-specific transgene systems are likely to reveal a surprising degree of complexity and functional specificity of isoforms in different biological processes. This approach is specifically useful in the absence of any prior knowledge, such as whether or not an AS event overlaps a protein domain with a previously characterized role. Such information will yield novel insights into protein diversification driven by AS or complementary evolutionary forces across species. With the advent of robust recombineering
protocols, it is highly advantageous to use fosmid-based transgenes for such experiments because they are more likely to recapitulate endogenous expression patterns of the gene of interest. Future work taking advantage of the relative ease of performing genetic analyses in *C. elegans*, combined with an increasing collection of deletion mutants (Moerman and Barstead, 2008), will enable rapid, *in vivo* functional exploration of any identified splice variant of interest.

In summary, the resources presented in this study will further facilitate the systematic investigation of AS in transcriptome regulation, evolution, and in the generation of functional complexity.
Chapter 5

Conclusions and Future Directions

5.1 Further dissection of AS differences between humans and chimpanzees

Continued exploration of the molecular changes that have contributed to human evolution remains a major goal. Results from my work profiling AS in humans and chimpanzees provided initial evidence that conserved exons from two closely related species have rapidly evolved differences in inclusion levels in transcripts. Since the publication of my work, additional studies have described profiling of splicing differences between these species using high-density exon junction microarrays and RNA-Seq (Blekhman et al., 2010; Lin et al., 2010; Shen et al., 2011). These recent studies have identified a larger repertoire of AS events differentially regulated between humans and chimpanzees, and have confirmed and further expanded results observed in our original study. The new data from Lin and colleagues suggests that exons with species-specific splicing differences have accelerated rates of substitutions in synonymous codon sites and in their flanking intron sequences, compared with exons showing no change in splicing patterns (Lin et al., 2010). Additional analysis of one candidate gene identified two nucleotide substitutions in humans that are responsible for the lineage-specific splicing patterns observed. These results collectively demonstrate that subtle changes in cis-regulatory elements can lead to rapidly evolving effects on splicing regulation in humans and chimpanzees.

In addition to the profiling of conserved exons, several studies have also examined recent exon gain and loss events through comparative genomics and transcriptome-wide analyses in primates (Huang et al., 2008; Kim and Hahn, 2011; Shen et al., 2011; Zhuo et al., 2007). In one of these studies, the extent to which Alu element-derived exons were alternatively spliced in human tissue samples was assessed by RNA-Seq (Shen et al., 2011). A significant fraction of these exons are included in their corresponding transcripts more than 50% of the time, suggesting that they are often not minor isoforms. Additionally, these Alu-derived exons appear to be enriched in zinc finger transcription factor genes and preferentially located in 5’ UTRs. Subsequent analysis of the effects of including several of these 5’ UTR exons in transcripts revealed that they frequently altered the translation efficiency of reporter transcripts.

In addition to our work, these latter examples reveal the importance of performing additional experiments aimed at identifying the functional consequences of species-specific AS events in relevant cell lines. However, one caveat is that most of these studies have been limited
to immortalized cell lines derived from lymphoblast or epithelial cell types and often not from relevant tissue-types. Recent breakthroughs in the ability to reprogram adult fibroblast cells into induced pluripotent stem cells (iPSCs) have opened the door to accessible primary cells from non-human primates, which can be differentiated into diverse cell types (Yamanaka and Blau, 2010; Zhong et al., 2011). Comparisons of potential differences between human and chimpanzee isoforms in otherwise equivalently treated cells from each species will allow more controlled and relevant investigation into functional consequences. Finally, in the most extreme case, model organisms such as the mouse can be used to transgenically express humanized or other species-specific isoforms to study their functional properties in vivo. This approach has yielded interesting results in studies of FOXP2, a transcription factor believed to have been subjected to positive selection in the human lineage and which is implicated in the evolution of language. Amazingly, when the human form of FOXP2 was expressed in mice (Enard et al., 2009), differences relative to control mice were observed in vocalization patterns, behavior, and synaptic plasticity in the striatum—a region of the brain known to be affected in humans with speech deficits. Extending these results, a recent comparative study of human and chimpanzee FOXP2 has demonstrated that these orthologous proteins exhibit different transcriptional regulatory activity (Konopka et al., 2009). Among the genes differentially regulated by human and chimp FOXP2 are those with critical functions in human central nervous system development. As transcriptome profiling methodologies continue to improve and we identify additional regulatory event differences between humans and chimpanzees, a combination of these above-mentioned approaches will prove quite valuable.

5.2 Detailed investigations into the mechanism of regulation by nSR100, its regulatory network, and its role in nervous system development

Several studies, including my work on nSR100, have suggested that proteins encoded by transcripts with AS events co-regulated by a given splicing factor also have a high propensity to physically associate with each other through protein-protein interactions (PPIs) (Calarco et al., 2009; Ule et al., 2005b; Warzecha et al., 2010). These observations suggest that tissue-specific AS networks may have evolved in part to fine-tune PPIs for a given cell or tissue-type, and are consistent with data indicating that a significant proportion of alternative exons map to the surfaces of proteins, potentially influencing PPIs (Wang et al., 2005b). Approaches that can be
performed on a large scale, such as yeast two-hybrid assays (Bruckner et al., 2009) or high-throughput mammalian co-immunoprecipitation assays (Barrios-Rodiles et al., 2005; Lievens et al., 2009), can be utilized to systematically test how PPIs are affected by individual splice variants in these networks. We are currently applying such approaches to study the influence of nSR100 on PPI networks in the nervous system (J. Ellis, M. Barrios-Rodiles, J. Calarco, J. Wrana, and B.J. Blencowe, unpublished data).

In addition to AS differences observed during our microarray profiling of nSR100-depleted neuroblastoma cells, we also identified transcripts with altered overall expression levels compared to control shRNA treated cells (data not shown). We speculated that these steady-state transcript level differences were either due to a direct role of nSR100 in regulating transcription, or through altered splicing regulation of one or more key transcription factors. Recent work from our lab has indicated that at least the latter possibility is true; nSR100 was found to regulate the inclusion of a neural-specific exon in transcripts encoding the RE-1 silencing transcription factor (REST) that alters the ability of the encoded protein to repress genes involved in neural differentiation (B. Raj and B.J. Blencowe, submitted). Intriguingly, there is considerable overlap between known REST-regulated transcripts and mRNAs showing reduced expression upon depletion of nSR100. Additionally, REST was observed to repress nSR100 levels, resulting in a regulatory feedback loop between these two factors. These results demonstrate that nSR100 coordinates a multi-layered network of transcriptome changes, both at the level of AS and through coupling with REST activity to regulate gene expression, in order to promote differentiation to the neural lineage.

Individual transcripts and isoforms of this multi-layered network should be tested for their possible role in neural differentiation. The use of RNA interference (RNAi) to knock down these transcripts in cellular models of neural differentiation will provide a suitable system to test their requirement in vitro. Alternatively, the use of morpholinos or other modified antisense oligonucleotides targeting these transcripts in developing zebrafish or mice followed by immunohistochemistry or FISH to monitor neuronal differentiation would serve as an excellent approach to characterize developmental effects in vivo. These assays will ultimately determine the relative contributions of transcripts and isoforms in the network to the phenotypes observed when nSR100 is depleted.
Finally, further insight into the mechanism by which nSR100 regulates neural-specific AS should be obtained. In particular, how nSR100 functions together with PTB and PTBP2, as well as other RNA binding proteins, to promote exon inclusion in neural cells, is still unclear. Recent mutagenesis experiments of cis elements in the intronic regions surrounding alternative exon 16 in Daam1 transcripts revealed several splicing regulatory motifs that play a role in its inclusion in neuroblastoma cells (Barash et al., 2010). These cis elements include those bound by Fox, MBNL, and CELF splicing regulators, indicating that multiple inputs play a role in the proper regulation of this exon. It will be interesting to determine if nSR100 physically interacts with these additional regulators. Additionally, the spliceosomal assembly step at which nSR100 stimulates exon recognition is currently unknown. Existing biochemical approaches using splicing extracts and in vitro transcribed substrates can be used to tease apart this mechanism. The RNA binding activity we identified in our initial study of nSR100 may play a key role in its regulation of AS. Thus, in addition to the above analyses, CLIP-Seq experiments should be performed to help identify specific locations in pre-mRNA transcripts that nSR100 associates with to promote exon inclusion, and possibly uncover a consensus binding sequence. Collectively, all of the experiments described here will provide a more detailed picture into how the evolution of nSR100 has contributed to vertebrate nervous system development and complexity.

5.3 Neuron- and developmental stage-specific splicing regulation in C. elegans

My observation that transcripts in the nervous system of C. elegans can display complex and often neuron-specific AS patterns is very exciting. These results strongly suggest that the splicing machinery in individual cells of the nervous system must also be differentially regulated to produce such AS patterns. Forward and reverse genetic screens can be performed to identify factors regulating these cell-specific splicing decisions. Worms expressing bichromatic reporters with neuron-specific AS patterns will be screened for observable deviations from wild-type patterns of red/green fluorescence intensities in cells of interest after mutagenesis or RNAi-based perturbation of gene function. These genetic screens will likely identify genes encoding known or putative splicing factors and RBPs, but may also identify novel candidates such as non-coding RNAs, chromatin modifying factors, or even signaling proteins, all of which have been shown to regulate AS (Allemand et al., 2008; Khanna and Stamm, 2010; Luco et al., 2010; Lynch, 2007;
Tripathi et al., 2010). Both classes of mutants will be interesting and would warrant follow-up experiments.

Comparison of the transcriptomes of wild-type and splicing factor mutant strains will help identify networks of AS events regulated in specific neuronal classes. Complementary to this approach, CLIP-Seq could be utilized to identify the set of transcripts directly bound by RBPs of interest and further classify potential regulatory targets. Importantly, the *C. elegans* genome contains approximately 500 genes predicted to encode RBPs, the majority of which are conserved across metazoans (Lee and Schedl, 2006). If any of the identified factors are conserved in mammals, analogous experiments to those performed in *C. elegans* could be conducted in mammalian systems. Ultimately, these identified candidate regulators will provide insights into the mechanisms governing neuronal subtype-specific AS, and provide entry points towards identifying isoforms required for these these neuronal cells to acquire their distinct identity and function.

Developmental stage-specific isoforms identified in our transcriptome analysis also provide a catalog of interesting gene regulatory events that likely play a role at distinct periods in the life of the organism. In agreement with important stage-specific functions of these isoforms, a small-scale comparison of temporally-regulated AS events in *C. elegans* and *C. briggsae* revealed a strong conservation in their dynamic splicing patterns across development (Rukov et al., 2007). As mentioned in Chapter 4, my fosmid based isoform-specific transgene rescue approach can be utilized to distinguish the roles of individual splice variants during development. If a given isoform is required at a specific stage in development, loss of this isoform will lead to a more pronounced phenotype at the corresponding stage that will not be rescued by expression of the alternative isoform. The use of this system will uncover previously under-appreciated roles for characterized genes, and guide insights into the function of uncharacterized genes. A similar strategy can also be used to interrogate cell type-specific AS events.
5.4 Promises and challenges ahead: tissue-specific splicing regulatory networks

5.4.1 Examining connections among splicing factors

Numerous biochemical studies have revealed that tissue-specific splicing factors act in concert with other tissue-specific or ubiquitously expressed splicing regulators to regulate AS (Chen and Manley, 2009). It is therefore becoming apparent that integrating regulatory networks of individual splicing factors is necessary in order to achieve an understanding of combinatorial control of tissue-specific AS patterns. Another emerging theme is that altered expression or kinetic activity of factors defined as “core” components of the spliceosome can lead to specific changes in AS patterns (Park et al., 2004; Saltzman et al., 2011; Yu et al., 2008; Zhang et al., 2008b). Future studies will be required to better understand how these core splicing regulators can interact with the landscape of other tissue-specific and generally expressed auxiliary factors.

As mentioned above, splicing factors can be dynamically regulated during development or in response to certain stimuli. Temporal regulation of the expression of CELF and MBNL regulators during heart development has been shown to play an integral role in establishing embryonic or adult splicing patterns (Kalsotra et al., 2008). Depolarization-induced changes in the nuclear accumulation of hnRNPA1 and Fox-1 in neurons lead to modulation of exon inclusion in target transcripts encoding synaptic proteins (An and Grabowski, 2007; Lee et al., 2009). A recent study has also demonstrated that SR proteins can dynamically alter the repertoire of target transcripts they interact with during neuronal differentiation (Anko et al., 2010). These examples indicate that most genome-wide analyses of splicing regulatory networks have only provided a snapshot of the landscape of transcripts bound to and influenced by a given splicing factor. Future studies are likely to focus on large-scale analyses over developmental time courses and under different physiological perturbations.

Splicing factors frequently regulate AS events within their own transcripts and in those encoding other RNA binding proteins. These auto- and cross-regulatory events often insert or remove PTC-containing exons that lead to the selective degradation of spliced transcripts by the NMD pathway, ultimately influencing the overall abundances of these factors (Lareau et al., 2007b; Ni et al., 2007; Saltzman et al., 2008; Spellman et al., 2007). Alternatively, the activity of splicing regulators was also found to be influenced by AS events in their own transcripts creating
dominant negative isoforms or variants with altered post-translational modification sites (Damianov and Black, 2010; Dredge et al., 2005). Collectively, these results suggest that altering the function of any single splicing factor may result in the perturbation of a network of factors. Understanding the functional relationships in these regulatory networks will greatly improve our ability to identify programs of exons controlled during specific developmental processes.

5.4.2 The generation of predictive models governing tissue-specific splicing

From a theoretical perspective, as all of these analyses are being performed, their cumulative integration provide resources to generate predictive models of tissue-specific splicing regulation. Recent work from the Frey and Blencowe groups demonstrated that the combination of large-scale AS profiling datasets and machine learning approaches can be used to infer combinations of features, such as cis-regulatory sequences, exon/intron lengths, splice site strengths, and others, that are maximally predictive of tissue-specific AS patterns (Barash et al., 2010; Appendix 4). The utility of this inferred tissue-dependent “splicing code” was demonstrated through its ability to identify known and predict new associations between sets of features, provide testable hypotheses for critical sequence elements involved in regulating AS patterns, and uncover novel classes of tissue-specific AS events from the analysis of genomic sequence alone (Barash et al., 2010). Recently, Zhang and colleagues combined several genome-wide datasets from studying Nova-regulated alternative exons to establish parameters that improve predictions of Nova-dependent AS events through a machine learning approach (Zhang et al., 2010). This approach was able to identify many more transcripts regulated by the Nova proteins than could be identified by each experimental dataset in isolation, predict combinatorial regulation with Fox splicing factors, and infer novel principles by which these regulators function.

These studies would not be possible without the generation of empirically-derived genome-wide datasets, from which the algorithms are usually trained, and decades of intensive research from the splicing community, from which parameters have been derived. As more detailed AS profiling data become available, new trans-acting factors and cis-regulatory elements are uncovered, and other features such as chromatin states are included in new algorithms, the predictive power of inferred models will continue to improve.
5.3 Concluding remarks

The inception of large-scale approaches to analyze gene expression has ushered in a new era in which biological phenomena can be studied both on a gene-by-gene basis and from a global perspective. A principle that has emerged from genome-wide transcriptome analyses is that master regulators have evolved to control networks of transcripts involved in common developmental and functional pathways. Splicing regulatory networks also appear to have evolved under similar organizing principles. Investigations of these networks have served two key purposes. First, they have enabled new insights into the mechanisms by which splicing factors regulate AS. Second, they have identified novel functional associations between gene products that would otherwise not have been illuminated by studies of transcription factor or miRNA regulated networks. As we move forward, these studies will lead to a comprehensive picture into how cellular complexity and species-specific differences have evolved through the interplay of multiple modes of gene regulation, and how the perturbation of these processes lead to various diseases.
References


Buckanovich, R.J., Posner, J.B., and Darnell, R.B. (1993). Nova, the paraneoplastic Ri antigen, is homologous to an RNA-binding protein and is specifically expressed in the developing motor system. Neuron 11, 657-672.


Appendix 1

Supporting data for Chapter 2
Table A1.1: Relevant information pertaining to the human, chimpanzee and macaque specimens used in this study.

Information includes institutional identification number, institutional source, age, sex, and tissue samples obtained for the purposes of this study.

<table>
<thead>
<tr>
<th>Institution number</th>
<th>Species</th>
<th>Source*</th>
<th>Sex</th>
<th>Age (years)</th>
<th>Tissue Analyzed^</th>
</tr>
</thead>
<tbody>
<tr>
<td>1029</td>
<td><em>H. sapiens</em></td>
<td>UMBTB</td>
<td>M</td>
<td>30</td>
<td>FP</td>
</tr>
<tr>
<td>1227</td>
<td><em>H. sapiens</em></td>
<td>UMBTB</td>
<td>M</td>
<td>53</td>
<td>FP</td>
</tr>
<tr>
<td>1570</td>
<td><em>H. sapiens</em></td>
<td>UMBTB</td>
<td>M</td>
<td>48</td>
<td>FP</td>
</tr>
<tr>
<td>1673</td>
<td><em>H. sapiens</em></td>
<td>UMBTB</td>
<td>M</td>
<td>42</td>
<td>FP</td>
</tr>
<tr>
<td>1903</td>
<td><em>H. sapiens</em></td>
<td>UMBTB</td>
<td>M</td>
<td>40</td>
<td>FP</td>
</tr>
<tr>
<td>859</td>
<td><em>H. sapiens</em></td>
<td>UMBTB</td>
<td>F</td>
<td>36</td>
<td>Heart</td>
</tr>
<tr>
<td>1455</td>
<td><em>H. sapiens</em></td>
<td>UMBTB</td>
<td>F</td>
<td>25</td>
<td>Heart</td>
</tr>
<tr>
<td>1475</td>
<td><em>H. sapiens</em></td>
<td>UMBTB</td>
<td>M</td>
<td>20</td>
<td>Heart</td>
</tr>
<tr>
<td>1542</td>
<td><em>H. sapiens</em></td>
<td>UMBTB</td>
<td>M</td>
<td>23</td>
<td>Heart</td>
</tr>
<tr>
<td>C0572</td>
<td><em>P. troglodytes</em></td>
<td>YNPRC</td>
<td>M</td>
<td>21</td>
<td>FP</td>
</tr>
<tr>
<td>C0421</td>
<td><em>P. troglodytes</em></td>
<td>YNPRC</td>
<td>M</td>
<td>27</td>
<td>FP</td>
</tr>
<tr>
<td>C0471</td>
<td><em>P. troglodytes</em></td>
<td>YNPRC</td>
<td>M</td>
<td>24</td>
<td>FP, Heart</td>
</tr>
<tr>
<td>C0425</td>
<td><em>P. troglodytes</em></td>
<td>YNPRC</td>
<td>M</td>
<td>28</td>
<td>FP, Heart</td>
</tr>
<tr>
<td>C0058</td>
<td><em>P. troglodytes</em></td>
<td>YNPRC</td>
<td>F</td>
<td>50</td>
<td>Heart</td>
</tr>
<tr>
<td>C0160</td>
<td><em>P. troglodytes</em></td>
<td>YNPRC</td>
<td>F</td>
<td>24</td>
<td>Heart</td>
</tr>
<tr>
<td>N840</td>
<td><em>M. mulatta</em></td>
<td>YNPRC</td>
<td>M</td>
<td>21</td>
<td>FP</td>
</tr>
<tr>
<td>RKe6</td>
<td><em>M. mulatta</em></td>
<td>YNPRC</td>
<td>M</td>
<td>7</td>
<td>FP</td>
</tr>
<tr>
<td>RU03</td>
<td><em>M. mulatta</em></td>
<td>YNPRC</td>
<td>M</td>
<td>13</td>
<td>FP</td>
</tr>
<tr>
<td>RHy6</td>
<td><em>M. mulatta</em></td>
<td>YNPRC</td>
<td>M</td>
<td>13</td>
<td>Heart</td>
</tr>
<tr>
<td>RY13</td>
<td><em>M. mulatta</em></td>
<td>YNPRC</td>
<td>M</td>
<td>14</td>
<td>Heart</td>
</tr>
<tr>
<td>RA06</td>
<td><em>M. mulatta</em></td>
<td>YNPRC</td>
<td>M</td>
<td>3</td>
<td>Heart</td>
</tr>
</tbody>
</table>

*a UMBTB, University of Maryland Brain and Tissue Bank for Developmental Disorders; YNPRC, Yerkes National Primate Research Center.
^Cerebral cortex samples were obtained from the frontal pole (FP) of the left hemisphere.
**Additional Supporting Data found on CD at end of thesis:**

**SD 2.1:** Tab delimited text file containing exons analyzed in comparative genomic alignment.
For each exon studied the columns provide the following information, from left to right: unigene cluster, ASAP II database identifier, gene name, gene symbol, human exon sequence and flanking upstream and downstream 150 nucleotides of intron sequence, orthologous chimpanzee exon/flanking intron sequence, percent sequence identity, percent of sequence that could be aligned, and alternative exon length in nucleotides.

**SD 2.2:** Data file containing microarray predicted percent inclusion values for the human-chimpanzee frontal cortex and heart datasets.
Unigene clusters and gene names are also listed, in addition to the upstream constitutive, alternative and downstream constitutive exon sequences for each of the splicing events being monitored.

**SD 2.3:** Primer sequences used in this study.
Appendix 2

Supporting data for Chapter 3
Figure A2.1: Enlarged view of cluster graph shown and described in Figure 3.1.
Figure A2.2: Additional alternative splicing events regulated by nSR100.
RT-PCR assays comparing the splicing patterns of eight AS events in Neuro2a cells expressing control shRNAs or nSR100 shRNAs. In all cases, knockdown of nSR100 leads to more exon skipping.
Figure A2.3: Additional motifs significantly overrepresented in the introns flanking nSR100-regulated alternative exons.
The diagram outlines the positions in the flanking introns that the particular sequence elements were enriched (FDR corrected \( p \) value < 0.01). Motifs were manually grouped together on the basis of sequence similarity.
Figure A2.4: The nucleotide sequence of mouse Daam1 exon 16 and 296 nucleotides of upstream intronic sequence.
The green region (region 1, nucleotides –280 to -251) overlaps with the pyrimidine-rich region identified by the minigene reporter analysis described in Figure 5. The pyrimidine residues in this region and all other regions of interest (highlighted in diagram and chosen on the basis of containing pyrimidine-rich sequence elements and/or overlap with n/brPTB CLIP tags) were mutated primarily to purine residues. A legend below the sequence describes the various combinations of sequences that were mutated in this study and tested for PTB and nPTB binding by RNA mobility shift assays.

Figure A2.5: ZnSR100 pre-mRNA splicing is blocked in embryos treated with splice blocking morpholino oligonucleotide (MOspl).
ZnSR100 transcript splicing patterns in the vicinity of exon 5 and intron 5 were analyzed in WT and nSR100 MOspl morphant embryos at 50hpf. A near complete disappearance of fully spliced transcripts along with the accumulation of transcripts with retained intron 5 could be observed in morphant embryos relative to WT embryos.
Figure A2.6: Phenotype of nSR100 MOspl morphant embryos.
A-B. Lateral views of control p53MO-injected (A) and nSR100 MOspl-injected (B) embryos at 30hpf. nSR100 morphant embryos exhibit only a slight delay in embryonic development, as exhibited by morphogenesis of the midbrain-hindbrain boundary (asterisks). C-D. Lateral (C, D) and ventral (C’, D’) views of control p53MO-injected (C, C’) and nSR100 MOspl-injected (D, D’) embryos at 48hpf. Note the expanded ventricle (arrowhead) and convoluted body axis of nSR100 morphant embryos. However, general embryonic patterning and morphogenesis of nSR100 morphants, including heart formation (arrows), appears relatively normal.

Figure A2.7: Morphology and islet-1 expression in control and nSR100 MOspl embryos at 24hpf.
Figure A2.8: Injection of ZnSR100 mRNA rescues neurite outgrowth defects in ZnSR100 MOspl morphant embryos.

A-C. Confocal projections of trigeminal ganglion in Tg(sensory:GFP) transgenic embryos at 30hpf. A. p53 + ZnSR100 MOspl injected embryo. B-C. p53 + ZnSR100 MOspl embryo, co-injected with ZnSR100 mRNA.
Additional Supporting Data found on CD at end of thesis:

SD 3.1: Database of mouse RS domain genes.
Each of the computationally identified mouse RS domain protein encoding sequences was manually annotated in Table S1 on the basis pf information available for the corresponding gene entry in the Mouse Genome Informatics (MGI) database. For functional annotations, a combination of information from available Gene Ontology terms and the published literature was used to broadly assign each protein to different functions categories (see also Figure 1). Additional information provided includes the presence of InterPro domains, the existence and percent identity shared with any human orthologs, and whether the ortholog contains at least one RSRS or SRSR tetrapeptide. For mouse protein sequences, the following quantitative information on RS domain characteristics is provided: (1) the total number of RS or SR repeats, the total number of RSRS or SRSR repeats, the maximum number of RS or SR repeats in a 20 amino acid-long or 60 amino acid-long window, and the maximum number of RSRS or SRSR repeats in a 20 and 60 amino acid-long window.

SD 3.2: Alternative Splicing Events Predicted by GenASAP to be Regulated by nSR100.
The first 81 entries represent AS events predicted to undergo more alternative exon skipping upon knockdown of nSR100. The remaining 34 entries correspond to AS events predicted to undergo more alternative exon inclusion upon nSR100 knockdown. The predicted %ex differences for each of the probe replicates is shown for the events. Additional columns contain relevant information pertaining to each splicing event, including the corresponding gene names, alternative exon sequences, flanking constitutive exon sequences, number of EST sequences supporting evidence for the included and skipped isoforms in human, conservation of the alternative exon and/or alternative splicing in human, overlap between alternative exons and flanking constitutive exons with known protein domains, whether the alternative exon is frame preserving, and whether the alternative splicing event has the potential to introduce a premature termination codon.

SD 3.3: Primers used in this study.
Appendix 3

Supporting data for Chapter 4
Figure A3.1: Analysis of how read depth affects junction identification

Total junctions identified versus new junctions identified were compared A. Truepositive junctions from Fig 2A, B. All adjacent junctions from WS200, and C. All non-adjacent junctions from WS200). In each case we increase the read depth by adding additional sequence reads and ask how many total reads we see and how many of these are new junctions and we notice that we saturate the potential junctions we can query. We also notice that while we saturate new junctions we can find, we can provide additional read support to junctions that were previously overlapped by single reads D. To estimate the accuracy of junction identification, we compared the accuracy of our algorithm by mapping the same set of reads to either a real database (blue diamonds) or to a random junction database (red diamonds). This shows that far fewer random junctions are mapped compared to real junction sequences E and that we have a very low false-positive rate misidentifying junction sequences (~2%).
Figure A3.2: Novel AS events identified by RNA-Seq
RT-PCR assays confirming six additional novel AS events identified by our RNA-Seq analysis. Primers are designed to anneal to flanking constitutive exons, leading to the amplification of two products. For each image, the slower migrating product represents the exon-included isoform, and the faster migrating product corresponds to the exon-skipped isoform.
Using SeedSearcher, we identified pentamer motifs located in the 50-nucleotide sequence upstream of temporally-regulated AS events. These 12 motifs were found to be statistically enriched in these temporally-regulated AS events compared to all AS events.
Figure A3.4: Motifs identified in the downstream region of temporally-regulated AS events
Using SeedSearcher, we identified pentamer motifs located in the 50-nucleotide sequence downstream of temporally-regulated AS events. We identify 19 motifs that are statistically enriched in these temporally-regulated AS events compared to all AS events.
Additional Supporting Data found on CD at end of thesis:

SD 4.1: Novel Splice junctions involving unannotated splice sites validated by RT-PCR assays.

SD 4.2: Alternative splicing events predicted to display statistically significant splicing differences based on RNA-Seq data during development.

SD 4.3: PATA derived %In values for temporally regulated AS events from our microarry analysis

SD 4.4: List of genes with alternative splicing events to be analyzed by bichromatic reporters

SD 4.5: Primers used in this study
Appendix 4

Published Manuscripts