High-Throughput Analysis of Alternative Splicing Regulatory Networks that Control Cell Fate

by

Hong Han

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

© Copyright by Hong Han 2016
High-Throughput Analysis of Alternative Splicing Regulatory Networks that Control Cell Fate

Hong Han

Doctor of Philosophy

Department of Molecular Genetics
University of Toronto

2016

Abstract

Advances in high-throughput profiling technologies have revolutionized our view of transcriptome complexity and regulation. Alternative splicing (AS) is increasingly recognized as a widespread mechanism for gene regulation and is responsible for greatly expanding transcriptome and proteome diversity in metazoans. It is often controlled in a precise spatiotemporal manner by an intricate RNA-protein network. Given the extensive roles of AS in development and disease, understanding the molecular mechanisms underlying AS regulation is of critical importance. The main focus of my thesis is to develop and apply high-throughput strategies to systematically discover AS networks that control cell fate, as well as the factors and mechanisms that regulate these AS networks. By combining high-throughput RNA sequencing and splicing code analyses, I describe the identification of the muscleblind-like RNA binding proteins, MBNL1 and MBNL2, as conserved negative regulators of a large program of alternative exons that are differentially spliced between embryonic stem cells (ESCs)/induced pluripotent stem cells (iPSCs) and diverse non-ESCs/tissues. My results from this study provided the first evidence that AS regulatory networks play a central role in the core circuitry required for ESC pluripotency and iPSC reprogramming. Next, I present the development of a new multiplexed functional genomics screen whereby thousands of gene perturbations can be
simultaneously assayed for effects on dozens of endogenous AS and gene expression events of interest. Using this system, I revealed a large network of novel positive and negative factors from multiple gene layers that regulate distinct subsets of AS events linked to ESC pluripotency, neural differentiation, and transitions in reprogramming. In summary, these studies have significantly broadened our knowledge of the global regulatory networks that determine cell fate decisions, and provided powerful tools for large-scale investigation of the regulation of AS and other RNA metabolic pathways in diverse cell types.


**Acknowledgments**

When writing this dissertation and looking back at my PhD journey, I am extremely grateful for the support, encouragement, and stimulating interactions of many individuals. Without them, I would not have been able to gain confidence to overcome the challenges along the way and to enjoy the invaluable research experience during my graduate studies.

First of all, I would like to especially thank my PhD advisors, Ben Blencowe and Jason Moffat, who have helped me grow not only as a student, but also as an independent thinker and a scientist. They have provided me with the opportunity to work on two exciting multidisciplinary projects, as well as to collaborate and integrate with both local and international research communities. Their insightful perspective and guidance have significantly broadened my approach to solving scientific questions and are instrumental for every bit of my progress in the past few years. Their passion, dedication, and professionalism for research and mentoring are contagious. I greatly appreciate the unwavering encouragement, support, and friendship they have offered, which will continue to inspire and motivate me in my future career.

I would also like to express my sincere gratitude towards my exceptional present and past committee members, Jeff Wrana, James Ellis, and Corey Nislow, for their availability, fruitful discussions, and critical insights. Their thoughtful advice and support through and beyond regular committee meetings have helped shape many aspects of my PhD thesis research. In particular, I must thank Jeff and James for providing access to their valuable lab resources and for collaborating on several related projects.

The unique culture of openness has made the Blencowe and Moffat labs a fantastic place to learn, share, and explore. I have been fortunate to have worked and interacted with many talented people in both groups. I would like to thank all members of the labs for their support, friendship, and great times we have had together. I am very grateful for the help from Mathieu Gabut, who introduced me to various experimental techniques to study splicing when I first joined the lab. I have also benefited enormously from stimulating conversations with Bushra Raj, John Calarco, Arneet Saltzman, Joanna Ip, Anthony Mak, Kim Blakely, Christine Misquitta-Ali, Razvan Nutiu, and Franco Vizeacoumar. I am truly thankful for the unflagging support and sage advice of Dave O’Hanlon throughout the years, as well as numerous practical help and suggestions from Patricia
Mero and Troy Ketela on microscopy and robotic systems, without which experiments would not proceed as smoothly as planned. I owe much to Manuel Irimia, Ulrich Braunschweig, Robert Weatheritt, Thomas Gonatopoulos-Pournatzis, Kevin Ha, Sandy Pan, and Nuno Barbosa-Moraí, whose input and effort have directly contributed to several aspects of this thesis work. It has been a wonderful experience collaborating and discussing with them, and I have learned a lot about computational and experimental approaches. Additionally, I would like to extend my thanks to Kevin Brown, Jonathan Ellis, Serge Gueroussov, Mathieu Quesnel-Vallieres, and Megha Chandrashekhar for providing helpful feedback and advice on my research. Andrew Best, Eesha Sharma, Tim Sterne-Weiler, Michael Aregger, Jonathan Roth, Cristel Cheung, and Kristina Ognjanovic have been great colleagues to discuss science, travel, food, and much more. It has also been a great pleasure for me to work, share experiences, and learn together with several amazing undergraduate students, Juli Wang, Emil Nachman, and Valentina Slobodeniuc, as well as multiple rotation students.

During the course of my PhD studies, I have had the privilege to collaborate with many great scientists in different fields. I would like to sincerely thank Calley Hirsch, Dan Trcka, Laurent David, and Azadeh Golipour from the Wrana lab, as well as Joel Ross and Tadeo Thompson from the Ellis lab for their extensive help with reprogramming studies; Hoon-Ki Sung, Iacovos Michael, and Andras Nagy for their expertise in stem cell biology; Babak Alipanahi and Brendan Frey for assistance with splicing code analysis; Ernest Radovani and Jack Greenblatt for stimulating interactions and help with ChIP-Seq; Eric Wang and Christopher Burge for sharing their CLIP-Seq data; Debashish Ray and Tim Hughes for providing several recombinant proteins for in vitro experiments; Alessandro Datti, Frederick Vizeacoumar, and Thomas Sun at the Lunenfeld-Tanebaum Research Institute robotics facility for help with high-throughput siRNA screens; Dax Torti, Danica Leung, and Graham O’Hanlon at the Donnelly Sequencing Centre for sequencing samples; and Tara Paton, Jo-Anne Herbrick, and Steve Scherer at the Centre for Applied Genomics, the Hospital for Sick Children for offering access to their PCR machines.

As funding sources are relatively limited for international students, I am very honored to have my PhD research supported by the Ontario Graduate Scholarship and the University of Toronto Open Fellowship. Moreover, I gratefully acknowledge the financial support received through the Donnelly Centre Jennifer Dorrington Award, the University of Toronto School of Graduate Studies Conference Grant, and the Society of Chinese Bioscientist in America Best Paper Award.
My heartfelt thanks also go to my friends and family for their constant encouragement and support along the way. I am deeply indebted to my parents for everything they have done for me over all these years and words cannot fully express the gratitude I feel. Importantly, their enthusiasm to life and kindness to people have had a profound impact on my attitude and practice. By offering unconditional love and support, they always encourage me to follow my heart and to choose my own path, in both my career and my personal life. Their persistent confidence and belief in me is a great inspiration for me to work hard and move forward.

Last but not least, I would like to thank my best friend and husband, Yiming Wang. Both being graduate students, we met and got to know each other on the U of T campus. I have been lucky enough to travel this entire journey together with him, supporting and comforting each other during challenging times and celebrating each other’s success. I am grateful for the love, understanding, reassurance, encouragement, and optimism you have provided, and for always being there whenever I need you. Thank you for spending many evenings and weekends with me in the library, and giving me valuable feedback on my projects and thesis. As my most important companion and partner, I look forward to unfolding next chapter of life with you.
Table of Contents

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

Table of Contents ..................................................................................................................................... vii

List of Figures ............................................................................................................................................ xi

List of Appendices ..................................................................................................................................... xiv

List of Abbreviations ................................................................................................................................. xv

Chapter 1 ................................................................................................................................................... 1

1 Introduction.............................................................................................................................................. 2

1.1 Pre-mRNA splicing and the spliceosome ......................................................................................... 2

1.1.1 Mechanism of pre-mRNA splicing ............................................................................................. 3

1.1.2 Assembly of the spliceosome .................................................................................................... 6

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

1.3 Function of alternative splicing ........................................................................................................ 11

1.4 Mechanisms of alternative splicing regulation ................................................................................ 13

1.4.1 Cis-acting regulatory elements .................................................................................................... 13

1.4.1.1 Core splicing signals ............................................................................................................. 13

1.4.1.2 Exonic/intronic splicing enhancers/slicencers ..................................................................... 14

1.4.2 RNA secondary structures ........................................................................................................... 15

1.4.3 Trans-acting splicing regulators .................................................................................................. 16

1.4.3.1 Core spliceosomal proteins .................................................................................................. 16

1.4.3.2 SR and SR-related proteins ................................................................................................... 17

1.4.3.3 hnRNPs ................................................................................................................................ 18

1.4.3.4 Cell- and tissue-specific splicing factors .............................................................................. 19

1.4.4 Post-translational modifications of splicing factors ..................................................................... 21

1.4.5 Signal transduction pathways ...................................................................................................... 22

1.5 Coordination of alternative splicing with other gene regulatory machineries ......................... 23

1.6 Experimental and computational approaches facilitating high-throughput alternative splicing analysis ............................................................................................................................................... 26

1.6.1 Microarray profiling and high-throughput RNA sequencing ..................................................... 26

1.6.2 Crosslinking and immunoprecipitation coupled to high-throughput sequencing .................. 27

1.6.3 Computational methods for large-scale AS analysis .................................................................. 28

1.6.4 High-throughput screening of cis-regulatory elements ............................................................. 29
1.6.5 High-throughput screening of trans-acting splicing regulators ............................................. 29

1.7 Alternative splicing in ESC pluripotency, differentiation and somatic cell reprogramming ................................................................. 33
  1.7.1 AS complexity and dynamics in stem cell fate decisions ...................................................... 33
  1.7.2 Functionally important AS events in ESCs ........................................................................ 34
  1.7.3 Regulation of AS networks in ESCs, differentiation, and reprogramming ......................... 38

1.8 Rationale and overview of thesis research .............................................................................. 39

Chapter 2 ........................................................................................................................................ 42

2 MBNL proteins repress ESC-specific alternative splicing and reprogramming ............. 43
  2.1 Abstract ................................................................................................................................ 43
  2.2 Introduction .............................................................................................................................. 44
  2.3 Materials and Methods .......................................................................................................... 45
    2.3.1 Cell lines and cell culture ................................................................................................. 45
    2.3.2 siRNA transfection .......................................................................................................... 45
    2.3.3 Protein extraction and western blotting .......................................................................... 47
    2.3.4 RNA extraction and (q)RT-PCR assays ......................................................................... 47
    2.3.5 Mbnl1 and Mbnl2 overexpression in mouse ESCs .......................................................... 47
    2.3.6 Mouse ESC differentiation .............................................................................................. 48
    2.3.7 Immunofluorescence ........................................................................................................ 48
    2.3.8 iPSC colony formation assays and imaging from secondary MEF reprogramming .......... 49
    2.3.9 Teratoma Analysis ............................................................................................................. 49
    2.3.10 Chimerism analysis ......................................................................................................... 50
    2.3.11 Generation and characterization of human iPS cells ...................................................... 50
    2.3.12 Clonal analysis by RNA-Seq during reprogramming ..................................................... 51
    2.3.13 RNA-Seq data ............................................................................................................... 52
    2.3.14 RNA-Seq data analysis and identification of ESC-differential AS events .................... 52
    2.3.15 Analyses of splicing factor expression ............................................................................. 54
    2.3.16 Splicing code analyses .................................................................................................... 55
    2.3.17 CLIP-Seq analysis ........................................................................................................... 55
    2.3.18 Evolutionary conservation of ESC-differential events .................................................. 56
    2.3.19 Analyses of function and protein domain enrichment .................................................. 57
  2.4 Results ..................................................................................................................................... 58
    2.4.1 Identification of novel regulators of ESC-differential AS in human and mouse ............... 58
    2.4.2 MBNL proteins regulate ESC-specific AS of FOXP1 ..................................................... 62
2.4.3 MBNL proteins directly and negatively regulate approximately half of the ESC-differential AS events.......................... 65
2.4.4 MBNL overexpression in ESCs suppresses pluripotency and promotes differentiation... 71
2.4.5 Knockdown of MBNL proteins enhances somatic cell reprogramming ..................... 73

2.5 Discussion .................................................................................................................................................. 79
2.5.1 The central role of MBNL proteins in regulating ESC-differential AS, pluripotency, and reprogramming .......................................................... 79
2.5.2 Coordinated regulation of AS networks in stem cell biology ...................................................... 80

Chapter 3 .................................................................................................................................................. 81

3 Systematic discovery of alternative splicing regulatory networks controlling cell fate ...82

3.1 Abstract ..................................................................................................................................................... 82

3.2 Introduction .............................................................................................................................................. 83

3.3 Materials and Methods .......................................................................................................................... 85
3.3.1 Cell lines and cell culture ..................................................................................................................... 85
3.3.2 High-throughput siRNA knockdown and RNA purification ............................................................ 85
3.3.3 Systematic Parallel Analysis of endogenous RNA regulation coupled to barcode Sequencing (SPAR-Seq) .................................................................................. 85
3.3.4 Events monitored in the screen ............................................................................................................ 87
3.3.5 Selection of candidate splicing-related genes .................................................................................... 87
3.3.6 Pipeline for high-throughput screen data analyses ........................................................................... 88
3.3.7 Correlation network analyses and functional cluster identification ............................................... 90
3.3.8 Hit frequencies of proteins with certain domains ............................................................................... 91
3.3.9 Principal component analysis ............................................................................................................ 91
3.3.10 Analysis of correlation within CORUM complexes ....................................................................... 91
3.3.11 RNA-Seq experiments and analysis ................................................................................................. 92
3.3.12 Gene Ontology analyses .................................................................................................................. 93
3.3.13 Analysis of features associated with Zfp871-regulated events ...................................................... 93
3.3.14 ChIP-Seq experiments and analysis ................................................................................................ 93
3.3.15 iCLIP .................................................................................................................................................. 94
3.3.16 In vitro splicing assays ....................................................................................................................... 95
3.3.17 Protein extraction and western blotting ............................................................................................ 96
3.3.18 Co-immunoprecipitation assay ........................................................................................................ 96
3.3.19 RNA extraction and (q)RT-PCR assays ............................................................................................ 96
3.3.20 Cloning and plasmids ...................................................................................................................... 97

3.4 Results ..................................................................................................................................................... 98
3.4.1 A high-throughput screen linking trans-acting factors to endogenous AS events .......... 98
3.4.2 Correlated AS changes reveal multilayered regulatory pathways and complexes ........ 104
3.4.3 Novel positive and negative roles of different gene regulatory layers in controlling AS 109
3.4.4 Chromatin and transcription factors function as AS regulators............................... 112
3.4.5 An extensive role for zinc finger proteins in AS regulation...................................... 114
3.4.6 Nacc1 regulates ESC-differential AS at multiple levels.......................................... 121

3.5 Discussion .................................................................................................................. 125

Chapter 4 ....................................................................................................................... 128

4 Conclusions and Future Directions ........................................................................... 129

4.1 Concluding remarks ................................................................................................. 129

4.2 Future directions ...................................................................................................... 130

4.2.1 Further characterization of biological functions of cell fate-associated AS networks and their key regulators................................................................. 130

4.2.2 Systematic analysis of stage-specific AS regulatory networks during cell fate transitions ............................................................................................................ 132

4.2.3 Broader applications of the novel high-throughput screening strategy developed in this study ......................................................................................................... 133

4.2.3.1 High-throughput screening of trans-acting factors orchestrating stage-specific AS events linked to neurogenesis and neurological disorders.......... 133

4.2.3.2 Chemical screens for small molecule modulators of AS and gene expression events relevant to development and disease.............................................. 134

References ..................................................................................................................... 137

Appendices ..................................................................................................................... 167
List of Figures

Figure 1-1 Core splicing signals and stepwise spliceosome assembly........................................5

Figure 1-2 Types of AS events. ........................................................................................................11

Figure 1-3 Mechanisms of AS regulation..........................................................................................23

Figure 1-4 Schematic structure of fluorescent and luminescent splicing reporters.........................31

Figure 1-5 Examples of functionally important AS events in ESCs. ..............................................37

Figure 2-1 Identification of ESC-differential AS events in human.....................................................59

Figure 2-2 Identification of ESC-differential AS events in mouse. ..................................................60

Figure 2-3 Expression profiling of human splicing factor genes. ......................................................61

Figure 2-4 Expression profiling of mouse splicing factor genes......................................................62

Figure 2-5 MBNL proteins regulate the ESC-specific switch exon in FOXP1.................................63

Figure 2-6 Comparison of effects of siRNA knockdown of MBNL1, MBNL2, and MBNL3 on ESC-differential AS..................................................................................................................64

Figure 2-7 MBNL proteins regulate approximately half of ESC-differential AS events..................66

Figure 2-8 Analysis of ESC-differential AS events affected by knockdowns of MBNL1 and MBNL2 in human 293T and mouse C2C12 cells. .........................................................................................67

Figure 2-9 Overexpression of Mbnl1 and Mbnl2 in ESCs promotes differentiated cell-like AS patterns for ESC-differential AS events..................................................................................................69

Figure 2-10 MBNL protein acts widely and directly to regulate ESC-differential AS.......................70

Figure 2-11 Mapping of Mbnl1 CLIP-Seq tags in the Foxp1 gene and sequence motifs enriched under Mbnl1 CLIP-Seq clusters. ..........................................................................................................71
Figure 2-12 Overexpression of Mbnl proteins suppresses pluripotency factors and promotes expression of differentiation markers. .............................................................................................................................. 73

Figure 2-13 Knockdown of Mbnl proteins enhances reprogramming efficiency and kinetics. .... 76

Figure 2-14 MBNL-regulated ESC-differential AS programs are important for transition to transgene independence. .............................................................................................................................. 76

Figure 2-15 Knockdown of MBNL1 enhances human somatic cell reprogramming ................ 77

Figure 2-16 Characterization of human iPSC lines generated from MBNL1-depleted fibroblasts. ........................................................................................................................................ 78

Figure 2-17 Model for the role of MBNL proteins in the regulation of ESC-differential AS, pluripotency, and iPSC reprogramming. .............................................................................................................................. 79

Figure 3-1 Systematic Parallel Analysis of endogenous RNA regulation coupled to barcode Sequencing (SPAR-Seq) ........................................................................................................................................ 99

Figure 3-2 Flowchart outlining pipeline for analysis of SPAR-Seq data. ................................. 100

Figure 3-3 Overall assessment of the SPAR-Seq strategy. ........................................................ 101

Figure 3-4 RT-PCR validations of PSI changes in CGR8 and N2A screens. ......................... 102

Figure 3-5 Systematic identification of splicing regulators in CGR8 and N2A screens......... 103

Figure 3-6 Correlations of AS changes in the CGR8 screen reveal protein complexes and pathways. ........................................................................................................................................ 105

Figure 3-7 Correlations of AS changes in the N2A screen reveal protein complexes and pathways. ........................................................................................................................................ 106

Figure 3-8 Complexes and pathways regulate AS in CGR8 and N2A cells. ......................... 107

Figure 3-9 The functional implication of U2 snRNP-associated factors. ............................... 109

Figure 3-10 Novel positive and antagonistic effects on AS. ..................................................... 111
Figure 3-11 Chromatin and transcription factors frequently regulate AS in N2A cells...........113

Figure 3-12 The functional implication of RNA Pol II-related factors......................................114

Figure 3-13 AS regulation by C2H2 ZnF proteins. ........................................................................116

Figure 3-14 The C2H2 ZnF factor Zfp871 regulates neural-enriched exons and microexons....117

Figure 3-15 RT-PCR validations of AS regulation by Zfp871. ............................................................118

Figure 3-16 GO-enrichment analysis of genes harbouring changing exons and microexons upon knockdown of Zfp871. ..................................................................................................................119

Figure 3-17 Zfp871 binds RNA sequences proximal to its target exons.................................120

Figure 3-18 The effect of Nacc1 on ESC-differential AS regulatory networks. ....................122

Figure 3-19 RNA-Seq analysis reveals extensive effect of Nacc1 on ESC-differential AS. ......122

Figure 3-20 Nacc1 binds proximal to the transcription start site of Mbnl1 and other target genes. ..................................................................................................................................................123

Figure 3-21 Nacc1 regulates AS through direct binding to RNA. ..............................................124
List of Appendices

Appendix 1 Supporting Data for Chapter 2 ................................................................. 167

Appendix 2 Supporting Data for Chapter 3 ................................................................. 175

Appendix 3 Published Manuscripts .......................................................................... 180
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3'ss</td>
<td>3' splice site</td>
</tr>
<tr>
<td>5'ss</td>
<td>5' splice site</td>
</tr>
<tr>
<td>AD</td>
<td>Alzheimer disease</td>
</tr>
<tr>
<td>AP</td>
<td>alkaline phosphatase</td>
</tr>
<tr>
<td>AP-MS</td>
<td>immunoaffinity purification coupled to mass spectrometry</td>
</tr>
<tr>
<td>Arglu1</td>
<td>arginine and glutamate rich 1</td>
</tr>
<tr>
<td>AS</td>
<td>alternative splicing</td>
</tr>
<tr>
<td>ASD</td>
<td>autism spectrum disorder</td>
</tr>
<tr>
<td>BFP</td>
<td>blue fluorescent protein</td>
</tr>
<tr>
<td>BPS</td>
<td>branch point sequence</td>
</tr>
<tr>
<td>BSA</td>
<td>bovine serum albumin</td>
</tr>
<tr>
<td>Chem-CLIP</td>
<td>chemical cross-linking and isolation by pull-down (Chem-CLIP)</td>
</tr>
<tr>
<td>ChIP</td>
<td>chromatin immunoprecipitation</td>
</tr>
<tr>
<td>circRNA</td>
<td>circular RNA</td>
</tr>
<tr>
<td>CLIP</td>
<td>UV-crosslinking immunoprecipitation</td>
</tr>
<tr>
<td>CLIP-Seq/HITS-CLIP</td>
<td>UV-crosslinking immunoprecipitation coupled to high-throughput sequencing</td>
</tr>
<tr>
<td>CRISPR</td>
<td>Clustered Regularly Interspaced Short Palindromic Repeats</td>
</tr>
<tr>
<td>cRPKM</td>
<td>corrected reads per kilobase pair and million mapped reads</td>
</tr>
<tr>
<td>CTD</td>
<td>carboxy-terminal domain</td>
</tr>
<tr>
<td>cTNT</td>
<td>cardiac troponin T</td>
</tr>
<tr>
<td>DM</td>
<td>myotonic dystrophies</td>
</tr>
<tr>
<td>DMEM</td>
<td>Dulbecco's Modified Eagle Medium</td>
</tr>
<tr>
<td>Dox</td>
<td>doxycycline</td>
</tr>
<tr>
<td>Dscam</td>
<td>Down syndrome cell adhesion molecule</td>
</tr>
<tr>
<td>EB</td>
<td>embryoid body</td>
</tr>
<tr>
<td>ECC</td>
<td>embryonal carcinoma cell</td>
</tr>
<tr>
<td>EEJ</td>
<td>exon-exon junction</td>
</tr>
<tr>
<td>EGFP</td>
<td>enhanced green fluorescent protein</td>
</tr>
<tr>
<td>EJC</td>
<td>exon junction complex</td>
</tr>
<tr>
<td>ESC</td>
<td>embryonic stem cell</td>
</tr>
<tr>
<td>ESE</td>
<td>exonic splicing enhancer</td>
</tr>
<tr>
<td>ESS</td>
<td>exonic splicing silencer</td>
</tr>
<tr>
<td>EST</td>
<td>expressed sequence tag</td>
</tr>
<tr>
<td>FACS</td>
<td>fluorescence-activated cell sorting</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>FBS</td>
<td>Fetal Bovine Serum</td>
</tr>
<tr>
<td>FDR</td>
<td>false discovery rate</td>
</tr>
<tr>
<td>FGF4</td>
<td>fibroblast growth factor 4</td>
</tr>
<tr>
<td>FISSEQ</td>
<td>fluorescent in situ RNA sequencing</td>
</tr>
<tr>
<td>FOXP1-ES</td>
<td>ESC-specific isoform of FOXP1</td>
</tr>
<tr>
<td>G runs</td>
<td>poly-guanine sequences</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>hnRNP</td>
<td>heterogeneous nuclear ribonucleoprotein</td>
</tr>
<tr>
<td>IBC</td>
<td>intron binding complex</td>
</tr>
<tr>
<td>iCLIP</td>
<td>individual nucleotide resolution CLIP</td>
</tr>
<tr>
<td>iPSC</td>
<td>induced pluripotent stem cell</td>
</tr>
<tr>
<td>ISE</td>
<td>intronic splicing enhancer</td>
</tr>
<tr>
<td>ISS</td>
<td>intronic splicing silencer</td>
</tr>
<tr>
<td>KH</td>
<td>K homology</td>
</tr>
<tr>
<td>LIF</td>
<td>leukemia inhibitory factor</td>
</tr>
<tr>
<td>LUMIER</td>
<td>luminescence-based mammalian interactome mapping</td>
</tr>
<tr>
<td>MALAT1</td>
<td>metastasis-associated lung adenocarcinoma transcript 1</td>
</tr>
<tr>
<td>MBNL</td>
<td>Muscleblind-like</td>
</tr>
<tr>
<td>MEF</td>
<td>mouse embryonic fibroblast</td>
</tr>
<tr>
<td>N2A</td>
<td>Neuro2A</td>
</tr>
<tr>
<td>Nacc1</td>
<td>Nucleus accumbens associated 1</td>
</tr>
<tr>
<td>ncRNA</td>
<td>non-coding RNA</td>
</tr>
<tr>
<td>NGS</td>
<td>next generation sequencing</td>
</tr>
<tr>
<td>NMD</td>
<td>nonsense-mediated mRNA decay</td>
</tr>
<tr>
<td>NP</td>
<td>neural progenitor</td>
</tr>
<tr>
<td>PAR-CLIP</td>
<td>photoactivatable ribonucleoside-enhanced CLIP</td>
</tr>
<tr>
<td>PAS</td>
<td>Periodic Acid Schiff</td>
</tr>
<tr>
<td>PBase</td>
<td>piggyBac transposase</td>
</tr>
<tr>
<td>PCA</td>
<td>principal component analysis</td>
</tr>
<tr>
<td>pol II</td>
<td>RNA polymerase II</td>
</tr>
<tr>
<td>POZ/BTB</td>
<td>Pox virus and Zinc finger/Bric-a-brac Tramtrack Broad complex</td>
</tr>
<tr>
<td>PPI</td>
<td>protein-protein interaction</td>
</tr>
<tr>
<td>PPlases</td>
<td>peptidyl-prolyl cis/trans isomerases</td>
</tr>
<tr>
<td>PPT</td>
<td>polypyrimidine tract</td>
</tr>
<tr>
<td>pre-mRNA</td>
<td>precursor messenger RNA</td>
</tr>
<tr>
<td>PRMT</td>
<td>protein arginine methyltransferase</td>
</tr>
<tr>
<td>PSI</td>
<td>percent spliced in</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>PTBP</td>
<td>polypyrimidine tract binding protein</td>
</tr>
<tr>
<td>PTC</td>
<td>premature termination codon</td>
</tr>
<tr>
<td>qPCR</td>
<td>quantitative PCR</td>
</tr>
<tr>
<td>RBD</td>
<td>RNA binding domain</td>
</tr>
<tr>
<td>RBFOX</td>
<td>RNA-binding fox protein</td>
</tr>
<tr>
<td>RBP</td>
<td>RNA binding protein</td>
</tr>
<tr>
<td>RFP</td>
<td>red fluorescent protein</td>
</tr>
<tr>
<td>RGG</td>
<td>Arg-Gly-Gly</td>
</tr>
<tr>
<td>RIPA</td>
<td>radioimmunoprecipitation assay</td>
</tr>
<tr>
<td>RNA-Seq</td>
<td>RNA sequencing</td>
</tr>
<tr>
<td>RNAi</td>
<td>RNA interference</td>
</tr>
<tr>
<td>RNP</td>
<td>ribonucleoprotein</td>
</tr>
<tr>
<td>RPM</td>
<td>reads per million read</td>
</tr>
<tr>
<td>RRM</td>
<td>RNA recognition motif</td>
</tr>
<tr>
<td>RS</td>
<td>Arg/Ser-repeat</td>
</tr>
<tr>
<td>rtTA</td>
<td>reverse tetracycline transactivator</td>
</tr>
<tr>
<td>SF1</td>
<td>splicing factor 1</td>
</tr>
<tr>
<td>shRNA</td>
<td>short hairpin RNA</td>
</tr>
<tr>
<td>siRNA</td>
<td>small interfering RNA</td>
</tr>
<tr>
<td>SMA</td>
<td>spinal muscular atrophy</td>
</tr>
<tr>
<td>SMC</td>
<td>structural maintenance of chromosomes</td>
</tr>
<tr>
<td>SMN</td>
<td>survival of motor neurons</td>
</tr>
<tr>
<td>snoRNA</td>
<td>small nucleolar RNA</td>
</tr>
<tr>
<td>snRNA</td>
<td>small nuclear RNA</td>
</tr>
<tr>
<td>snRNP</td>
<td>small nuclear ribonucleoprotein</td>
</tr>
<tr>
<td>SPAR-Seq</td>
<td>Systematic Parallel Analysis of endogenous RNA regulation coupled to barcode Sequencing</td>
</tr>
<tr>
<td>SR protein</td>
<td>serine/arginine-rich protein</td>
</tr>
<tr>
<td>SRE</td>
<td>splicing regulatory element</td>
</tr>
<tr>
<td>SSMD</td>
<td>Strictly Standardized Mean Difference</td>
</tr>
<tr>
<td>TRAP</td>
<td>translating ribosome affinity purification</td>
</tr>
<tr>
<td>TTP</td>
<td>thiamine pyrophosphate</td>
</tr>
<tr>
<td>U2AF</td>
<td>U2 auxiliary factor</td>
</tr>
<tr>
<td>UTR</td>
<td>untranslated region</td>
</tr>
<tr>
<td>ZnF</td>
<td>zinc finger</td>
</tr>
</tbody>
</table>
Chapter 1
1 Introduction

A few years after the landmark discovery of the DNA double-helix structure (Franklin and Gosling, 1953; Watson and Crick, 1953; Wilkins et al., 1953), Francis Crick proposed the “central dogma” of molecular biology, which provided a simplistic model for the flow of genetic information between nucleic acids (DNA and RNA) and proteins (Crick, 1958). Since then, numerous technological advances and groundbreaking studies have further transformed our understanding of gene expression, which is now viewed as a highly complex and coordinated process, involving extensive physical and functional interactions between multiple gene regulatory pathways (Braunschweig et al., 2013; Maniatis and Reed, 2002; Moore and Proudfoot, 2009). In eukaryotic gene expression, splicing, in particular alternative splicing (AS), is a key and distinct layer of regulation at the stage of RNA processing (Blencowe, 2006; Licatalosi and Darnell, 2010). Its complexity, function, regulation, as well as integration with orthogonal gene regulatory layers will be discussed in detail in this chapter.

1.1 Pre-mRNA splicing and the spliceosome

RNA splicing was first discovered in 1977, independently by Phillip Sharp and Richard Roberts (Berget et al., 1977; Chow et al., 1977). When visualizing RNA-DNA hybrids by electron microscopy, both research groups found that adenoviral mRNAs were complementary to discontinuous segments on the viral genome. They proposed that certain segments from a long precursor messenger RNA (pre-mRNA) are ligated together to generate the mature mRNA, while intervening sequences are excised during the processing. Shortly thereafter, the split gene structure was also shown to be common in eukaryotes, and the terms “exons” and “introns” were introduced to define the sequences encoding the final mRNAs and the noncoding intervening sequences, respectively (Gilbert, 1978). Splicing has profound implications in biology and medicine. Since the initial discovery, remarkable progress has been made to understand how splicing occurs as well as its regulation and dysregulation. In this section, I briefly describe the mechanism of pre-mRNA splicing and highlight the important aspects of spliceosome assembly, dynamics, and catalysis.
1.1.1 Mechanism of pre-mRNA splicing

As eukaryotic genes frequently contain introns, ensuring the efficiency, fidelity, and regulation of pre-mRNA splicing is of crucial importance to gene expression. Splicing consists of two sequential transesterification reactions that are catalyzed by the spliceosome, a highly dynamic macromolecular ribonucleoprotein (RNP) machine. Spliceosome assembly is mediated by cis-acting splicing signals in the pre-mRNA and involves intricate RNA-RNA, RNA-protein, and protein-protein interactions (Matlin and Moore, 2007; Wahl et al., 2009).

The core splicing signals are a set of conserved short sequences present in the introns (Figure 1-1A). The 5′ splice site (5′ss) and 3′ splice site (3′ss) mark the 5′ and 3′ ends of each intron, respectively. The branch point sequence (BPS) is typically located ~15-50 nucleotides upstream of the 3′ss and in metazoans, the polypyrimidine tract (PPT) is located between the BPS and 3′ss. In contrast to the situation in budding yeast *Saccharomyces cerevisiae*, the consensus sequences of core splicing signals are highly degenerate in metazoans, where AS is prevalent (Lee and Rio, 2015). This feature provides great potential for regulatory flexibility. Additional auxiliary cis-regulatory elements can further facilitate efficient and accurate recognition/selection of splice sites by the spliceosome (see section 1.4.1.2 for more details).

The spliceosome is one of the most complicated machines in the cell. It is composed of five uridine-rich small nuclear RNAs (U1, U2, U4, U5, and U6 snRNAs) and over 170 different proteins (Wahl et al., 2009). It exhibits remarkable compositional and conformational dynamics at different stages of the splicing cycle. The snRNAs are associated with a common set of core factors (seven Sm or LSm ring proteins) and a variable number of particle-specific proteins to form the U1, U2, U4/U6, and U5 small nuclear ribonucleoprotein (snRNP) particles. These are the main subunits essential for spliceosome assembly and catalysis (Jurica and Moore, 2003; Matera and Wang, 2014). A plethora of non-snRNP proteins, such as serine/arginine-rich (SR) and heterogeneous nuclear ribonucleoprotein (hnRNP) families of proteins (see sections 1.4.3.2 and 1.4.3.3 for further details), DExD/H-type RNA-dependent ATPases/helicases, peptidyl-prolyl cis/trans isomerases (PPIases), kinases, and phosphatases, can significantly affect the recruitment and stability of snRNPs and further facilitate the dynamic rearrangements of the spliceosome. In addition to the major (U2-dependent) spliceosome, which is responsible for removing >99.5% of introns, some metazoans and plants contain a second, minor (U12-
dependent) spliceosome to excise relatively rare U12-type introns (Turunen et al., 2013). The minor spliceosome is composed of the U11, U12, and U4atac/U6atac snRNPs that are distinct from but functionally analogous to the U1, U2, and U4/6 snRNPs in the major spliceosome, whereas the U5 snRNP is shared between the two types of spliceosomes.

A dynamic network of snRNA-snRNA and snRNA-pre-mRNA interactions is required during splicing, which plays key roles in spliceosome assembly, activation, and catalytic reactions (Wahl et al., 2009). Pre-mRNA splicing by the spliceosome shares striking mechanistic similarities to the self-splicimg reactions of group II introns, and they are thought to be evolutionarily related (Robart et al., 2014; Toor et al., 2008). Moreover, it has been shown that U2 and U6 snRNAs can catalyze splicing-like reactions in the absence of protein components, albeit at lower efficiency (Valadkhan and Manley, 2001; Valadkhan et al., 2009; Valadkhan et al., 2007). Taken together, converging lines of evidence strongly support that the catalytic core of the spliceosome is RNA-based (Fica et al., 2013). Nevertheless, the protein components of the spliceosome are indispensable, and extensive interplay between RNA and proteins is essential for splicing catalysis at the active site (Wahl et al., 2009). For example, splicing defects caused by RNA mutations can be suppressed by rescuing mutations in the U5 snRNP protein Prp8 (Collins and Guthrie, 1999; Liu et al., 2007; Umen and Guthrie, 1996). Recent studies using single-particle cryogenic electron microscopy have provided further insights into the structure and mechanism of the spliceosome, suggesting it is a protein-controlled ribozyme (Hang et al., 2015; Nguyen et al., 2015; Yan et al., 2015). The complex protein composition allows sufficient flexibility of the spliceosome, especially in metazoans, to regulate diverse constitutive and AS events. In contrast, the yeast spliceosome is less complex and contains fewer proteins (Fabrizio et al., 2009).
Figure 1-1 Core splicing signals and stepwise spliceosome assembly.

(A) Consensus sequences of metazoan and yeast core splicing signals. The 5' splice site (5' ss), branch point sequence (BPS), polypyrimidine tract (PPT), and 3' splice site (3' ss) are indicated. R, purine; Y, pyrimidine; N, any nucleotide. (B) Canonical cross-intron assembly of the major spliceosome. Intermediate complexes at different stages of spliceosome assembly are shown with special attention paid to stepwise interactions of snRNPs (U1, U2, U4/U6, and U5) and a few additional key proteins (SF1, U2AF65, and U2AF35).
1.1.2 Assembly of the spliceosome

A large number of in vitro studies using mammalian and yeast extracts have revealed that the spliceosome assembles in an ordered stepwise manner on the pre-mRNA (reviewed in Matlin and Moore, 2007; Wahl et al., 2009). Biochemical approaches have been widely used to isolate and characterize distinct intermediate complexes during the splicing cycle. The consensus view of the stepwise spliceosome assembly pathway is briefly discussed below, mainly focusing on mammalian systems (Figure 1-1B).

Spliceosome assembly begins with the ATP-independent formation of early E complex. It involves the binding of the U1 snRNP to the 5'ss through base-pairing interactions between the U1 snRNA and the pre-mRNA (Kramer et al., 1984; Mount et al., 1983; Seraphin and Rosbash, 1989). This interaction is relatively weak and is stabilized by U1 snRNP proteins (Zhang and Rosbash, 1999) as well as additional factors, such SR proteins (Cho et al., 2011; Staknis and Reed, 1994). In addition, the BPS is recognized by splicing factor 1/branchpoint binding protein (SF1/BBP) at this stage, while the PPT and 3'ss are bound by the U2 auxiliary factor heterodimer, the 65 kDa and 35 kDa subunits (U2AF65 and U2AF35), respectively (Berglund et al., 1998; Kramer and Utans, 1991; Merendino et al., 1999; Wu et al., 1999; Zamore et al., 1992; Zorio and Blumenthal, 1999). These cooperative interactions formed in the E complex are particularly important for the initial splice site recognition and the progression of the subsequent assembly steps.

The U2 snRNP is then recruited to the BPS in an ATP-dependent manner, leading to the formation of the pre-spliceosome or A complex, and SF1/BBP is displaced (Chiara et al., 1996; Konarska and Sharp, 1988; Liu et al., 2001; MacMillan et al., 1994; Rutz and Seraphin, 1999). The base-pairing interactions between the U2 snRNA and the BPS are facilitated by the U2 snRNP-associated heteromeric SF3a and SF3b complexes (Gozani et al., 1996; Kramer et al., 2005) and by U2AF65 (Valcarcel et al., 1996). A short U2 snRNA/BPS intermolecular duplex is formed and causes the branch point adenosine to protrude, which will serve as the nucleophile for the first step of splicing catalysis (Query et al., 1994). The SF3b subunit, SF3b14a/p14, directly contacts the bulged adenosine and it also binds to SF3b155 to form a bridging interaction with U2AF65 (Gozani et al., 1998; Spadaccini et al., 2006; Will et al., 2001).
Following A complex assembly, a U4/U6.U5 tri-snRNP, which is pre-assembled from the U4/U6 and U5 snRNPs, and many additional proteins, including components of the Prp19 and Prp19-related complexes are recruited to form the pre-catalytic B complex (Bindereif and Green, 1987; Deckert et al., 2006; Konarska and Sharp, 1987). Although all of the snRNP components are present at this stage, the B complex is catalytically inactive and must undergo a series of conformational and compositional rearrangements in order to activate the spliceosome, generating the B* complex. The major rearrangements include displacement of the U1 snRNP from the 5′ss; unwinding of the U4/U6 snRNA duplex and dissociation of the U4 snRNP; and formation of base paring interactions between the U6 snRNA and the 5′ss as well as between the U6 and U2 snRNAs (reviewed in Matlin and Moore, 2007; Wahl et al., 2009). Multiple RNA helicases are required for these rearrangements and the activated B* complex is now competent to carry out the first transesterification reaction of splicing. The 2′-hydroxyl of the bulged branch point adenosine nucleophilically attacks the phosphodiester bond at the 5′ss, leading to a free 3′-hydroxyl on the 5′ exon and an intron lariat-3′ exon intermediate.

Upon completion of the first catalytic step of splicing, the C complex is generated. It undergoes additional ATP-dependent rearrangements in transition to the second catalytic step, which is mediated by helicases and other proteins (Matlin and Moore, 2007; Wahl et al., 2009). Within the catalytic center of the C complex, the lariat intermediate is displaced and the 5′ss and 3′ss are repositioned in close proximity (Konarska et al., 2006; Rhode et al., 2006; Umen and Guthrie, 1995). The second transesterification reaction then takes place where the 3′-hydroxyl of the 5′ exon attacks the phosphodiester bond at the 3′ss, resulting in ligation of the two exons and release of the intron lariat. At this point, the exon junction complex (EJC) is deposited upstream of the exon-exon junction, which can affect subsequent metabolism of the spliced mRNA (Tange et al., 2004). Lastly, the spliceosome is disassembled, and the U2, U5, and U6 snRNPs are released and recycled for additional rounds of splicing.

During the stepwise assembly of the highly dynamic spliceosome, the splice sites and BPS are recognized multiple times by both snRNAs and proteins, and a cooperative combination of weak interactions is involved in the formation and stability of the intermediate complexes (Wahl et al., 2009). Together, these features greatly contribute to the fidelity of the splicing process and meanwhile, offer remarkable flexibility for regulation at different stages along the spliceosome assembly pathway. Single molecule fluorescence microscopy has been applied to real-time
analysis of splicing of individual pre-mRNAs in vitro and provided further support for the stepwise model of spliceosome assembly (Abelson et al., 2010; Crawford et al., 2008; Hoskins et al., 2011). In addition, results from chromatin immunoprecipitation (ChIP) and live cell imaging experiments have demonstrated the sequential interactions of individual snRNPs with the pre-mRNA, confirming that the spliceosome assembles in a similar manner in vivo (Gornemann et al., 2005; Huranova et al., 2010; Lacadie and Rosbash, 2005). However, as suggested by other observations, alternative spliceosome assembly pathways different from the canonical stepwise model may also exist. For example, a penta-snRNP complex containing all snRNPs and the Prp19 complex was isolated from yeast extracts, indicating that the spliceosome is recruited to the pre-mRNA in a pre-assembled form (Stevens et al., 2002). In the penta-snRNP model, the various intermediate complexes may reflect the stepwise stabilization instead of the stepwise recruitment of different snRNPs (Brow, 2002). In higher eukaryotes, most genes contain relatively short exons that are separated by long introns and therefore, spliceosomal components first assemble across exons (referred to as exon definition; Berget, 1995). Subsequently, the exon definition complex is converted to the intron definition complex and several studies have indicated that the transition step has an important impact on regulated splicing events (Bonnal et al., 2008; House and Lynch, 2006; Schneider et al., 2010; Sharma et al., 2008). Future studies will be required to fully understand the mechanism of this poorly characterized cross-exon to cross-intron transition.

1.2 Alternative splicing

In 1978, soon after the discovery of the split gene structure, Walter Gilbert speculated that the presence of introns could accelerate evolution (Gilbert, 1978). Notably, he hypothesized that multiple protein variants could be produced from a single pre-mRNA by joining exons in different combinations to evolve modified or novel gene functions, a phenomenon now known as AS. Gilbert also proposed that AS could be regulated in a cell-specific manner, potentially contributing to cell differentiation. These revolutionary hypotheses have since been confirmed by numerous subsequent studies. AS patterns were first observed in several viral gene transcripts (Berk and Sharp, 1978a, b). Shortly thereafter, AS was discovered in cells; for example, in the transcripts of immunoglobulin IgM proteins (Alt et al., 1980; Early et al., 1980; Rogers et al., 1980). Structurally and functionally distinct isoforms, which were generated through alternative
splice site selection, were identified during B cell differentiation and found to encode the membrane-bound and secreted forms of IgM proteins.

Together with conventional approaches, advances in high-throughput methodologies have provided further insights into the prevalence, complexity, and importance of AS. Following the sequencing of genomes of model organisms, such as *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*, it was realized that these organisms have comparable numbers of protein-coding genes, which do not scale with their cellular complexity (Adams et al., 2000; Consortium, 1998; Lander et al., 2001). This unanticipated observation sparked an expanded interest in AS as a crucial mechanism for gene regulation, which dramatically increases the transcriptomic and proteomic diversity of complex metazoans from a limited repertoire of genes. The advent of transcriptome profiling technologies along with computational tools has paved the way for the global identification of AS events in different cell and tissue types of diverse species (Blencowe, 2006; Lee and Rio, 2015; Xiao and Lee, 2010). Recent high-throughput RNA sequencing (RNA-Seq) studies have estimated that transcripts from more than 95% of human multi-exon genes are alternatively spliced (Pan et al., 2008; Wang et al., 2008a). In contrast, transcripts from ~60% and ~25% of genes undergo AS in flies and worms, respectively (Graveley et al., 2011; Ramani et al., 2011). In addition, deep sequencing of organ transcriptomes from several vertebrate species has revealed extensive variation of AS complexity with the highest complexity observed in primates, further supporting that AS contributes in a major way to phenotypic diversity among species (Barbosa-Morais et al., 2012; Merkin et al., 2012). There are different types of AS events and they vary in frequency (Figure 1-2). In higher eukaryotes, cassette alternative exons are the most common type, which accounts for ~40% of AS events. Other types of AS include alternative 5′ and 3′ splice sites, intron retention, mutually exclusive alternative exons, alternative promoter usage, and alternative polyadenylation. More recently, systematic analysis of AS by RNA-Seq has expanded a specific class of AS events called microexons, which were frequently missed or underestimated in previous transcriptome profiling studies (Irimia et al., 2014; Li et al., 2015). Microexons are characterized as ultra-short exons (typically 3-27 nucleotides in length) and are highly conserved, showing strong enrichment for neural-specific inclusion patterns. Lastly, complex AS patterns have also been reported including extreme cases; for example, *Down syndrome cell adhesion molecule (Dscam)*
and neurexin transcripts (Missler and Sudhof, 1998; Park and Graveley, 2007), which can produce thousands of isoforms to diversify the repertoires of proteins.

The realization of the prevalence and general importance of AS has stimulated increased efforts toward understanding the physiological outcomes and functional significance of AS events. Despite the frequent species and lineage specificity of AS, many AS events are conserved and regulated in a cell-, tissue-, and developmental stage-specific manner (Barbosa-Morais et al., 2012; Merkin et al., 2012). These conserved AS events more often preserve reading frame to generate alternative protein isoforms (Sugnet et al., 2006; Xing and Lee, 2005a; Xing and Lee, 2005b; Yeo et al., 2005). The sequences surrounding these alternative exons are also strongly enriched in splicing regulatory motifs. Research over the past few decades has demonstrated that AS plays critical roles in a wide range of biological processes, including proliferation, apoptosis, pluripotency, differentiation, development, and response to stresses (Kalsotra and Cooper, 2011; Kelemen et al., 2013). Misregulation of AS is frequently associated with various human diseases, particularly neurological disorders, muscular dystrophies, and cancer (Cooper et al., 2009; Daguenet et al., 2015; David and Manley, 2010). AS can impact these processes in many different ways and in the following section, some important examples and principles of AS functions will be highlighted.
1.3 Function of alternative splicing

AS has the capacity to greatly diversify gene function and regulatory potential (Blencow, 2006; Irimia and Blencow, 2012; Kalsotra and Cooper, 2011; Kelemen et al., 2013). For example, AS events can insert or delete functionally defined domains and critical regions in proteins to influence diverse aspects of protein functions, including binding properties to proteins, nucleic acids or small ligands, enzymatic activities, subcellular localization, and post-translational modifications (reviewed in Kelemen et al., 2013). The effects of AS changes range from a complete loss of function or acquisition of a novel function to relatively modest modifications. Notably, a substantial fraction of AS events can introduce premature termination codons (PTCs), which result in reduction of transcript abundance by the process of nonsense-mediated mRNA decay (NMD) (Lareau et al., 2007a; Lewis et al., 2003). AS coupled to NMD (AS-NMD) has
been reported to regulate the expression of a variety of genes, especially splicing factors and core spliceosomal proteins (Lareau et al., 2007b; Ni et al., 2007; Saltzman et al., 2008; Saltzman et al., 2011 for examples). Moreover, AS can occur both in translated and untranslated regions (UTRs) of transcripts to affect sequence elements that are required for the regulation of mRNA stability, 3’ end processing, transport, and translation (Braunschweig et al., 2013; Licatalosi and Darnell, 2010).

In addition to individual examples characterized by focused studies, global analyses using experimental and computational approaches have provided further functional insights into regulated AS. When systematically investigating the functional properties of tissue-regulated AS events, it was revealed that these events are significantly underrepresented in functionally defined protein domains, but are highly enriched in intrinsically disordered regions located on the surface of proteins (Buljan et al., 2012; Ellis et al., 2012; Weatheritt et al., 2012). These regions often contain short linear motifs and are known to mediate protein-protein interactions (PPIs) and ligand interactions (Dunker et al., 2008). Remarkably, employing a high-throughput PPI assay, it was observed that approximately one third of the analyzed neural-specific alternative exons positively or negatively affected partner interactions and that proteins containing these exons tended to be central in interaction networks (Ellis et al., 2012). Taken together, these results strongly support the widespread role of regulated AS in remodeling interactions to establish tissue-specific PPI networks. Other studies have indicated that tissue-regulated alternative exons are enriched in post-translational modification sites, such as phosphorylation sites (Buljan et al., 2012; Merkin et al., 2012; Zhang et al., 2010).

Another important finding stemming from high-throughput AS analysis is that cell-, tissue- and developmental stage-regulated alternative exons are typically organized into functionally and biologically coherent networks (Calarco et al., 2011; Jangi and Sharp, 2014; Kalsotra and Cooper, 2011; Licatalosi and Darnell, 2010). For example, analysis of AS in brains from wild-type and Nova-null mice revealed that Nova-2-regulated neural-specific AS events were significantly enriched in genes functionally associated with synapse activity and axon guidance (Ule et al., 2005). Groups of genes containing tissue-regulated alternative exons generally do not overlap with those genes differentially regulated in the same contexts at the transcription level, suggesting that networks of coordinated AS events represent a distinct layer of gene regulation.
and act in parallel with other gene regulatory pathways to control cell- and tissue-specific functions (Pan et al., 2004).

Given the profound functional effects of AS in both physiological and pathological contexts, it is particularly important to understand the diverse mechanisms underlying AS regulation as well as how it is coordinated with other layers of gene regulation, which will be discussed in detail in the sections to follow.

### 1.4 Mechanisms of alternative splicing regulation

#### 1.4.1 Cis-acting regulatory elements

##### 1.4.1.1 Core splicing signals

The core splicing signals, the 5’ss, BPS, PPT, and 3’ss, not only direct constitutive splicing, but also play a key role in regulating AS by affecting the efficiency of spliceosome formation. As already mentioned, in metazoans, the poor conservation and higher degeneracy of these core signals have been speculated as a key feature responsible for the complexity and regulatory flexibility of AS (Lee and Rio, 2015). Consistent with this, several studies have reported that the splice sites of alternatively spliced exons significantly more often deviate from the consensus sequences than those of constitutively spliced exons, resulting in weaker or suboptimal splice sites that are inefficiently recognized by the snRNAs and spliceosomal proteins (Itoh et al., 2004; Zheng et al., 2005). Moreover, variations of the BPS and PPT properties, such as the nucleotide composition, length, and position relative to the 3’ss, can also alter the splice site strength and consequently contribute to AS regulation (Chiara et al., 1997; Merendino et al., 1999; Reed, 1989; Schwartz et al., 2008). Highly complex intron-exon architecture is observed for mammalian genes, which typically comprise multiple relatively short exons flanked by much longer introns, ranging from a few hundred to several thousand nucleotides. Therefore, the core splicing signals are necessary but not sufficient for the spliceosome to distinguish authentic pairs of splice sites from a vast number of cryptic or pseudo splice sites (Chasin, 2007). Additional auxiliary sequence elements are required to facilitate the recruitment of spliceosomal components and to ensure the high splicing efficiency, fidelity, and specificity, especially for alternative exons.
1.4.1.2 Exonic/intronic splicing enhancers/silencers

Extensive work has been carried out to identify and characterize additional \emph{cis}-acting splicing regulatory elements (SREs). Depending on their position in the pre-mRNA and function, these short (typically 5-10 nucleotides) and degenerate sequences are conventionally classified as exonic splicing enhancers/silencers (ESEs/ESSs) and intronic splicing enhancers/silencers (ISEs/ISSs) (Matlin et al., 2005) (Figure 1-3). SREs are usually bound by \emph{trans}-acting RNA binding proteins (RBPs) (see section 1.4.3) to promote or inhibit splice site recognition and spliceosome assembly through various mechanisms, including interacting with spliceosomal components, sterically blocking access to splice sites, and modulating splice site competition (Chen and Manley, 2009). While SREs have been implicated in constitutive splicing, they are more highly enriched to regulate alternative exons and to suppress pseudoxons (Wang et al., 2004; Zhang and Chasin, 2004). In addition to many SREs regulating exon skipping and alternative 5' or 3' splice site usage, a specific subset of ESSs are likely involved in the regulation of intron retention (Wang et al., 2006).

Advances in high-throughput experimental assays and the development of sophisticated computational approaches (see section 1.6), together with focused studies, have rapidly expanded our understanding of the general principles by which \emph{cis}-acting elements regulate diverse AS patterns. The activities of SREs are both position- and context-dependent (Wang and Burge, 2008; Xiao and Lee, 2010). For example, Nova binding motifs (YCAY) can function as ESSs when present in exons, but function as ISSs or ISEs when located in upstream or downstream flanking introns, respectively (Ule et al., 2006). In addition, the actions of SREs can be further modulated by the strength of adjacent splice sites, the presence of neighboring \emph{cis}-regulatory elements, or the local structure of the pre-mRNA (see section 1.4.2). For instance, the ISE activity of hnRNP H associated poly-guanine sequences (G runs) are dependent on the 5'ss strength, with the highest splicing enhancing effect when adjacent to intermediate strength 5'ss (Xiao et al., 2009). An increasing amount of evidence suggests that AS is frequently controlled by a combinatorial regulation of multiple enhancer and silencer elements (Chen and Manley, 2009; Xiao and Lee, 2010). These \emph{cis}-elements and their cognate \emph{trans}-acting factors can function in a cooperative or competitive manner to specifically regulate and determine AS networks in different contexts. Furthermore, combinatorial control could also buffer the effect of
mutations and ensure robust splicing outcomes of functionally significant genes during evolution.

One of the ultimate goals in the splicing field is to integrate complex combinations of cis-regulatory features toward a comprehensive splicing code that can reliably predict splicing patterns of any transcript under any physiological or pathological condition from genomic sequence alone. Despite formidable challenges ahead, initial progress has been made, focusing on specific questions. For example, different splicing code models have been applied to predict cell-, tissue-, and species-specific AS patterns as well as the effects of genomic variations and mutations on splicing in various human diseases, including neurological disorders and cancer (Barash et al., 2010; Barbosa-Morais et al., 2012; Rosenberg et al., 2015; Xiong et al., 2015). The integrative network analysis of multiple types of datasets has also been performed to infer an expanded Nova-dependent AS network (Zhang et al., 2010).

1.4.2 RNA secondary structures

Recent discoveries have illustrated the role of RNA secondary structures in the regulation of AS and shed light on a variety of mechanisms involved (see Buratti and Baralle, 2004; McManus and Graveley, 2011 for reviews). First, local pre-mRNA structures can directly affect the accessibility of core splicing signals and additional cis-acting SREs. For example, a stem-loop structure is found at the 5'ss of the human tau exon 10, stabilization of which prevents 5'ss recognition by the U1 snRNP and decreases exon 10 inclusion (Donahue et al., 2006). In addition, the structure context of the fibronectin EDA exon can influence the recruitment of SR proteins by an ESE element (Buratti et al., 2004b). Second, long-range RNA-RNA interactions are proposed to bring distant splice signals into closer proximity to regulate AS. A classical example comes from the complex splicing of the exon 6 cluster of the Drosophila Dscam gene, where competing intronic RNA structures ensure that the 48 exons within the cluster are spliced in a mutually exclusive manner (Graveley, 2005). In a recent large-scale study, long-range interactions have been reported in Rbfox-regulated AS events, suggesting the common usage of this mechanism (Lovci et al., 2013). Third, RNA structures can mediate dynamic interactions with some trans-acting splicing regulators (see below). For instance, the muscleblind-like splicing factor, MBNL1, recognizes and stabilizes a RNA hairpin structure directly upstream of the human cardiac troponin T (cTNT) exon 5, and therefore inhibits the binding of U2AF65 to
the PPT and causes the exclusion of exon 5 (Warf and Berglund, 2007). Lastly, RNA structure elements can potentially regulate AS in response to small molecule binding. In the fungus Neurospora crassa, thiamine pyrophosphate (TTP) binding riboswitches within introns have been shown to affect the AS of transcripts of thiamine metabolism genes in order to control their expression levels (Cheah et al., 2007). The conformation changes of these riboswitches are thus thought to provide an important negative feedback mechanism to regulate TTP synthesis at the AS level. However, it is not known whether riboswitches can also mediate AS regulation in higher eukaryotes and it will be interesting to address this question in future studies.

1.4.3 Trans-acting splicing regulators

1.4.3.1 Core spliceosomal proteins

It is emerging that beyond their general roles in splicing, core spliceosomal components also contribute to the regulation of AS. Recent studies have indicated that changes in the abundance and/or activity of these core factors can specifically affect a subset of AS events. For example, an RNAi screen in Drosophila cells revealed that depletion of core spliceosomal proteins, including multiple components of snRNPs and both U2AF subunits, can lead to altered splicing patterns of specific alternative exons (Park et al., 2004). Moreover, it was shown that these proteins exhibit different expression levels during development and in certain tissues of Drosophila. Similar expression variations were also observed for several snRNP components in mammalian cells and tissues (Grosso et al., 2008). Using a mouse model of spinal muscular atrophy, deficiency of the survival of motor neurons (SMN) protein, which is part of the SMN complex required for snRNP biogenesis, was found to cause tissue-specific alternations in the stoichiometry of snRNAs and splicing patterns, including cases of specific AS events (Zhang et al., 2008b). Another study showed that the core snRNP factor SmB/B', a subunit of Sm ring, can promote the inclusion of a PTC-introducing alternative exon in its own pre-mRNA to auto-regulate its expression (Saltzman et al., 2011). In addition, knockdown of SmB/B' can affect a large network of AS events, which are significantly enriched in genes encoding RNA processing/binding factors. Consistent with their functional importance, many core spliceosomal components are frequently mutated in human diseases, particularly in retinitis pigmentosa and in hematopoietic disorders (Daguenet et al., 2015). These mutations can result in specific effects on splice site choice and AS patterns, which then cause or contribute, at least in part, to the disease phenotypes. More recently, high-throughput screening in HeLa cells has demonstrated distinct regulatory roles of core
spliceosomal components in AS, depending on their recruitment order and duration in the splicing reaction (Papasaikas et al., 2015; Tejedor et al., 2015). In addition, results from the screen described in my thesis work provide further insights into the regulation of AS by the core components of the splicing machinery (see Chapter 3).

In the future, it will be important to systematically identify cis-features that are responsible for the differential sensitivity of splicing events to fluctuations in the concentration and activity of core spliceosomal components. Moreover, studies will be required to fully explore how these components function in conjunction with additional ubiquitously expressed splicing factors, SR proteins and hnRNPs, as well as cell- and tissue-specific splicing regulators, to regulate AS.

1.4.3.2 SR and SR-related proteins

SR and SR-related proteins are a major class of splicing factors in metazoans, which regulate constitutive and alternative splicing (Blencowe, 2000; Lin and Fu, 2007; Long and Caceres, 2009) (Figure 1-3). All members of this superfamily contain a signature protein domain rich in alternating arginine and serine dipeptides, termed the RS domain. SR proteins also have a modular structure consisting of one or two RNA recognition motifs (RRMs). SF2/ASF was the first SR protein shown to complement splicing-deficient HeLa S100 extracts and to alter splice site selection in vitro, supporting its dual function as a constitutive and alternative splicing factor (Ge and Manley, 1990; Krainer et al., 1990a, b). Since then many other classical and additional SR proteins have been identified. In addition to SR proteins, a great number of RS domain-containing proteins are collectively referred as to SR-related proteins, which may or may not contain RRM s or other RNA binding domains (RBDs), such as both U2AF subunits, the U1 snRNP protein U1-70K, and the SRm160/300 splicing coactivators (Blencowe et al., 1999; Lin and Fu, 2007). Some SR-related proteins may possess domains for enzymatic activities, including DEAH box and kinase domains. Although SR and SR-related proteins are ubiquitously expressed in general, evidence has suggested that variations in the relative expression levels and/or activity of these proteins can influence cell- and tissue-specific AS (Hanamura et al., 1998; Zahler et al., 1992).

SR proteins are commonly known as positive regulators of AS and bind to ESEs in a sequence-specific manner through their RRM s. It is believed that RS domains can mediate both protein-protein and protein-RNA interactions to facilitate spliceosome assembly (Hertel and Graveley,
Several mechanistic models have been proposed for the roles of SR proteins in promoting exon inclusion. First, ESE-bound SR proteins can recruit/stabilize the U2AF heterodimer and the U1 snRNP to the upstream 3′ss and the downstream 5′ss, respectively, which facilitates the assembly of spliceosomal complexes across exons, through their RS domains (Graveley et al., 2001; Zahler and Roth, 1995). Second, ESE-bound SR proteins can antagonize the activity of ESS-dependent negative splicing regulators, such as hnRNP proteins (Kan and Green, 1999; Zhu et al., 2001). Lastly, SR proteins can form a protein-protein interaction network with snRNP components and SR-related splicing coactivators to promote cross-intron bridging interactions (Blencowe et al., 2000; Eldridge et al., 1999).

Recent studies have reported that SR proteins can also inhibit alternative exon inclusion through their strong interactions with adjacent constitutive exons or through competition between different SR proteins (Han et al., 2011b; Pandit et al., 2013). Beyond their significant roles in splicing regulation, SR and SR-related proteins participate actively in various other processes, such as transcription, mRNA stability, translation, and non-coding RNA (ncRNA) biogenesis, and are therefore considered as multitasking regulators of gene expression (Anko, 2014).

1.4.3.3 hnRNPs

Heterogeneous nuclear ribonucleoproteins (hnRNPs) are abundant in metazoans and comprise a diverse set of RBPs. They were initially defined as proteins that are associated with nascent nuclear RNAs of high molecular weight (Dreyfuss et al., 1993). While exhibiting considerable structural divergence, hnRNPs typically contain one or more RBDs, such as RRM, quasi-RRMs, or K homology (KH) domains, and auxiliary domains, such as glycine-rich domains or Arg-Gly-Gly (RGG) boxes (Han et al., 2010; Martinez-Contreras et al., 2007). Unsurprisingly, hnRNPs play a wide variety of roles in RNA metabolism, and many members of this protein family have been demonstrated to be involved in splicing regulation (Figure 1-3).

Compared to SR proteins, hnRNPs frequently bind to ESSs/ISSs and negatively regulate AS. Depending on the position and/or context (described above), different mechanisms are utilized to inhibit splice site recognition and spliceosome assembly. For example, binding of hnRNP proteins to silencer elements that are located close to core splicing signals or splicing enhancer elements can sterically block the access of snRNP components and the U2AF subunits, or antagonize the effect of SR proteins and other positive regulators (Buratti et al., 2004a; Expert-
Moreover, binding of hnRNPs, such as hnRNP A1, at high affinity silencer sites can promote cooperative binding and multimerization along the pre-mRNA and inhibit the interaction of spliceosomal components to splicing signals at a distant location (Zhu et al., 2001). Another mechanism proposes that interactions of hnRNPs bound at different sites can loop out the alternative exon and bring the distal pair of splice sites in close proximity (Blanchette and Chabot, 1999; Martinez-Contreras et al., 2006). Ultimately, hnRNP proteins, such as PTBP1 (hnRNP I), can prevent spliceosome assembly during exon and intron definition (Spellman and Smith, 2006). Nevertheless, recent high-throughput analyses have revealed that in addition to inhibiting exon inclusion, hnRNPs can also function as positive regulators of AS (Huelga et al., 2012; Hung et al., 2008; Konig et al., 2010; Rossbach et al., 2014). Future studies are needed to better understand how hnRNP and SR families of proteins act in a combinatorial manner to regulate complex AS decisions in different contexts.

1.4.3.4 Cell- and tissue-specific splicing factors

In addition to widely expressed trans-acting regulators, multiple cell-, tissue-, or developmental stage-specific splicing factors have been identified and their functions have been covered in recent reviews (Calarco et al., 2011; Chen and Manley, 2009; Kalsotra and Cooper, 2011; Raj and Blencowe, 2015; Ye and Blelloch, 2014) (Figure 1-3). Well-characterized examples include neural or brain-specific splicing factors NOVA1/2, PTBP1/2, and nSR100 (also known as SRRM4), epithelial cell-specific splicing factors ESRP1/2 (also known as RBM35a/b), as well as splicing factors such as CELF, RBFOX, and MBNL proteins, which exhibit enriched or highly-restricted expression patterns in particular cells and tissues. Splicing regulators that play significant roles in controlling cell fate, especially pluripotency, differentiation, and reprogramming, will be discussed in detail in section 1.7 and Chapter 2. In the current section, I highlight some general principles and regulatory mechanisms shared between different cell- and tissue-specific splicing factors, mainly using MBNL proteins as an example.

Muscleblind-like (MBNL) proteins are a highly conserved family of tissue-specific and developmentally regulated RNA binding factors, containing tandem CCCH zinc finger (ZnF) domains (Fernandez-Costa et al., 2011; Konieczny et al., 2014). Three paralogs, MBNL1, MBNL2, and MBNL3, have been identified in mammals (Fardaei et al., 2002; Kanadia et al.,
2003). MBNL1 and MBNL2 are expressed in a variety of adult tissues, predominantly in muscle, heart, and brain, and have been shown to regulate specific AS networks during terminal differentiation of these tissues. In contrast, MBNL3 expression is more restricted and primarily in the placenta. More recently, my thesis work has demonstrated that MBNL1 and MBNL2 are differentially expressed between ESCs/iPSCs and differentiated cells/tissues, and that they function as direct negative regulators of a large program of ESC-differential AS events to control ESC pluripotency and iPSC reprogramming (Han et al., 2013; Chapter 2). Therefore, MBNL proteins play multifaceted roles in various developmental transitions. Misregulation of MBNL has been implicated in human disease, most importantly myotonic dystrophies (DM). In DM, sequestration of MBNL proteins by expanded CUG and CCUG repeats leads to loss of MBNL activity and aberrant splicing of their normal RNA targets, switching AS patterns from adult isoforms back to fetal isoforms (Lin et al., 2006; Miller et al., 2000).

Notably, transcripts of MBNL proteins contain several developmentally regulated alternative exons. Extensive AS adds further complexity to the functions of MBNL isoforms, which vary in subcellular localization, RNA binding affinity, dimerization, and splicing activity (Lin et al., 2006; Terenzi and Ladd, 2010; Tran et al., 2011). Moreover, some of these alternative exons, such as MBNL1 exon 5, are subject to autoregulation. Cross-regulation between MBNL proteins and by other cell/tissue-specific splicing factors, such CELF1 and RBFOX2, has also been reported (Kalsotra et al., 2008; Lee et al., 2013; Venables et al., 2013). Similar complexity and regulation have emerged from studies of other splicing factors and RBPs. For instance, a neural-specific exon 10 in Ptbp2 is autoregulated and cross-regulated by other factors, including splicing activator nSR100 and repressor Ptbp1 through AS-NMD (Boutz et al., 2007b; Calarco et al., 2009; Makeyev et al., 2007; Raj et al., 2014; Spellman et al., 2007). Recent high-throughput RNA-Seq studies and UV-crosslinking immunoprecipitation (CLIP) experiments have identified transcriptome-wide MBNL targets in brain, heart, muscle tissues, and myoblasts (Charizanis et al., 2012; Wang et al., 2012a). The resulting nucleotide resolution RNA splicing maps support the position-dependent AS regulation by MBNL proteins. Specifically, binding of MBNL to YGCGY motifs in downstream introns tends to promote exon inclusion, while binding within alternative exons and upstream introns represses exon inclusion. This type of position-dependent activity has also been observed in RNA maps of many other cell/tissue-specific splicing factors (reviewed in Fu and Ares, 2014). One exception is nSR100, which mainly binds to intronic
sequences proximal to the upstream 3’ss to promote exon inclusion (Raj et al., 2014). Together, these factors can act cooperatively or competitively to modulate cell/tissue-specific and developmentally regulated AS networks. Lastly, like most RNA binding splicing factors, MBNL proteins are multifunctional and have profound impacts on numerous additional cellular pathways, such as mRNA localization, stability, polyadenylation, and circular RNA (circRNA) biosynthesis (Ashwal-Fluss et al., 2014; Batra et al., 2014; Wang et al., 2012a).

1.4.4 Post-translational modifications of splicing factors

The function of splicing factors in AS regulation is further affected by a variety of post-translational modifications, including phosphorylation, methylation, acetylation, ubiquitylation, and sumoylation (Heyd and Lynch, 2011; Risso et al., 2012). The best characterized post-translational modification is phosphorylation. For example, phosphorylation of RS domains in SR proteins has been shown to influence their protein-protein and protein-RNA interactions, subcellular localization, stability, and splicing activity (Chen and Manley, 2009; Lai et al., 2003). Different kinases have been identified to phosphorylate SR proteins, such as members of the SRPK and Clk/Sty families of kinases (Colwill et al., 1996; Ngo et al., 2005). Akt, a key mediator of PI3K signaling, and topoisomerase I have also been implicated in SR protein phosphorylation (Patel et al., 2005; Rossi et al., 1996). In addition, phosphatases such as PP1, have been shown to dephosphorylate SR proteins (Misteli and Spector, 1996). Consequently, the outcome of AS can be modulated by the balance between kinase and phosphatase activity. Similar to SR proteins, hnRNP proteins, core spliceosomal factors, and cell/tissue-specific splicing factors are also regulated by widespread post-translational modifications. For instance, protein kinases PKA and PKC have been reported to phosphorylate PTBP1 and CELF1, respectively (Kuyumcu-Martinez et al., 2007; Xie et al., 2003). Moreover, protein arginine methyltransferases (PRMTs) have been linked to AS regulation, through modifying RGG-box containing hnRNP proteins and core snRNP components, Sm proteins (Bedford and Richard, 2005). There is certainly more to learn about the diverse sets of enzymes responsible for post-translational modifications, and the mechanisms by which they regulate splicing factors and AS patterns.
1.4.5 Signal transduction pathways

A large body of recent work has suggested the critical roles of signal transduction pathways in regulating AS (Blaustein et al., 2007; Heyd and Lynch, 2011; Li et al., 2007; Tarn, 2007). As one of the mechanisms, cellular signals can alter post-translational modifications of key splicing factors (described above), which in turn, modulate alternative splice site usage. In a simple manner, signaling cascades can also regulate the expression of splicing regulators to affect AS. For example, T cell activation increased hnRNP LL mRNA and protein levels, which then led to an AS switch in CD45 (Oberdoerffer et al., 2008). Moreover, signaling pathways can influence co-transcriptional regulation of AS (see Section 1.5 for details) through different mechanisms, such as changing RNA polymerase II (Pol II) modifications and chromatin structure. Remarkably, many signal-mediated AS events occur within transcripts that encode cell surface adhesion and signaling proteins, such as CD44, CD45, and NMDARI, to form feedback loops (Lee et al., 2007; Lynch, 2004). Some AS regulation is achieved through extensive cross-talk between different signaling transduction pathways (Blaustein et al., 2007). In the past decades, signal-mediated AS regulation has been implicated in diverse biological processes, including apoptosis, cell cycle, stress response, and neuronal depolarization. Nevertheless, the details of the underlying mechanisms and the functional consequences remain to be fully explored in future studies.
The regulation of AS is governed by the intricate interplay of *cis*-acting RNA sequence elements, RNA secondary structures, and *trans*-acting protein regulators. In addition to core splicing signals, numerous auxiliary *cis*-regulatory elements (ESEs/ESSs, exonic splicing enhancers/silencers; ISEs/ISSs, intronic splicing enhancers/silencers) are required to ensure the efficiency and fidelity of splicing. There are two basic types of protein regulators: ubiquitously expressed splicing factors (e.g. core spliceosomal proteins; SR and SR-related, serine/arginine-rich proteins; hnRNP, heterogenous nuclear ribonucleoproteins) as well as cell- and tissue-specific splicing factors (e.g. MBNL, nSR100/SRRM4, and other RBP, RNA binding proteins). AS is often tightly controlled in a spatio-temporal manner by differentially expressed splicing regulators and/or by the changes in the concentration and activity of ubiquitously expressed splicing regulators, which in turn can be modulated by specific post-translational and signaling pathways. Together, the combined action of these factors can promote or inhibit recognition/selection of particular splice sites by the spliceosome. Furthermore, AS is controlled by complex mechanisms involving coordination and integration with other gene regulatory layers, such as chromatin, transcription, and non-coding RNA regulation.

1.5 Coordination of alternative splicing with other gene regulatory machineries

In addition to the regulation by a wide variety of splicing factors and their cognate *cis*-elements, AS is also controlled by complex, multi-layered mechanisms involving coordinate interactions with other gene regulatory pathways (Braunschweig et al., 2013; Kornblihett et al., 2013; Luco and Misteli, 2011). There is accumulating evidence that different steps of pre-mRNA processing, including capping, splicing, and polyadenylation can occur co-transcriptionally. Two coupling
mechanisms, which are not mutually exclusive, have been proposed to explain the impact of transcription on AS. First, ‘recruitment coupling’ involves the association of splicing factors with the transcription machinery. The Pol II complex, especially the carboxy-terminal domain (CTD) of its largest subunit, serves as a key player in the coupling of the transcription and splicing machineries (Hsin and Manley, 2012; Munoz et al., 2010). The Pol II CTD consists of repeated heptapeptides with the consensus sequence YSPTSPS and is subject to dynamic post-translational modifications to regulate transcription and AS. Several splicing factors have been found to associate with the CTD. Recently, Mediator, a transcriptional coactivator complex, has also been linked to AS (Huang et al., 2012). Its subunit MED23 physically interacts with the splicing factor hnRNP L and regulates a subset of hnRNP L-dependent AS events. Second, ‘kinetic coupling’ model proposes that the CTD and other factors can alter the rate of Pol II elongation to modulate AS by influencing the time by which splice sites and additional regulatory sequences are exposed to the splicing machinery. In support of this model, it has been shown that ultraviolet irradiation results in reduced Pol II elongation rate and, in turn, specifically affects AS of gene transcripts involved in apoptosis (Munoz et al., 2009). Furthermore, a large-scale study has revealed that Pol II elongation inhibition by different mechanisms causes changes in AS events that are highly enriched in genes encoding splicing factors and RBPs (Ip et al., 2011).

Chromatin structure and histone modifications have also been implicated in AS regulation (Luco et al., 2011). Genome-wide analysis has revealed that chromatin features are nonrandomly distributed and in particular, certain histone modifications are enriched over exons even after normalizing for nucleosome occupancy (Andersson et al., 2009; Dhami et al., 2010; Kolasinska-Zwierz et al., 2009; Schwartz et al., 2009; Spies et al., 2009). Consistent with this, evidence supports that histone marks can facilitate the recruitment of splicing regulators to the nascent pre-mRNA through chromatin adaptor complexes in order to modulate splice site selection. This mechanism is exemplified by the tissue-specific mutually exclusive AS of the FGFR2 transcript, where IIIb and IIIc alternative exons are predominantly included in epithelial and mesenchymal cells, respectively. H3K36me3 is enriched over the FGFR2 alternative spliced region in mesenchymal cells and importantly, this histone mark specifically binds to the adaptor protein MRG15, which in turn recruits the splicing factor PTBP1 to exon IIIb to repress its inclusion (Luco et al., 2010). Comparative analysis using RNA-Seq further demonstrated that
H3K36me3/MRG15 preferentially affects a subset of AS events with weak PTBP1-binding sites, suggesting that histone modifications are important to strengthen the effect of splicing factors and their cognate cis-elements on AS regulation. On the other hand, it has been proposed that chromatin structure can impact AS outcomes by controlling Pol II elongation kinetics. For example, the chromatin-remodeling complex SWI/SNF subunit Brm can induce accumulation of Pol II with modified Ser5 phosphorylation of the CTD over the alternative exons of CD44, which blocks Pol II elongation and increases exon inclusion (Batsche et al., 2006). Moreover, the DNA binding protein CTCF has been reported to regulate the inclusion of CD45 exon 5 by causing local Pol II pausing during lymphocyte maturation, and CTCF binding to exon 5 can be inhibited by DNA methylation (Shukla et al., 2011).

Intriguingly, ncRNAs have emerged as a novel layer of regulation for AS. First, a number of studies have documented that microRNAs, such as miR-124, miR-133, and miR-23a/b can regulate the expression of key splicing factors to coordinate networks of AS changes during development and differentiation (Boutz et al., 2007a; Kalsotra et al., 2010; Makeyev et al., 2007). Second, the brain-specific small nucleolar RNA (snoRNA) HBII-52 has been proposed to interact directly with the target pre-mRNA via sequence complementarity to regulate AS (Kishore and Stamm, 2006). The binding of HBII-52 to a silencer element in the serotonin receptor 5-HT2cR exon Vb promotes inclusion, likely through inhibiting the recruitment of a splicing repressor. Third, another study has indicated that small interfering RNAs (siRNAs) targeting sequences near alternative exons can generate closed chromatin structure and reduce Pol II processivity, leading to increased exon inclusion (Allo et al., 2009). Long ncRNAs have also been implicated in regulating AS. For example, metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) can sequester SR proteins in nuclear speckles and impact AS by modulating the levels of active SR proteins (Tripathi et al., 2010). Recently, an antisense long ncRNA from the human FGFR2 locus (asFGFR2) has been shown to regulate tissue-specific FGFR2 AS through influencing histone modifications and the recruitment of chromatin adaptor complexes (Gonzalez et al., 2015).

In summary, these discoveries have profound consequences for our understanding of the diversity and complexity of AS regulation. A major goal for future studies will be to fully explore the molecular connections between transcription factors, chromatin regulators, and ncRNAs as well as the extent to which these regulators are integrated to contribute to cell-,
tissue- and developmental stage-specific AS patterns. Reciprocally, splicing plays an essential part in modulating other gene regulatory pathways, including upstream chromatin organization and transcription, other steps of RNA processing, and downstream events of mRNA turnover and translation (Braunschweig et al., 2013; described above). Therefore, it will be critically important to understand how different layers of regulation are communicated and coordinated with each other to specifically control gene expression in numerous biological processes.

1.6 Experimental and computational approaches facilitating high-throughput alternative splicing analysis

Our understanding of AS has greatly expanded over the past decade. Genome-wide approaches have proved indispensable for gaining global insights into the complexity, evolution, function, and regulation of AS. In the current section, I highlight several key methodologies that have been developed to facilitate high-throughput AS studies.

1.6.1 Microarray profiling and high-throughput RNA sequencing

Various custom microarray platforms have been developed for the large-scale discovery and analysis of AS since 2002. These splicing-sensitive microarrays typically contain oligonucleotide probe sets designed to target exon and/or splice junction sequences, and are therefore capable of monitoring thousands of AS events simultaneously (Calarco et al., 2007a; Moore and Silver, 2008). Using this profiling technology, AS patterns associated with different cell and tissue types, physiological and pathological conditions, as well as species have been characterized (Calarco et al., 2007b; Ip et al., 2007; Johnson et al., 2003; Pan et al., 2004; Pleiss et al., 2007; Yeakley et al., 2002 for examples). Moreover, these microarrays have been employed to detect AS changes upon modulation of the levels of several differentially or ubiquitously expressed splicing factors, systematically uncovering the regulatory targets of the corresponding factor (Blanchette et al., 2005; Calarco et al., 2009; Ule et al., 2005 for examples). These studies have provided the first glimpse into global features and general principles of AS regulation (Ben-Dov et al., 2008; Blencowe, 2006). Interestingly, AS and transcription networks represent distinct layers of gene regulation that function in coordinated manner to establish and/or maintain cell- and tissue-specific characteristics (Le et al., 2004; Pan et al., 2004).
The introduction of RNA-Seq has provided a revolutionary tool for transcriptome profiling and it has since been widely used to characterize RNA regulation and complexity. In RNA-Seq, millions of sequence reads are generated using next generation sequencing (NGS) technologies and subsequently aligned to reference transcriptomes and splice junction databases (Blencow et al., 2009; Wang et al., 2009b). This new approach overcomes several limitations of splicing-sensitive microarrays. It can detect novel exons and splice junctions without prior knowledge, and analyze the transcriptome at unprecedented depth, specificity, and sensitivity. Recent high-throughput RNA-Seq studies have clearly demonstrated that AS is the rule rather than the exception, providing a wealth of information about the complexity and functional implications of AS regulation (Barbosa-Morais et al., 2012; Braunschweig et al., 2014; Irimia et al., 2014; Merkin et al., 2012; Pan et al., 2008; Wang et al., 2008a for examples). In addition, RNA-Seq profiling of whole RNA populations can yield further insights into the communication and integration of AS networks with various other regulatory steps during RNA metabolism (de Klerk and t Hoen, 2015; Han et al., 2011a; Licatalosi and Darnell, 2010). With continuous improvements in the library preparation, read coverage and length, as well as analysis software and pipelines, RNA-Seq is rapidly evolving to power a wide spectrum of applications in RNA research.

1.6.2 Crosslinking and immunoprecipitation coupled to high-throughput sequencing

UV crosslinking immunoprecipitation has been coupled to high-throughput sequencing (CLIP-Seq or HITS-CLIP) to uncover transcriptome-wide protein-RNA interactions (Licatalosi et al., 2008; Yeo et al., 2009). In CLIP assays, living cells or tissues are irradiated with UV light, which induces covalent crosslinks between directly interacting proteins and RNA (within ~1 Ångstrom). Subsequently, protein-RNA adducts are immunoprecipitated under stringent conditions and purified by denaturing gel electrophoresis. The RNA is then released and prepared for high-throughput sequencing, and short sequence tags are mapped back to reference genomes (Zhang and Darnell, 2011). Several modifications to the original CLIP-Seq protocol, including photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP) and individual nucleotide resolution CLIP (iCLIP) have been made to increase the crosslinking efficiency and mapping resolution (Hafner et al., 2010; Huppertz et al., 2014; Konig et al., 2010). These approaches have been applied for transcriptome-wide identification of in vivo RNA targets and
binding sites of a number of splicing factors and other RBPs (reviewed in Fu and Ares, 2014; Konig et al., 2011; Witten and Ule, 2011). Importantly, integrating the information of protein-RNA interactions with the regulatory data from AS profiling can generate RNA splicing maps, which reveal global regulatory principles and functional insights of specific RNA-binding splicing factors. Examples of these were mentioned previously in section 1.4.3.4. In addition to CLIP-based approaches, high-throughput in vitro methods have also been developed for comprehensive analysis of the RNA binding specificity of RBPs, such as RNAcompete and RNA Bind-n-Seq (Lambert et al., 2014; Ray et al., 2009; Ray et al., 2013).

1.6.3 Computational methods for large-scale AS analysis

In parallel with high-throughput experimental technologies, computational and bioinformatic approaches have emerged for large-scale investigations of AS and its regulation. At first, these methods have been utilized to predict the properties of core splicing signals and to systematically identify large repertoires of auxiliary SREs, such as splicing enhancer and silencer elements (Fairbrother et al., 2002; Yeo and Burge, 2004; Yeo et al., 2007a; Zhang and Chasin, 2004 for examples). In addition to general motifs, computational approaches have defined many cis-acting SREs that are associated with cell- and tissue-specific AS events (Brudno et al., 2001; Fagnani et al., 2007; Sugnet et al., 2006 for examples). Together, these studies have deciphered a set of basic rules and principles by which cis-regulatory elements modulate AS decisions (Wang and Burge, 2008; Xiao and Lee, 2010). Furthermore, computational and bioinformatic analyses have served as powerful tools for revealing the origins and evolution of AS (Keren et al., 2010). More recently, advanced machine learning algorithms and integrative network analyses have been developed to infer a splicing code that governs regulated AS (Barash et al., 2010; Rosenberg et al., 2015; Weyn-Vanhentenryck et al., 2014; Xiong et al., 2015; Zhang et al., 2010). For example, using massive microarray profiling data of AS in diverse mouse tissues and a large compendium of RNA features, a deep learning algorithm was developed to determine specific combinations of cis-regulatory features that are maximally predictive of tissue-dependent AS patterns (Barash et al., 2010). Later, a similar approach was applied to infer a human splicing code, which reliably predicts the consequences of genetic variations and mutations on splicing in human disease (Xiong et al., 2015). On the other hand, an integrative method using Bayesian networks was developed to probabilistically model diverse experimental, computational, and evolutionary datasets in order to comprehensively define regulatory networks of specific splicing
factors, such as Nova and Rbfox proteins (Weyn-Vanhentenryck et al., 2014; Zhang et al., 2010). Since AS is also controlled by complex mechanisms involving other layers of gene regulation (Luco and Misteli, 2011), it will be worth considering integrating features associated with transcription, chromatin, and ncRNA regulation into the splicing code in future studies.

1.6.4 High-throughput screening of cis-regulatory elements

Reporter-based screens have been developed to identify cis-acting SREs from a random sequence library. In these methods, random oligonucleotides are inserted into an exon or an intron in splicing reporter constructs (Wang and Wang, 2014). For example, minigene reporters were first used for selection of sequences that enhance alternative exon inclusion during splicing in vitro or in vivo (Coulter et al., 1997; Tian and Kole, 1995). GFP fluorescent reporters have then been introduced to facilitate screening SREs that affect splicing in transfected cells, and different splicing reporters have been specifically designed for screen of ESSs, ESEs, ISSs, and ISEs (Wang et al., 2012b; Wang and Wang, 2014; Wang et al., 2013; Wang et al., 2004). The GFP gene is split into two exons in splicing reporters and the spliced variant encoding functional GFP is produced upon inserting SREs. Therefore, GFP-positive cells can be sorted out by flow cytometry to recover inserted sequences and deduce SREs. More recently, splicing reporters have been combined with high-throughput sequencing to measure the effect of millions of synthetic sequences on splice site usage and subsequently, the vast datasets have enabled the development of a predictive model of AS (Rosenberg et al., 2015). A limitation of these splicing reporter systems is that the function of the identified sequences could be dependent on the reporter or the cell line used. In the future, high-throughput genome editing tools can be potentially used to insert synthetic sequences into endogenous gene loci surrounding alternative exons and to facilitate massively parallel analysis of SREs across diverse cell types. In addition, these approaches can be integrated with complementary screens of trans-acting splicing factors to better understand the mechanisms of AS regulation.

1.6.5 High-throughput screening of trans-acting splicing regulators

Continuous efforts have been made to develop and apply high-throughput screening strategies to uncover trans-acting regulators of AS. To visualize and analyze AS patterns of specific transcripts in cell-based screens, a variety of fluorescent- and luciferase-based splicing reporter minigenes have been designed (Figure 1-4; see below). Using these reporter systems, large-scale
RNA interference (RNAi), cDNA overexpression, and chemical screens have been carried out to identify genetic and small molecule modulators of AS events in different biological contexts. Initially, a GFP-based splicing reporter coupled with an expression cloning strategy was employed, and positive individual cDNA clones that regulate AS were isolated through multiple rounds of division and subdivision of cDNA pools (Kar et al., 2006; Wu et al., 2006). Alternatively, a two-component splicing reporter that utilizes transcriptional synergy to amplify signal and a flow cytometry method were applied to select GFP-positive cDNA clones (Topp et al., 2008). Recent advances in high-throughput assays and high-content screening have greatly facilitated genome-scale screens for AS regulators. For example, using a luciferase-based splicing reporter, a genome-wide cDNA overexpression screen identified ESRP1 and ESRP2 as epithelial-specific splicing factors regulating the mutually exclusive AS event of FGFR2 (Warzecha et al., 2009). Moreover, a genome-wide siRNA screen was performed using fluorescent minigene reporters, which identified many regulators of Bcl-x and Mcl1 AS, revealing the functional link between cell-cycle control and apoptosis AS (Moore et al., 2010). To further increase the sensitivity and specificity, two-color or bichromatic fluorescent splicing reporters were developed and applied for a high-throughput overexpression screen of AS regulators of the post-synaptic density protein Dlg4 (Psd-95) (Zheng et al., 2013). Notably, similar two-color reporter systems were also employed in transgenic model organisms, such as C. elegans, Drosophila, and mouse, to investigate cell- and tissue- specific AS regulation (Norris et al., 2014; Ruan et al., 2015; Takeuchi et al., 2010). In addition to genetic screens, fluorescent- and luciferase-based splicing reporters have been used to facilitate high-throughput screening of small-molecule compounds that affect specific splicing events (Naryshkin et al., 2014; Stoilov et al., 2008; Younis et al., 2010). For example, a recent large-scale chemical screen identified selective compound modifiers of SMN2 AS, which can improve motor function and life expectancy in a mouse model of spinal muscular atrophy (SMA), providing potential therapeutics for SMA patients (Naryshkin et al., 2014).
Figure 1-4 Schematic structure of fluorescent and luminescent splicing reporters.

(A) Single-color splicing reporters. By shifting the reading frame, introducing a stop codon, or splitting the fluorescent protein/luciferase into two inactive parts, these splicing reporters are designed in such a way that only one specific isoform produces fluorescent/luminescent signals.

(B) Two-color splicing reporters. Top, the bichromatic fluorescent splicing reporter contains mutually exclusive GFP and RFP open reading frames, ORF1 and ORF2, respectively. When the alternative exon is skipped, GFP is expressed in frame, followed by a stop codon. Inclusion of the alternative exon shifts the reading frame to express RFP, while GFP is in the non-fluorescent reading frame that lacks stop codons. Bottom, the start codon of the modified GFP (mGFP) is split between two exons and additional AUG codons are mutated. When the alternative exon is skipped, the start codon is formed and GFP is expressed. When included, GFP loses its start codon.
codon and ribosomes initiate translation of the downstream RFP. White boxes, constitutive exons; blue boxes, alternative exons; green boxes, GFP (green fluorescent protein); red boxes, RFP (red fluorescent protein); purple boxes, fluorescent proteins or luciferase (F/L). Start, stop codons, and untranslated regions are also indicated.

Despite these great advances, generating functional splicing reporter minigenes for diverse mammalian AS events can be laborious. Furthermore, splicing reporters may not completely recapitulate physiological regulation of AS, such as complex mechanisms involving chromatin regulation. Consequently, methodologies of high-throughput analysis of endogenous AS changes are particularly valuable for large-scale screening to further expand the repertories of trans-acting splicing regulators. In one study, isoform-specific antibodies and flow cytometry were employed in functional screening with a splicing factor-directed lentiviral short hairpin RNA (shRNA) library in order to identify AS regulators of the transmembrane phosphatase CD45 (Oberdoerffer et al., 2008). Additionally, automated RT-PCR platforms together with capillary electrophoresis have been used to increase the throughput of AS measurements of endogenous transcripts. For example, these methods facilitated defining regulators of cancer- and apoptosis-associated AS events (Papasaikas et al., 2015; Venables et al., 2009; Venables et al., 2008). Along the same lines, a high-throughput quantitative PCR (qPCR) approach was developed to carry out a genome-wide screen using mutant strains of budding yeast, which uncovered novel regulators affecting the splicing efficiency of several yeast transcripts (Albulescu et al., 2012). In contrast to yeast, genome-scale screens to systematically identify regulators of endogenous AS events in mammalian cells are considerably more challenging. Recently, a new PCR-based barcode sequencing strategy has been applied to a genome-wide siRNA screen for AS regulators of the endogenous Fas/CD95, revealing connections between iron homeostasis and AS regulation (Tejedor et al., 2015). In my own work, I have developed a novel high-throughput screen that allows thousands of perturbations (e.g. genetic manipulations and drug treatments) to be simultaneously tested for effects on dozens of endogenous AS and expression events of interest, which will be presented in greater detail in Chapter 3.

In summary, the advent of high-throughput technologies and screening strategies has dramatically transformed AS studies, generating enormous amount of data on AS regulatory networks in a wide variety of contexts. Increasingly sophisticated computational and
bioinformatic methods have greatly facilitated interpreting the large-scale experimental data, which in turn have helped infer general regulatory principles and splicing codes. Furthermore, the mechanistic and functional discoveries stemming from computational approaches can be systematically tested and validated by experimental studies, so that a comprehensive splicing code can be ultimately deciphered with improved predictive power and accuracy.

1.7 Alternative splicing in ESC pluripotency, differentiation and somatic cell reprogramming

Pluripotent stem cells, including ESCs and iPSCs, have the capacity to self-renew indefinitely and differentiate into all cell types of the three germ layers in the body. Therefore, they serve as a remarkable model system for studying embryonic development and hold great promise for regenerative medicine, disease modeling, and drug screening (Tabar and Studer, 2014; Zhu and Huangfu, 2013). Substantial progress has been made in dissecting the molecular mechanisms underlying stem cell biology, primarily focusing on the regulation by transcription factors, signaling pathways, chromatin modifiers, and ncRNAs (Bao et al., 2013; Ng and Surani, 2011; Rinn and Chang, 2012; Young, 2011). Nevertheless, the impact of AS, a widespread mechanism of gene regulation, on stem cells has been largely overlooked until recently. With advances in high-throughput AS analysis, an increasing number of studies have emerged that suggest a prominent role of AS regulatory networks in controlling ESC pluripotency, differentiation, as well as somatic cell reprogramming. In this section, I discuss current and ongoing progress in our understanding of AS complexity, function, and regulation in determining these cell fate decisions.

1.7.1 AS complexity and dynamics in stem cell fate decisions

Early studies of expressed sequence tag (EST) libraries represent the first genome-wide effort to analyze AS patterns in ESCs, revealing a glimpse into the significant contribution of AS towards increasing the transcriptome complexity of ESCs (Lemischka and Pritsker, 2006; Pritsker et al., 2005). Shortly afterwards, microarray and RNA-Seq profiling coupled with bioinformatics approaches have been widely used for global investigation of the AS patterns in human and mouse ESCs, progenitor cells, differentiated cells/tissues, during differentiation of ESCs into distinct lineages in vitro, and across different stages of embryonic development in vivo (Gabut et al., 2011; Kolle et al., 2011; Revil et al., 2010; Salomonis et al., 2009; Salomonis et al., 2010;
Wu et al., 2010; Yeo et al., 2007b for examples). These studies have confirmed the extensive AS complexity in ESCs and demonstrated dynamic AS changes during differentiation and development. Interestingly, it has been shown that splice isoform diversity is high in ESCs, and becomes reduced or restricted upon differentiation (Wu et al., 2010). This phenomenon of “isoform specialization” could indicate that highly diverse and complex AS patterns are required to establish and maintain ESC pluripotency, while restricted expression of specialized isoforms are important for commitment to specific cell lineages during differentiation. Recently, an AS switch has been implicated in reprogramming of somatic cells to iPSCs (Gabut et al., 2011). It is noteworthy that an ESC-like AS pattern can be acquired during reprogramming and play a critical role in efficient iPSC formation.

While a wealth of AS profiling data is being generated, systematic approaches are necessary to build a comprehensive map of coordinated AS events that contribute to fundamental aspects of ESC biology, including self-renewal, pluripotency, and differentiation into specific lineages. In addition, it is important to apply RNA-Seq and complementary methods to globally identify AS patterns that are associated with human and mouse iPSCs as well as different phases during reprogramming, such as initiation, maturation, and stabilization (Golipour et al., 2012; Samavarchi-Tehrani et al., 2010). By comparing human and mouse data, it is possible to distinguish between conserved core AS networks and species-specific AS events in ESC biology and reprogramming. Together, the results from these studies can provide substantial groundwork for further investigation of AS function and regulation in cell fate determinations.

1.7.2 Functionally important AS events in ESCs

AS has been shown to affect genes associated with multiple biological processes in ESCs, including pluripotency, signaling, and cell cycle (Kolle et al., 2011; Salomonis et al., 2010; Wu et al., 2010; Yeo et al., 2007b). Several known examples of AS events that are functionally significant to ESC biology are summarized below (Figure 1-5).

Notably, a few key pluripotency transcription factors are regulated by AS to generate functionally distinct protein isoforms. OCT4 (also known as POU5F1) is a POU domain transcription factor and functions as a master regulator of self-renewal and pluripotency (Tantin, 2013). Three splice isoforms, named OCT4A, OCT4B, and OCT4B1, have been identified (Wang and Dai, 2010). The nuclear localized OCT4A is restrictedly expressed in pluripotent
cells and regulates genes responsible for stemness (Atlasi et al., 2008; Boyer et al., 2005). By contrast, a relatively low level of OCT4B has been detected in both pluripotent and differentiated cells. It exhibits different subcellular localization and DNA binding properties from OCT4A (Lee et al., 2006; Wang et al., 2009a). Moreover, OCT4B is not capable of maintaining self-renewal of ESCs and likely functions in stress response. OCT4B1 is highly expressed in ESCs and embryonal carcinoma cells (ECCs), and decreases upon differentiation (Atlasi et al., 2008). However, its exact localization and function remain to be explored. Oct2 (also known as Pou2f2) is a paralog of Oct4, which is expressed in brain and blood (Tantin, 2013). It has been shown that multiple Oct2 isoforms are generated by AS and they behave differently during ESC differentiation (Theodorou et al., 2009). Interestingly, when overexpressed in mouse ESCs, Oct2.2 and Oct2.4 isoforms can induce and block neuronal differentiation, respectively. Sall4, a member of the spalt-like gene family, is a zinc finger transcription factor essential for ESC pluripotency (Zhang et al., 2006). A recent study has indicated that two splice variants of Sall4, Sall4a and Sall4b, have differential roles, but maintain pluripotency in a cooperative manner in mouse ESCs (Rao et al., 2010). While Sall4b is important for the expression of pluripotency genes, Sall4a predominantly contributes to the regulation of differentiation and patterning genes. Sall4b, but not Sall4a, can partially rescue the ESC state upon knockdown of both isoforms of endogenous Sall4. Tcf3 (also known as Tcf7l1) is a Wnt/β-catenin-responsive transcription factor, which is intimately connected to the core regulatory circuitry of ESCs to control pluripotency and differentiation (Cole et al., 2008). Two isoforms have been reported for mouse Tcf3, the long isoform, Tcf3(l) and the short isoform, Tcf3(s) (Salomonis et al., 2010). While Tcf3(l) is enriched in ESCs and down-regulated following embryoid body (EB) differentiation, the level of Tcf3(s) remains relatively constant. Further analysis has revealed that Tcf3(l) and Tcf3(s) isoforms regulate overlapping but different sets of target genes. In addition, knockdown of each individual isoforms has distinct effect on differentiation pathways, suggesting that the two isoforms might be important for commitment to different lineages.

FOXP1, a member of the forkhead box transcription factor family, has been implicated as a key player in development and cancer (Benayoun et al., 2011). However, its functional importance in ESC pluripotency and differentiation is unclear. A recent study from our lab has identified an evolutionally conserved ESC-specific AS event in FOXP1, which changes its DNA binding specificity and regulatory roles (Gabut et al., 2011). It has been shown that in addition to the
canonical isoform of FOXP1, which is ubiquitously expressed, a novel FOXP1 isoform is expressed in ESCs and rapidly decreased during differentiation. The ESC-specific isoform of FOXP1 (FOXP1-ES), but not the canonical FOXP1 isoform, promotes the expression of core pluripotency transcription factor genes, including OCT4 and NANOG, while repressing genes required for differentiation. Intriguingly, further experiments have demonstrated that the FOXP1-ES isoform is critical for maintaining ESC self-renewal and pluripotency, and is required for efficient somatic cell reprogramming. Therefore, this study provides strong evidence that an AS switch can significantly contribute to stem cell fate control through regulating core transcriptional networks.

In addition to transcription factors, other genes that play an important role in ESC biology are also subject to regulation at the level of AS. Fibroblast growth factor 4 (FGF4) is an autocrine regulator and two splice isoforms, a full-length FGF4 and a novel truncated isoform FGF4si, are co-expressed in human ESCs (Mayshar et al., 2008). In contrast to the continuously expressed FGF4si, the full-length FGF4 is down-regulated following differentiation. Particularly, FGF4 promotes the undifferentiated state of human ESCs, whereas FGF4si functions as an antagonist of FGF4 and inhibits ESC self-renewal by blocking FGF4-induced Erk1/2 phosphorylation. Serca2 encodes a calcium pump, which is crucial for cardiac development and function (Greene et al., 2000; Misquitta et al., 2002). Two isoforms, Serca2a and b, with different 3′UTRs are generated by AS. While Serca2b with a long 3′UTR is expressed at similar levels in ESCs and differentiated EBs, Serca2a with a short 3′UTR is predominantly expressed in EBs (Salomonis et al., 2010). Importantly, the 3′UTR of Serca2b, but not Serca2a, is targeted by multiple microRNAs required for cardiac differentiation, such as miR-200b, miR-214, and miR-30 (Ivey et al., 2008; Salomonis et al., 2010). These results suggest that AS can alter gene regulatory potential and that the Serca2a isoform is used in EBs to avoid repression by microRNAs.

Although striking examples has been shown, the functions of the vast majority of AS events linked to ESCs, differentiation, and reprogramming have not been well characterized. An important goal in the future is to combine high-throughput assays/screens with focused studies in order to fully explore the biological functions of AS events in controlling cell fate decisions. Moreover, we still know very little about the repertoires of cis- and trans-acting factors that regulate these AS programs.
Figure 1-5 Examples of functionally important AS events in ESCs.

Schematic representation of the AS events discussed in section 1.7.2. For each case, different isoforms and their functional importance in ESC biology are shown. White and colored boxes represent constitutively and alternatively spliced exons/regions, respectively. Stop codons and untranslated regions are also indicated.
1.7.3 Regulation of AS networks in ESCs, differentiation, and reprogramming

Microarray profiling of AS events different between human ESCs and neural progenitors (NPs), together with comparative sequence analysis, has revealed that RBFOX binding motif is enriched in ESC-associated alternative exons, suggesting a potential role of RBFOX splicing factors in human ESC biology (Yeo et al., 2007b). Subsequently, using CLIP-Seq and RNA-Seq, the same group generated an RNA splicing map of the RBFOX2 protein in human ESCs (Yeo et al., 2009). RBFOX2 has been shown to regulate itself and many other splicing factors, and therefore may function as an upstream regulator of the AS network in human ESCs. Notably, RBFOX2 is critical for the survival of human ESCs, but not of neural stem cells or other cell lines. Ptbp1 is another splicing factor that has been implicated in ESC function (Shibayama et al., 2009). Knockout of Ptbp1 in mouse ESCs leads to a prolonged G2/M phase and a proliferation defect. As Ptbp1 has multiple roles in RNA metabolism, it is unclear about the contribution of AS regulation to the phenotype. Nevertheless, Ptbp1 has been demonstrated to regulate AS during neuronal differentiation (Boutz et al., 2007b; Li et al., 2014; Linares et al., 2015; Makeyev et al., 2007; Zheng et al., 2012). Ptbp1 globally represses neural-specific AS events in non-neuronal cells. During neuronal differentiation, Ptbp1 is targeted by the neural-specific miR-124. Down-regulation of Ptbp1 then promotes proper splicing and stable expression of its neural-enriched homolog, Ptbp2, and activates a neural-specific AS network. More recently, it has been shown that the function of Ptbp1 in neuronal differentiation is also influenced by an AS event in its own transcript and by the presence of additional neural splicing factors, such as nSR100 (Calarco et al., 2009; Gueroussov et al., 2015; Raj et al., 2014).

Collectively, AS networks that control cell fate are subject to complex regulation by combinations of trans-acting splicing factors and their cognate cis-elements as well as diverse other direct and indirect factors, majority of which remains to be identified and characterized. Integrated experimental and computational approaches need to be developed in order to systematically uncover landscapes of AS regulation in ESC pluripotency, differentiation, and somatic cell reprogramming. The massive AS profiling and protein-RNA interaction datasets generated by microarrays, RNA-Seq, and CLIP-Seq can be used to train sophisticated machine learning algorithms to decipher cis-regulatory codes that govern AS programs linked to cell fate decisions. The transcriptomic and proteomic expression profiles obtained using RNA-Seq,
ribosome profiling, and mass spectrometry, provide good opportunities to classify potential cell fate-associated splicing factors that are differentially expressed between pluripotent and differentiated cells and/or across different stage of lineage differentiation or somatic cell reprogramming. Interestingly, RNA processing factors including spliceosomal components display dynamic changes at both the mRNA and protein levels during reprogramming, especially at the early and late stages (Hansson et al., 2012; Polo et al., 2012). It suggests that regulated AS might play an important part in cell-state transitions during the reprogramming process. Moreover, high-throughput methods, such as mRNA interactome capture, have been applied to identify comprehensive lists of proteins that directly bind to mRNA in ESCs and differentiated cells (Baltz et al., 2012; Castello et al., 2012; Kwon et al., 2013). The increasing amount of information about RBPs is valuable for further selection of putative splicing factors. As mentioned in the previous sections, the expression and function of splicing factors are also influenced by post-translation and signaling pathways, and extensive interconnections are observed between the splicing, chromatin, transcription, as well as other co-/post-transcriptional and translational regulation. Therefore, it is important to integrate large-scale datasets from different levels to fully understand the coordinated gene regulation in ESC biology. Lastly, functional genomic approaches, such as RNAi, cDNA overexpression, and CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) screens, can greatly facilitate the systematic discovery of novel direct and indirect AS regulators in the context of cell fate control.

1.8 Rationale and overview of thesis research

The studies summarized above emphasize the prevalence and functional significance of AS. It is often tightly controlled in a spatio-temporal manner and disruption of splicing regulation can lead to various diseases. I am particularly interested in AS regulatory networks that determine cell fate decisions. In recent years, methodological advances have paved the way for large-scale identification of biologically relevant AS events in ESC pluripotency, differentiation, and during reprogramming of somatic cells to iPSCs. Nevertheless, relatively little is known about the factors and mechanisms dictating these AS outcomes, and many important questions remain to be addressed in the field. For example, which trans-acting regulators are responsible for controlling, directly or indirectly, the diverse patterns of cell fate-associated AS? How do these trans-acting factors interplay with combinations of cis-regulatory elements to establish complex AS regulatory networks? To what extent are these AS programs integrated and cross-regulated
with other layers of gene regulation to contribute to cell fate-specific functions. Accordingly, the main objective of my thesis research was to develop and apply high-throughput strategies so as to systematically uncover repertoires of factors that regulate AS events with important functions in cell fate control, including pluripotency, differentiation, and reprogramming, and thereby facilitate focused investigations of the underlying regulatory mechanisms.

In Chapter 2, I first describe the utilization of a large panel of RNA-Seq data to comprehensively define AS events that are differentially regulated in ESCs/iPSCs compared to diverse differentiated cells/tissues in both human and mouse. By integrating information from the splicing code and expression profiling data, we discovered that MBNL1 and MBNL2 are differentially expressed in pluripotent cells and function as regulators of ESC-differential AS events. Subsequently, using RNA-Seq profiling, RT-PCR assays, and CLIP-Seq, I demonstrated that MBNL proteins directly and negatively regulate approximately half of these ESC-differential AS events, including an ESC-specific AS switch in the forkhead family transcription factor FOXP1 that controls pluripotency. In line with their central and negative regulatory role in the establishment and maintenance of the pluripotent state, I found that overexpression of MBNL proteins in ESCs leads to increased kinetics of lineage differentiation. Notably, I also showed that MBNL knockdown significantly enhances the expression of key pluripotency genes and the formation of iPSCs during somatic cell reprogramming. These findings thus revealed for the first time that splicing regulators play a prominent role in the control of pluripotency, differentiation, and reprogramming (Han et al., 2013).

To develop a more complete picture of the regulatory networks controlling AS events in cell fate determination, I present in Chapter 3 the development of a novel high-throughput screen, which is capable of comprehensively linking trans-acting factors to endogenous AS and gene expression events of interest. In brief, I employed an advanced multiplex RT-PCR and barcode sequencing strategy to assay the effect of knockdown of genes with diverse functions on the regulation of a set of 50 or more prioritized AS events that have been previously implicated in ESC pluripotency, neural differentiation, and somatic cell reprogramming. I performed the screen in both mouse ESCs and neuroblastoma cell lines. The results revealed hundreds of factors from different gene regulatory layers that positively or negatively impact distinct subsets of the AS events monitored. In addition to splicing and RNA processing factors, many transcription factors, particularly a high fraction of C2H2 zinc finger proteins, were identified as novel direct
and indirect regulators of cell fate-associated AS events. In the last part of this chapter, I describe the regulatory roles of two striking examples, Zfp871 and Naac1, in controlling neural- and ESC-differential AS networks, respectively, and illustrate complex, multi-layered mechanisms underlying AS regulation.

Taken together, these studies have provided fundamental new insights into the global regulatory landscapes of AS events that are critical for cell fate determination. The high-throughput approaches developed here hold promise for unearthing the full repertoires of trans-acting factors that control AS and other RNA regulatory processes in different cell types under both physiological and pathological conditions.
This chapter is adapted from the following published article:

M. Irimia performed bioinformatic analyses with input from N.L. Barbosa-Morais and myself. P.J. Ross and T. Thompson performed human reprogramming experiments and iPSC characterization with assistance from myself and M. Gabut. H-K. Sung performed teratoma assays. B. Alipanahi generated splicing code data. I.P. Michael, H-K. Sung, and D. O’Hanlon assisted with ESC overexpression and differentiation experiments. L. David and A. Golipour assisted with secondary MEF reprogramming experiments and clone characterization, and D. Trcka generated secondary MEF lines and performed chimerism testing. E. Wang generated and analyzed CLIP-Seq data. E. Nachman and V. Slobodeniuc assisted with RT-PCR validation experiments. I performed all other experiments and data analyses presented in this chapter.
2 MBNL proteins repress ESC-specific alternative splicing and reprogramming

2.1 Abstract

Previous investigations of the core gene regulatory circuitry that controls embryonic stem cell (ESC) pluripotency have largely focused on the roles of transcription, chromatin, and non-coding RNA regulators. Alternative splicing (AS) represents a widely acting mode of gene regulation, yet its role in regulating ESC pluripotency and differentiation is poorly understood. Here, I identify the muscleblind-like RNA binding proteins, MBNL1 and MBNL2, as conserved and direct negative regulators of a large program of cassette exon AS events that are differentially regulated between ESCs and other cell types. Knockdown of MBNL proteins in differentiated cells causes switching to an ESC-like AS pattern for approximately half of these events, whereas over-expression of MBNL proteins in ESCs promotes differentiated cell-like AS patterns. Among the MBNL-regulated events is an ESC-specific AS switch in the forkhead family transcription factor FOXP1 that controls pluripotency. Consistent with a central and negative regulatory role for MBNL proteins in pluripotency, their knockdown significantly enhances the expression of key pluripotency genes and the formation of induced pluripotent stem cells (iPSCs) during somatic cell reprogramming.
2.2 Introduction

A core set of transcription factors that includes OCT4, NANOG, and SOX2, together with specific microRNAs and long non-coding RNAs, controls the expression of genes required for the establishment and maintenance of ESC pluripotency (Bao et al., 2013; Chen et al., 2008; Kim et al., 2008; Rinn and Chang, 2012; Silva et al., 2009; Young, 2011). Alternative splicing (AS), the process by which splice sites in primary transcripts are differentially selected to produce structurally and functionally distinct mRNA and protein isoforms, provides a powerful additional mechanism with which to control cell fate (Irimia and Blencowe, 2012; Kalsotra and Cooper, 2011; Nilsen and Graveley, 2010), yet its role in the regulation of pluripotency has only recently begun to emerge. In particular, the inclusion of a highly conserved ESC-specific “switch” exon in the FOXP1 transcription factor changes its DNA binding specificity such that it stimulates the expression of pluripotency transcription factors, including OCT4 and NANOG, while repressing genes required for differentiation (Gabut et al., 2011). However, the trans-acting regulators of this and other AS events (Mayshar et al., 2008; Rao et al., 2010; Salomonis et al., 2010) implicated in ESC biology are not known. These factors are important to identify, as they may control regulatory cascades that direct cell fate, and likewise they may also control the efficiency and kinetics of somatic cell reprogramming.

In this study, we utilized integrated splicing code and RNA-Seq profiling analyses to identify the muscleblind-like RNA binding proteins, MBNL1 and MBNL2, as potential regulators of a large network of alternative exons that are differentially spliced between ESCs/iPSCs and diverse non-ESC lines and tissues in both human and mouse, referred to below as “ESC-differential AS”. Subsequently, I performed siRNA knockdown, cDNA overexpression, RT-PCR, and RNA-Seq profiling experiments, and my results showed that MBNL proteins negatively regulate approximately 50% of the ESC-differential AS events, including the aforementioned ESC-specific exon in FOXP1. The direct binding of MBNL to their targets was further confirmed by CLIP-Seq. In addition, I found that overexpression of MBNL proteins in ESCs promotes lineage differentiation, whereas their knockdown during somatic cell reprogramming stimulates the expression of key endogenous pluripotency genes and results in a significant increase in iPSC formation. This study thus represents the first demonstration of a central role for trans-acting splicing regulators in the control of pluripotency, differentiation, and reprogramming.
2.3 Materials and Methods

2.3.1 Cell lines and cell culture

HeLa, 293T and C2C12 cell lines were maintained in Dulbecco’s Modified Eagle Medium (DMEM) supplemented with 10% Fetal Bovine Serum (FBS) and antibiotics (penicillin/streptomycin). Neuro2A (N2A) cells were grown in DMEM supplemented with 10% FBS, sodium pyruvate, MEM non-essential amino acids, and penicillin/streptomycin. H9 human ESCs, CGR8 and R1 mouse ESCs were cultured as described previously (Gabut et al., 2011). Secondary mouse embryonic fibroblasts (MEFs) were maintained in DMEM supplemented with 10% FBS, L-glutamine and penicillin/streptomycin on 0.1% gelatin-coated plates. During reprogramming, secondary MEFs were grown in mouse ES media and induced to express OKMS factors using 1.5 µg/mL of doxycycline (Dox) as described previously (Polo and Hochedlinger, 2010).

2.3.2 siRNA transfection

Cells were transfected with SMART-pool siRNAs (Dharmacon) using DharmaFECT1 reagent (Dharmacon), as recommended by the manufacturer. A non-targeting siRNA pool was used as a control. Cells were harvested 48 or 72 hours post transfection. In reprogramming experiments, secondary MEFs were transfected with siRNA pools using Lipofectamine™ RNAiMAX (Invivogen), as described previously (Samavarchi-Tehrani et al., 2010), and the OKMS transgenes were induced the day after with Dox.

List of siRNA target sequences (5’-3’)

siRNAs targeting human MBNL1 (M-014136-01-0010)

GAAAATGATTGTCCGTTTGG

ACAAATGTATGTCGCGTTG

TCACCTGTGTATCGGATTA

CCATAATATCTGCGGAACA

siRNAs targeting human MBNL2 (M-032198-00-0010)
GATGAAGAATGCAAATTTG
GCCACAAACTGCATGCTAA
GACAACCGTAAACCGTTT
GAGAGAGAAGCTGCAAGTAT

*siRNAs targeting human MBNL3 (M-015373-00-0010)*

CGTGATACCAAGTGGCTGA
GAGATTAATGGGCGGAACA
GGCTTTCAATCCTTACATA
GTAGAGAATTCAGAGAGG

*siRNAs targeting mouse Mbnl1 (M-065216-01-0010)*

TAAATGGGCAGAATAACTT
GAAAGGTCGTTGCTCCAGA
CCACCAGGCTCAATATTGT
GACCAGACACGGAATGTAA

*siRNAs targeting mouse Mbnl2 (M-065217-01-0010)*

GGAAACAATTTGATCCAGCA
CAGATGCAATTTATGTTTC
CTATGAGCTTGGCTCCTTA
ACAAACGACAACACCGTTA

*Non-targeting siRNA pool (D-001810-10-20)*
2.3.3 Protein extraction and western blotting

Cell pellets were lysed in radioimmunoprecipitation assay (RIPA) buffer by brief sonication. 30-150 µg of protein lysate was separated on a 10% SDS-polyacrylamide gel and transferred to a PVDF membrane. The membranes were blotted with the following antibodies: anti-Flag M2 (1:1500, Sigma), anti-MBNL1 (1:500, Abcam), anti-MBNL2 (1:200, Santa Cruz Biotechnology) and anti-α-tubulin (1:5000, Sigma). Secondary antibodies (GE Healthcare) and chemiluminescence reagents (Perkin Elmer) were used as per the manufacturer’s instructions.

2.3.4 RNA extraction and (q)RT-PCR assays

Total RNA was extracted using Tri Reagent (Sigma) or RNeasy columns (Qiagen). RT-PCR assays were performed using the OneStep RT-PCR kit (Qiagen), as per the manufacturer’s instructions. 20 ng total RNA or 1 ng of polyA+ RNA was used per 10 µL reaction. Radiolabeled reactions contained 0.3 µCi of α-32P-dCTP per 10 µL reaction. The number of amplification cycles was 22 for ACTIN and Gapdh, and 27-32 for all other transcripts analyzed. Reaction products were separated on 1-3% agarose gels. Quantification of isoform abundance was performed using either ImageQuant (GE Healthcare) or ImageJ software. To selectively amplify the FOXP1/Foxp1 isoforms (Figures 2-5C and 2-5D), primers specific for splice junctions were used. All primer sequences are available upon request.

For quantitative (q)RT-PCR, first-strand cDNAs were generated from 1-3 µg of total RNA or 100 ng of polyA+ RNA using SuperScript III Reverse Transcriptase (Invitrogen), as per the manufacturer’s recommendations, and diluted to 20 ng/µL and 1ng/µL, respectively. qPCR reactions were performed in a 384 well format using 1 µL of each diluted cDNA and FastStart Universal SYBR Green Master (Roche Applied Science). Primers used for qRT-PCR reactions are available upon request.

2.3.5 Mbnl1 and Mbnl2 overexpression in mouse ESCs

PiggyBac vectors “PB-TAG” and “PB-TAC” were generated to overexpress 3xFlag-Mbnl1 and 3xFlag-Mbnl2 proteins in mouse R1 ESCs. These vectors are a derivative of PB-TET (Woltjen et al., 2009) that co-express EGFP (PB-TAG) and mCherry (PB-TAC). Both vectors contain a tetracycline-regulated promoter controlling the expression of a bicistronic transgene comprising the gene of interest and EGFP or mCherry, respectively.
Overexpression R1 ESC lines were generated by co-transfection of: (1) PB-3xFlag-Mbnl1 and/or PB-3xFlag-Mbnl2, (2) PB-rtTA, a PB plasmid that includes the reverse tetracycline transactivator (rtTA) and the neomycin-resistance gene (neo) (Li et al., 2013), and (3) a plasmid coding for piggyBac transposase (PBase) (Wang et al., 2008b), followed by neomycin selection at 160 µg/mL. A constitutive PB-EGFP, PB-rtTA and PBase co-transfection was used as a control.

The overexpression of 3xFlag-Mbnl1 and/or 3xFlag-Mbnl2 was induced by 1.5ug/mL of Dox, as determined by western blot analysis (Figure 2-9A). To analyze the effect the 3xFlag-Mbnl1 and/or 3xFlag-Mbnl2 overexpression on ESC-differential AS, the cells were cultured with Dox for 48 hours. Cell populations with low and high EGFP and/or mCherry expression were collected on a MoFlo Astrios Flow Cytometer Cell sorter (Beckman Coulter) then lysed and analyzed by RT-PCR.

### 2.3.6 Mouse ESC differentiation

Multiple independent clonal mouse ESC (R1) line that co-express Mbnl1 and Mbnl2 at different levels under Dox inducible control were characterized. Following Dox induction, cells were cultured in mouse ESC media without LIF on 0.1% gelatin-coated plates. At day 6 following LIF withdrawal, expression of pluripotency factors in the clonal lines were analyzed by qRT-PCR and immunostaining. The parental R1 line, which was also treated with Dox, was used as a control.

Embryoid body (EB) differentiation was performed as described previously (Watanabe et al., 2005). ESCs were trypsinized and plated in 10 cm non-adherent plates at a density of 5 X 10⁵ cells per 10 mL in SFEB media. Cells were fed at days 3 and 5 by replacing 50% of the media. At day 7, the EB aggregates were seeded onto 0.1% gelatin-coated plates containing mouse ES media without LIF for another 4 days before harvesting for qRT-PCR analysis.

### 2.3.7 Immunofluorescence

For immunofluorescence experiments, cells were fixed in 4% PFA for 10 min at room temperature, washed with PBS, and permeabilized for 10 min at 4°C with 0.1% Triton X100. After one hour of blocking, cells were incubated with primary antibodies overnight at 4°C, and then with secondary antibodies for 1 hour at room temperature. Nuclei were stained with
Hoechst 33258 (1:5000, Sigma-Aldrich). Primary antibodies used in this study are: mouse IgM anti-SSEA1 (1:500, BD Biosciences), mouse anti-Oct4 (1:200, BD Biosciences), rabbit anti-Nanog (1:200, Cosmo Bio), goat anti-Dppa4 (1:250, R&D), mouse IgM anti-TRA1-60 (1:100, Invitrogen), rabbit anti-NANOG (1:400, Cell Signalling), mouse anti-SSEA4 (1:100, Invitrogen), rabbit anti-OCT4 (1:200, Abcam), mouse anti-alpha fetoprotein (1:200, R&D), mouse anti-smooth muscle actin (1:200, Invitrogen), and mouse anti-beta-III-tubulin (1:200, Millipore). Secondary antibodies used in this study are: anti-mouse IgM Alexa555 (1:1000, Molecular Probes), anti-mouse IgG Alexa555 (1:1000, Molecular Probes), anti-rabbit IgG Alexa594 (1:1000, Molecular Probes), anti-rabbit IgG Alexa488 (1:500, Molecular Probes), and anti-goat IgG Alexa546 (1:1000, Molecular Probes).

2.3.8 iPSC colony formation assays and imaging from secondary MEF reprogramming

Secondary MEFs were seeded in 12-well plates, transfected with siRNA pools, and treated with Dox for 5 days before fixing and staining. The plates were imaged (for both SSEA1-immunostained and DAPI channels) using an IN Cell Analyzer 2000 (GE Healthcare) with a 4X objective. For each well, 20 non-overlapping fields were captured and images were analyzed using the Columbus System (PerkinElmer). A custom script was generated to identify SSEA1-positive and DAPI-positive colonies. The overall signal in each well was determined using the sum of the overlap area for the 20 fields captured.

To assay the formation of Dox-independent colonies, secondary MEFs transfected with siRNA pools were treated with Dox for 8 days. Cell counting was performed before and at each passage after siRNA transfection and doubling rates were determined not to significantly change (data not shown). At day 8, the same number of cells were passaged into Dox-free mESC media on 12-well plates and cultured until day 13, when they were fixed and stained with alkaline phosphatase for colony counting.

2.3.9 Teratoma Analysis

Cells were suspended in PBS and Matrigel (BD Bioscience) mixed solution, and 1 x 10^6 cells in 100 µl were injected subcutaneously into both dorsal flanks of nude mice (CBByJ.Cg-Foxn1nu/J) anaesthetized with isoflurane. Four to 5 weeks after injection, mice were sacrificed and teratomas were dissected, fixed overnight in 10% buffered formalin phosphate, and embedded in
paraffin. Three to 4 µm thick sections were deparaffinized and hydrated in distilled water. Sections were stained either with haematoxylin and eosin for regular histological examination, or with the following dyes; 0.1 % Safranin O solution (cartilage, mesoderm-derived tissue) or 0.5 % Periodic Acid Schiff (PAS) solution (glycoprotein-producing intestinal cell, endoderm-derived tissue). For immunohistochemistry, sections were deparaffinized and hydrated, and antigen retrieval process was performed. After blocking, sections were incubated overnight at 4°C with primary monoclonal antibody (1:100, Millipore MAB377, clone A60) specific for neuronal nuclear antigen (NeuN, ectoderm-derived tissue), followed by washing in PBS. After 1 hour of incubation with secondary anti-mouse-HRP conjugated antibody (1:500, Jackson ImmunoResearch, 115-035-003), signal was visualized by DAB (3,3’-diadiminobenzidine; Vector Laboratories, SK-4100) substrate for 5-20 minutes. Sections were counter-stained with haematoxylin.

2.3.10 Chimerism analysis

Chimera aggregation and whole mount staining were performed as described previously (Golipour et al., 2012). Chimeras were obtained through aggregation of siMbnl iPS cell clumps with diploid Hsd:ICR(CD-1) embryos. E10.5 embryos were dissected after Dox treatment in utero via ingestion 24 hours prior to dissection. After dissection, embryos were fixed with 0.25% glutaraldehyde, rinsed in wash buffer [2mM MgCl2, 0.01% sodium deoxycholate, and 0.02% Nonidet-P40 in PBS] and then stained overnight in LacZ staining solution [20mM MgCl2, 5mM K3Fe(CN)6, 5mM K4Fe(CN)6 and 1mg/ml X-gal in PBS]. Embryos were embedded in paraffin, sectioned and counterstained with nuclear fast red.

2.3.11 Generation and characterization of human iPS cells

Human BJ foreskin fibroblasts (Stemgent) were reprogrammed using published protocols (Hotta et al., 2009a; Hotta et al., 2009b), with the following modifications. BJ fibroblasts were first infected with lentivirus vectors encoding both a puromycin resistance gene and doxycycline (dox)-inducible shRNA targeting either GFP (negative control, target sequence: 5’ GCAAGCTGACCCTGAAGTTCAT 3’) or MBNL1 shRNA (target sequence: 5’ GCCTGCTTTGATTCATTGAAA 3’). Lentiviral vector preparations and infections were performed as described (Hotta et al., 2009a). After selection with 1 µg/ml puromycin, shRNA-encoding BJ fibroblasts were infected with a second lentivirus vector (obtained from Addgene)
co-expressing both the mouse retrovirus receptor mSle7a1 and the blasticidin resistance gene (Takahashi et al., 2007). During transient selection with 5 µg/ml blasticidin, puromycin was reduced to 0.5 µg/ml and maintained at this concentration for 6 days after infection with retroviral reprogramming vectors.

pMXs-based retrovirus vectors encoding the four reprogramming factors hOCT4, hSOX2, hKLF4, hCMYC (OSKM) (Takahashi et al., 2007), were obtained from Adgene and packaged exactly as described (Hotta et al., 2009a). Puromycin/blasticidin-resistant BJ fibroblasts were infected in triplicate, using three separate preparations of retrovirus vectors. shRNA expression was induced by treatment with 2 µg/ml Dox, which was initiated contemporaneously with retrovirus vector infection; control cells were treated with vehicle only.

Six days after retrovirus infection, BJ fibroblasts were seeded on a monolayer of feeder cells. Embryonic day 12.5 fibroblasts from Tg(DR4)1Jae/J mice (Jackson Laboratory) were seeded on collagen-coated 6 well plates at a density of 3x10^5 cells per well as described (Hotta et al., 2009a); retrovirus-infected BJ fibroblasts were seeded at a density of 2x10^4 cells per well. At day 28 of reprogramming, quantification (by whole-well morphological examination and by TRA1-60 immunostaining) of human iPSC colonies was performed by investigators who were blinded as to the experimental conditions. To count TRA1-60 positive colonies, the plates were imaged (for both TRA1-60-immunostained and DAPI channels) using an IN Cell Analyzer 2000 (GE Healthcare) with a 4X objective. For each well, 64 non-overlapping fields were captured and images were analyzed using the Columbus System (PerkinElmer). Knockdown of MBNL1 resulted in an approximate two-fold increase in TRA1-60 immunostaining colonies over the control knockdown with GFP-targeting shRNA (Figure 2-15 and data not shown). Additional OSKM retrovirus-infected BJ fibroblasts were seeded in parallel; individual colonies from dox-treated plates were manually isolated four weeks post-infection, and seeded on feeders in collagen-coated 24 well plates (Hotta et al., 2009a). Cells from these colonies were expanded, and subsequently characterized as described (Cheung et al., 2011).

2.3.12 Clonal analysis by RNA-Seq during reprogramming

In a single-cell assay, secondary MEFs were plated in individual wells of a 96-well plate, OKSM factors were induced by Dox treatment and the clonal derivatives were cultured for 21 days. Removal of Dox at day 21 revealed that approximately 50% of clones produced abundant AP-
positive colonies (transgene-independent clones), while the rest yielded little or no colonies (transgene-dependent clones). RNA-Seq analysis was performed for 3 transgene-independent and 5 transgene-dependent clones at day 21 after Dox induction (SD 1-1).

Using RNA-Seq derived PSI values (see below), the inclusion levels of mouse ESC-differential cassette alternative exons were quantified for each of these clones. 51 ESC-differential AS events with sufficient read coverage in all samples and with a ≥ 25 PSI difference between iPSCs and MEFs were compared between the two types of clones (Figure A1-6; SD 1-5).

2.3.13 RNA-Seq data

We used RNA-Seq data from 36 and 32 different human and mouse samples, respectively. Details and sample sources are provided in SD 1-1. The samples comprise, for human: 5 ESC lines (3 different cell lines and 2 replicates), 2 iPSC lines, 7 non-ESC lines, and 22 adult tissues (18 different tissues and 4 replicates); for mouse: 6 ESCs, 2 iPSC colonies, 8 non-ESC lines (5 different cell lines and 3 replicates) and 16 adult tissues (10 different types and 6 replicates).

2.3.14 RNA-Seq data analysis and identification of ESC-differential AS events

Information on intron-exon structures was extracted from Ensembl annotations (release 65) for human (hg19) and mouse (mm9) genomes. In order to minimize the differences in annotation depth between mouse and human, mouse coordinates orthologous to human alternative exon-exon junctions (EEJs) were added to the annotations, using the lift-over tool from Galaxy (https://main.g2.bx.psu.edu/). From the resulting datasets, a Bowtie library of non-redundant EEJ sequences was generated for each species by combining every possible (forward combination) of splicing donor and acceptor within each gene. For each EEJ sequence, we determined the effective number of unique mappable positions. We extracted the L-k+1 (L being the EEJ length) k-mers from each EEJ sequence and then aligned the full set of k-mers against the EEJ library plus the respective genome using Bowtie (Langmead et al., 2009), allowing for a maximum of two mismatches along the entire length of the read. The number of k-mers with one unique alignment was counted for each EEJ; this corresponds to the junction’s effective number of unique mappable positions for a given set of RNA-Seq k-mers.
RNA-Seq reads from the different samples were then mapped to the EEJ libraries using Bowtie (Langmead et al., 2009) with \(-m 1 --v 2\) parameters. Reads were trimmed to 50 nucleotides, if longer, and reads mapping to the genome were previously discarded (since bona fide EEJs do not exist in the genome). A minimum of eight mapped nucleotides overlap was required across a given EEJ. Next, the outputs were parsed to identify cassette alternative exons (i.e. exons that are either included or fully excluded from the transcripts) by identifying exons that have associated reads mapping to (i) both EEJs formed by inclusion of the exon (i.e. a constitutive upstream exon \([C1]\) joined to an alternative exon \([A]\) and an \([A]\) exon joined to a constitutive downstream exon \([C2]\)), referred to as C1A and AC2 junctions below, and (ii) the EEJ formed by exclusion of the exon (C1C2).

The inclusion level of a cassette alternative exon was defined as the percentage of gene transcripts in which the exon is spliced in (PSI). To calculate this, we used a two-step approach. First, for each possible combination of C1 and C2 exons, PSI was estimated using read counts mapping to EEJs, corrected for the mappable positions, following the formula:

$$\text{PSI} = 100 \times \frac{C1A + AC2 + \sum (C_iA) + \sum (AC_j)}{C1A + AC2 + \sum (C_iA) + \sum (AC_j) + 2 \times (C1C2 + \sum (C_iC2) + \sum (C1C_j))}$$

where \(C_i\) is any possible splice donor upstream of the alternative exon (not including C1), and \(C_j\) any possible splice acceptor downstream of the alternative exon (not including C2). A reference pair of C1 and C2 exons was then selected (if more than one) as the combination yielding the maximum number of reads for the sum of the C1A, AC2 and C1C2 junctions. Alternative splicing events were only included when they met a minimal transcript coverage requirement of (i) \(\geq 15\) corrected reads mapping to the sum of exclusion EEJs, or (ii) \(\geq 15\) mappability-corrected reads mapping to one of the two inclusion EEJs (C1A+C_iA or AC_j+AC2), and \(\geq 10\) to the other inclusion EEJ. For alternative exons with multiple acceptor/donor splice sites, we used the splice site combination with the highest read support. A similar approach to calculate PSIs was recently used by Pervouchine et al. (Pervouchine et al., 2013).

To identify alternative exons differentially regulated in ESCs, we first calculated a single averaged PSI value for tissues of similar origin (see SD 1-1). Only events with enough coverage
in at least two ESC samples and three distinct tissue types were considered. ‘ESC-differential AS events’ were defined as those with a mean PSI difference of \( \geq 25 \) between ESCs and differentiated tissues. To account for AS events potentially related to cell proliferation, we also required a mean PSI difference of \( \geq 25 \) between ESC lines and non-ESC lines, when the event had sufficient coverage in at least one cell line sample. The set of background AS events used throughout the study are alternatively spliced exons (defined here as exons with PSI values of \(<95\%\) and \(>5\%\) in at least one sample) that meet the same expression requirement (i.e. in \( \geq 2 \) ESCs and \( \geq 3 \) differentiated tissue types) and that show an average difference in PSI level of \(<5\%\) between ESCs and differentiated tissue samples, and between the ESC and non-ES cell lines.

### 2.3.15 Analyses of splicing factor expression

A total of 221 human and 214 mouse genes were selected for analysis based on literature mining for previously described splicing functions, “splicing”- and/or “spliceosome”-associated Gene Ontology (GO) terms, and/or the presence of a PFAM-annotated RNA-binding domain (SD 1-4). To calculate the mRNA expression values for each sample, we used corrected (for mappability) Reads Per Kilobase pair and Million mapped reads values (cRPKMs) of the “stable” (as defined by BioMart) Ensembl transcript for each gene, as previously described (Labbe et al., 2012).

Splicing factor genes were ranked according to the relative extent of their differential expression (as determined by cRPKM values) in ESCs and iPSCs versus non-ESC lines and tissues by comparing summed ranks of each gene in all ESC/iPSCs across the full range of samples. Based on this approach, human \( MBNL1 \) and \( MBNL2 \) showed the 1\(^{st} \) and 2\(^{nd} \) lowest overall rank in ESCs/iPSCs, respectively, and mouse \( Mbnl1 \) and \( Mbnl2 \) showed the 1\(^{st} \) and 3\(^{rd} \) lowest overall rank in ESCs/iPSCs, respectively.

To assess the statistical significance of the differential expression of individual splicing factor genes, we compared their cRPKM values in ESCs/iPSCs to the cRPKMs in all other cell lines and differentiated tissues using a Wilcoxon rank-sum test after quantile normalization. Splicing factors with Bonferroni-corrected p-values < 0.05 were considered significantly differentially expressed (SD 1-4).
2.3.16 Splicing code analyses

The feature vectors for each species were produced by extracting sequence-based features from alternatively spliced exons, their adjacent constitutive exons, and 300 nucleotides of flanking intronic sequence. The features used were a subset of those defined in (Xiong et al., 2011), with the following differences: (a) all sequence length features are now in the log domain; (b) due to a lack of comprehensive transcript libraries and the corresponding uncertainty about downstream consequences of frame shifts, premature termination codon (PTC) features were excluded; and (c) conservation scores and conservation-weighted motifs were excluded from the feature set. In addition, related features (i.e. consensus recognition sequences for a given splicing factor inferred by different methods) were combined and included as independent features.

To identify features strongly associated with ESC-differential exon inclusion or exclusion, we compared 172 ESC-differential exons and 908 background (BG) exons for human, along with 102 ESC-differential exons and 811 BG exons for mouse. Associations between features and ESC-differential splicing were detected using Pearson correlation. For each feature, we computed the correlation between its value and the difference in average PSI values in ESCs and non-ESCs, across exons. To obtain more accurate correlation values, we considered two scenarios: (a) a positive scenario in which the differences in average PSI values in ESCs and non-ESCs are larger than 25%, and (b) negative scenario in which the differences in average PSI values in ESCs and non-ESCs are smaller than -25%.

2.3.17 CLIP-Seq analysis

We used recently described Mbnl1 CLIP-Seq data from C2C12 cells (Wang et al., 2012a). In order to estimate the fractions of ESC-differential and background AS events that are associated with Mbnl1 binding, we asked whether CLIP binding clusters overlap the alternative exon and/or flanking intron sequences of each event. CLIP binding clusters were defined as previously described (Wang et al., 2012a). In short, CLIP-Seq tags were trimmed of adapters and then collapsed to remove redundant sequences. These tags were mapped to genome and a database of splice junctions using Bowtie. To identify CLIP clusters lying within genic regions, gene boundaries were first defined using RefSeq, Ensembl, and UCSC tables. For each window of 30 nucleotides covered by at least one CLIP-Seq tag, a test was performed to assess whether the tag density in the window exceeded that which is predicted by a simple Poisson model which
accounts for gene expression and pre-mRNA length. An AS event was considered to have an overlapping Mbnl1 binding cluster if the mid-point of the cluster is located within the alternative exon, within 300 nt of the 5´ or 3´ ends of upstream or downstream flanking introns, and/or within 30 nt within the 3´ end of C1 exon or the 5´ end of C2 exon. Only AS events that had significant read coverage (see above) in at least one of two C2C12 samples used were analyzed. In total, 57 ESC-differential AS events and 601 background AS events were compared.

To generate an RNA regulatory map (Licatalosi et al., 2008) highlighting Mbnl1 binding sites in relation to ESC-differential AS events with either higher (ESC-included) or lower (ESC-excluded) exon inclusion levels in ESCs versus other cell lines and tissues, we applied the following procedure: for each nucleotide position from the regions described above and from three sets of AS events (ESC-included, ESC-excluded and background), we counted the average number of Mbnl1 CLIP-Seq tags. To minimize the impact of outliers with extreme read density, we limit the maximum count per event to an average of 10 reads per position within each region. In order to normalize the length of the AS exon, we divided each exon into 100 bins and uniquely assigned each position to the integer of 100*position/exon_length, with a relative weight inversely related to the length of the AS exon. To draw the map, we used sliding windows of 30nt for the intronic regions and 25nt for the length-corrected exons (total of four windows shown).

2.3.18 Evolutionary conservation of ESC-differential events

We analyzed three different aspects of conservation of the human and mouse ESC-differential alternative exons (Irimia et al., 2009). To determine whether the alternative exon is conserved at the genomic level, we performed a lift-over of the exon coordinates using Galaxy (https://main.g2.bx.psu.edu/). Exons with a unique lift-over hit in the other species, and with AG (splicing acceptor) and/or GT (splicing donor) sites were considered to be Genome-conserved in the other species. In addition, if the orthologous exon has a PSI of <95% and >5% in at least one sample from each species, AS of the exon was defined as conserved. Finally, to assess whether ESC-differential regulation is conserved, we applied two criteria: (i) the exons are independently detected as ESC-differential in human and mouse using the criteria as described above (total = 25 AS events), and (ii) the orthologous exons must meet minimal read coverage requirements (also as described above) to afford direct comparison.
2.3.19 Analyses of function and protein domain enrichment

To investigate whether ESC-differential events are significantly enriched in genes with specific functional associations and/or protein domains, we used the online tool DAVID (http://david.abcc.ncifcrf.gov/) (Huang da et al., 2009a, b) (with annotations from levels 3, 4 and 5 in the GO hierarchy), KEGG pathways and InterPro domains. As background, we used the genes with at least one AS event that met the minimal expression criteria described above (i.e. detection in ≥ 2 ESCs and ≥ 3 differentiated cell/tissue types). The main clusters of functionally related genes enriched in both human and mouse (as well as among the conserved) ESC-differential events (SD 1-3) are associated with: actin cytoskeleton, plasma membrane (including cell junctions), and protein kinase-associated terms.
2.4 Results

2.4.1 Identification of novel regulators of ESC-differential AS in human and mouse

To identify regulators of AS networks required for pluripotent stem cell, we first used high-throughput RNA sequencing (RNA-Seq) data to define human and mouse cassette alternative exons that are differentially spliced between ESCs/iPSCs and diverse differentiated cells and tissues, ESC-differential AS. A splicing code analysis (Barash et al., 2010) was then performed to identify cis-elements that may promote or repress these exons. The RNA-Seq data used to profile AS were also used to detect human and mouse splicing factor genes that are differentially expressed between ESCs/iPSCs and non-ESCs/tissues. By integrating these data sources, we sought to identify differentially expressed splicing regulators with defined binding sites that match cis-elements predicted by the code analysis to function in ESC-differential AS.

We identified 181 human and 103 mouse ESC-differential AS events, with comparable proportions of exons that are ≥25% more included or more skipped in ESCs versus the other profiled cells and tissues (Figures 2-1A, 2-1B, and 2-2A; SD 1-1 and SD 1-2). When comparing orthologous exons in both species, 25 of the human and mouse ESC-differential AS events overlapped (p<2.2e-16; hypergeometric test). The human and mouse ESC-differential AS events are significantly enriched in genes associated with the cytoskeleton (e.g. DST, ADD3), plasma membrane (e.g. DNM2, ITGA6), and kinase activity (e.g. CASK, MARK2 and MAP2K7) (SD 1-3). They also include the aforementioned FOXP1 ESC-switch AS event, and previously unknown AS events in other transcription or chromatin regulatory factors (e.g. TEAD1 and MTA1) that have been implicated in controlling pluripotency (Lian et al., 2010; Liang et al., 2008). These results suggest a considerably more extensive role for regulated AS in ESC biology than previously appreciated.

The splicing code analysis revealed that motifs corresponding to consensus binding sites of the conserved Muscleblind-like (MBNL) proteins are the most strongly associated with ESC-differential AS in human and mouse. The presence of MBNL motifs in downstream flanking intronic sequences is associated with exon skipping in ESCs, whereas their presence in upstream flanking intronic sequences is associated with exon inclusion in ESCs (Figure 2-1C, human code; Figure 2-2B, mouse code). To a lesser extent, features resembling binding sites for other splicing
regulators, including Polypyrimidine tract binding protein (PTBP) and RNA-binding fox (RBFOX) proteins, may also be associated with ESC-differential AS.

Figure 2-1 Identification of ESC-differential AS events in human.
(A) Heatmap of percent spliced in (PSI) values for 95 representative ESC-differential AS events in transcripts that are widely expressed across human ESCs/iPSCs, non-ESC lines, and differentiated tissues. (B) Examples of human ESC-differential AS events. Representative RT-PCR assays monitoring AS patterns in human ESCs (H9 and hES2), fetal (heart, brain, and liver) and adult (heart, brain, liver, and stomach) tissues. Red and blue exons indicate those exons with higher and lower inclusion in ESCs compared to differentiated tissues/cell lines, respectively. Primers specific to the constitutive exons were used. Asterisks indicate brain-specific MTA1 isoform. (C) Splicing code features that are significantly associated with ESC-differential AS. Features are ranked according to Pearson correlation p-values (y-axis) for alternative exons with either lower (top) or higher (bottom) inclusion in ESCs/iPSCs. Dashed lines indicate 300 nucleotide intervals from splice sites.
Figure 2-2 Identification of ESC-differential AS events in mouse.

(A) Heatmap of percent spliced in (PSI) values for 85 representative mouse ESC-differential AS events in transcripts that are widely expressed across ESCs, iPSCs, non-ESC lines, and differentiated tissues. (B) Splicing code features that are significantly associated (p<0.001, Pearson correlation) with mouse ESC-differential AS. Features are ranked according to Pearson correlation p-values (y-axis) for alternative exons with either lower (top) or higher (bottom) inclusion in ESCs/iPSCs.
From RNA-Seq expression profiling 221 known or putative splicing factors, eleven genes showed significant differential expression between human ESCs/iPSCs and other cells and tissues (Bonferroni-corrected p < 0.05, Wilcoxon rank-sum test) (Figure 2-3A; SD 1-4). Remarkably, *MBNL1* and *MBNL2* had the lowest relative mRNA levels in ESCs/iPSCs compared to other cells and tissues (Figures 2-3A and 2-3B; Materials and Methods). Quantitative RT-PCR assays confirmed this observation (Figure 2-3C). Similar results were obtained when analyzing mouse expression data (Figure 2-4; SD 1-4). PTBP, RBFOX, and other splicing factors potentially associated with ESC-differential AS by the splicing code analysis did not exhibit significant differences in mRNA levels between ESCs/iPSCs and other cells or tissues. Collectively, these results suggest a conserved and prominent role for MBNL1 and MBNL2 in ESC-differential AS.

**Figure 2-3 Expression profiling of human splicing factor genes.**

(A) Heatmap of Z-scores of mRNA expression (cRPMK) levels for splicing factors. Twenty-five splicing factors with the lowest or highest relative mRNA expression levels in ESCs/iPSCs compared to non-ESCs/tissues are shown. cRPMK, corrected reads per kilobase exon model per million reads (Labbe et al., 2012). (B and C) RNA-Seq and qRT-PCR data showing expression levels of human *MBNL1* and *MBNL2* mRNAs.
Figure 2-4 Expression profiling of mouse splicing factor genes.

(A) Heatmap of Z-scores of mRNA expression (cRPKM) levels for the 25 mouse splicing factors with the lowest or highest overall rank of ESC/iPSC mRNA expression compared to differentiated tissues and cell lines. Additional information in SD 1-4. (B and C) RNA-Seq and qRT-PCR data showing expression levels of mouse Mbnl1 and Mbnl2 mRNAs.

Because MBNL proteins are expressed at minimal levels in ESCs compared to other cell types, we proposed that they may repress ESC-differential exons in non-ESCs, and/or activate the inclusion of exons in non-ESCs that are skipped in ESCs. Indeed, previous studies have shown that in differentiated cells, MBNL proteins suppress exon inclusion when they bind upstream flanking intronic sequences, and they promote inclusion when binding to downstream flanking intronic sequences (Charizanis et al., 2012; Wang et al., 2012a). The results of the splicing code analysis are consistent with this mode of regulation, when taking into account that MBNL proteins are depleted in ESCs relative to differentiated cells and tissues (Figures 2-1C and 2-2B).

2.4.2 MBNL proteins regulate ESC-specific AS of FOXP1

To test the above hypothesis, I used siRNAs to knockdown MBNL1 and MBNL2 (to ~10% of their endogenous levels), individually or together, in human (293T and HeLa) and mouse (neuro2A [N2A]) cells (Figures 2-5A and A1-1A; see below). For comparison, knockdowns were performed in human (H9) and mouse (CGR8) ESCs. RT-PCR assays were used to monitor the ESC-switch exon of FOXP1/Foxp1 (human exon 18b/mouse exon 16b), which is partially
included in ESCs and fully skipped in differentiated cell types (Gabut et al., 2011). The splicing code analysis suggested that this exon is associated with conserved regulation by MBNL proteins, through possible direct disruption of splice site recognition (Figure 2-5B; see legend and below). Knockdown of MBNL2 in 293T or HeLa cells resulted in a <1% increase in FOXP1 exon 18b inclusion, whereas knockdown of MBNL1 alone, or together with MBNL2, resulted in increases in PSI, from zero to 2-2.4% and 6.5-8.5%, respectively (Figures 2-5C, 2-6, 2-7C, and A1-1B). More pronounced effects were observed for Foxp1 exon 16b in N2A cells (PSI shift from 0 to 15.1 for the double knockdown; Figures 2-5D and A1-1C). Knockdowns in ESCs had modest effects on exon 18b/16b splicing, consistent with the low levels of MBNL/Mbnl expression in these cells (Figures A1-1D and A1-1E). Knockdown of a third MBNL family member, MBNL3, which has a more restricted cell-type distribution compared to MBNL1 and MBNL2 (Pascual et al., 2006), had no detectable effect on exon 18b splicing (Figure 2-6). These results suggest that MBNL1 and MBNL2 proteins have conserved and partially redundant roles in the negative regulation of FOXP1/Foxp1 exon 18b/16b inclusion.

![Figure 2-5 MBNL proteins regulate the ESC-specific switch exon in FOXP1.](image)

(A) Western blots confirming efficient knockdown of MBNL1 and MBNL2 proteins in human 293T cells transfected with siRNA pools targeting these factors (siMBNL1+2, lane 6). Lane 5, lysate from cells transfected with a non-targeting siRNA pool (siControl). Lanes 1-4, serial
dilutions (1:1, 1:2, 1:4, and 1:8) of lysate from cells transfected with siControl. (B) Splicing code map highlighting genomic locations of MBNL, RBFOX, and PTBP motifs associated with ESC-specific AS of FOXP1/Foxp1 exon 18b/16b, the inclusion of which forms the FOXP1-ES/Foxp1-ES isoform. Human (black), mouse (grey), or conserved features (red) are indicated. Note that conserved MBNL motifs are associated with possible direct interference of exon 18b/16b splice site regulation. (C) RT-PCR assays monitoring mRNA levels of FOXP1 canonical (blue exon) and FOXP1-ES (red exon) isoforms in 293T cells transfected with siControl, siMBNL1, siMBNL2, or siMBNL1+2. RT-PCR employed splice junction-specific primers, as indicated. Expression levels of ACTIN are shown as loading controls. (D) mRNA levels of murine Foxp1 canonical and Foxp1-ES isoforms were assayed as in (C) in N2A cells. Expression levels of Gapdh are shown as loading controls.

![RT-PCR Assays](image)

**Figure 2-6** Comparison of effects of siRNA knockdown of MBNL1, MBNL2, and MBNL3 on ESC-differential AS.

Representative ESC-differential AS events were analyzed by RT-PCR assays following transfection of HeLa cells with a control siRNA, or with single or combinations of siRNA pools that target MBNL1, MBNL2, and MBNL3 (refer also to Figure 2-1 legend).
2.4.3 MBNL proteins directly and negatively regulate approximately half of the ESC-differential AS events

To assess the extent to which ESC-differential AS events are controlled by MBNL proteins, MBNL1+2 were knocked down in HeLa cells, and RNA-Seq profiling was used to detect AS changes (Figure 2-7). Of 119 profiled ESC-differentially spliced exons, nearly half are affected by knockdown of MBNL proteins, with a $\geq 15$ PSI change towards an ESC-like AS pattern (Figure 2-7A). A strong overall association ($p<2.2e^{-16}$, one-sided binomial test) was observed when comparing PSI changes for exons differentially spliced between ESCs and non-ESCs/tissues, and PSI changes for the same exons following knockdown of MBNL proteins (Figure 2-7B). RT-PCR experiments confirmed all analyzed MBNL knockdown-dependent and -independent PSI changes (Figures 2-7C and A1-2A). The specificity of the knockdown experiments was further demonstrated by comparing individual siRNAs that target different sequences within MBNL1 transcripts (Figure A1-2B). Comparable results were observed when MBNL1+2 proteins were simultaneously knocked down in 293T cells, and in undifferentiated C2C12 mouse myoblast cells (Figure 2-8). Conversely, overexpression of Mbnl1 and/or Mbnl2 proteins in mouse ESCs promoted differentiated cell-like patterns for all analyzed ESC-differential AS events, including a switch to the exclusive use of the canonical (i.e. non-ESC) exon 16 in Foxp1 transcripts, and maximal effects were generally seen when two proteins were co-expressed (Figure 2-9; see below).
Figure 2-7 MBNL proteins regulate approximately half of ESC-differential AS events.

(A) Venn diagram showing the proportion of ESC-differential AS events (green) that display $\geq 15$ PSI change between HeLa cells transfected with siRNA pools targeting MBNL1+2 (siMBNL1+2) versus siControl pool (orange). (B) High association ($p<2.2e^{-16}$, one-sided binomial test between quadrants) between differences in PSIs of ESCs versus differentiated cells/tissues, and differences in PSIs of siMBNL1+2 knockdown versus siControl treatments. (C) Representative RT-PCR validations for ESC-differential AS events that have PSI changes in HeLa cells following siMBNL1+2 transfection and for ESC-differential AS events that do not change upon siMBNL1+2 knockdown (c-); splicing patterns in human H9 ESCs are shown for comparison.
Figure 2-8 Analysis of ESC-differential AS events affected by knockdowns of MBNL1 and MBNL2 in human 293T and mouse C2C12 cells.

(A and B) Venn diagrams showing proportions of human and mouse ESC-differential AS events (green) that display ≥15 PSI change between 293T or C2C12 cells transfected with siMbnl1+2 versus a siControl control (orange). (C and D) High association (p < 2.4e-7 (C) and p < 7.6e-6
(D), one-sided binomial tests between quadrants) between differences in PSIs of ESCs versus differentiated cells/tissues, and between differences in PSIs of siMbnl1+2 knockdown versus siControl treatments of human 293T (C) or mouse C2C12 cells (D). Linear regression lines are shown. (E and F) RT-PCR validation of RNA-Seq-detected changes in PSI for ESC-differential AS events, following siMbnl1+2 knockdown in 293T (E) and in C2C12 cells (F). ESC-differential AS events detected by RNA-Seq not to change upon siMbnl1+2 knockdown are included as specificity controls (c-): CRTC2, SLK in (E) and Csnk1g3 in (F).
Figure 2-9 Overexpression of Mbnl1 and Mbnl2 in ESCs promotes differentiated cell-like AS patterns for ESC-differential AS events.

(A) Cell lysates prepared from mouse R1 ESCs over-expressing 3xFlag-Mbnl1 and/or 3xFlag-Mbnl2 under Dox inducible control, or EGFP (“GFP control”) were immunoblotted and probed with anti-Flag antibody (refer to Materials and Methods for details). (B) RNA isolated from cell populations with low and high EGFP and/or mCherry expression (collected by flow cytometer cell sorter) was analyzed by RT-PCR to assess effects of Mbnl proteins on ESC-differential AS events. Exogenous (ex) expression of Mbnl1 and Mbn2 transcripts was analyzed by RT-PCR in the same samples, as indicated.

Mapping of Mbnl protein binding to endogenous transcripts using UV-crosslinking coupled to immunoprecipitation and sequencing (CLIP-Seq or HITS-CLIP (Licatalosi et al., 2008)) in undifferentiated C2C12 myoblast cells (Wang et al., 2012a) confirmed that these proteins directly target ESC-differential AS events, including Foxp1 exon 16b (Figures 2-10A and 2-11A). Of 57 mouse ESC-differential exons expressed in C2C12 cells, ~34 (60%) are associated with overlapping or proximal clusters of Mbnl CLIP-Seq tags (“binding clusters”), whereas binding clusters are associated with 72/601 (12%) of exons that are not differentially regulated in ESCs (p < 2.2e-16, proportion test; Figure 2-10A). The binding clusters associated with ESC-differential AS are significantly enriched in consensus binding sites for MBNL proteins (Figure 2-11B) (Charizanis et al., 2012; Fernandez-Costa et al., 2011; Wang et al., 2012a). Moreover, consistent with the splicing code analysis (Figures 2-1C and 2-2B) and previous results (Charizanis et al., 2012; Wang et al., 2012a), the locations of Mbnl binding clusters correlate with whether the target exons are more or less included in ESCs compared to other cells and tissues (Figure 2-10B). Collectively, the results so far demonstrate that MBNL proteins act widely and directly to regulate ESC-differential AS.
Figure 2-10 MBNL protein acts widely and directly to regulate ESC-differential AS.

(A) Percentage of AS events with overlapping Mbnl1 CLIP-Seq binding clusters (Wang et al., 2012a) in C2C12 cells for ESC-differential or non-ESC-regulated alternative exons. (B) Merged map of Mbnl binding clusters in transcripts with ESC-differential AS events. Maps of Mbnl binding sites with respect to exons that have higher or lower inclusion in ESCs/iPSCs, relative to non-ESCs/tissues.
Figure 2-11 Mapping of Mbnl1 CLIP-Seq tags in the Foxp1 gene and sequence motifs enriched under Mbnl1 CLIP-Seq clusters.

(A) CLIP-Seq tags for Mbnl1 from murine myoblast C2C12 cells are shown mapped to the vicinity of Foxp1 exon 16b, the inclusion of which forms the Foxp1-ES isoform. The highest density of mapped tags is adjacent to the 5’ splice site of exon 16b, and binding of Mbnl1 in this region in differentiated cells is predicted to interfere with recognition by U1 snRNP. (B) Histogram and box-plot of the ratios of observed versus expected occurrences of 4-mers in Mbnl1 CLIP clusters in ESC-differential events in C2C12 cells, relative to randomized sequences from the same regions. An Mbnl RNA binding consensus (boxed) derived from the literature (refer to main text) is shown for comparison.

2.4.4 MBNL overexpression in ESCs suppresses pluripotency and promotes differentiation

To test the consequence of Mbnl protein overexpression in ESCs, I generated mouse R1 lines that express Mbnl1 and Mbnl2, singly or in combination, under Dox-inducible control (Figure 2-9A). In addition, I also characterized multiple clonal mouse R1 lines that co-express Mbnl1 and Mbnl2, and investigated the effects of Mbnl1+2 over-expression on pluripotency and differentiation. Intriguingly, overexpression of Mbnl1+2 in ESCs led to increased kinetics for silencing of expression of Oct4 and other analyzed key pluripotency factors following leukemia inhibitory factor (LIF) withdraw (Figures 2-12A and 2-12B). By comparing results from independent clonal lines that co-express Mbnl1 and Mbnl2 at different levels, I observed a strong inverse relationship between the levels of total Mbnl1+2 expression and the levels of pluripotency factors. This result is consistent with our observation that Mbnl proteins suppress pluripotency through repression of the ESC-specific splicing of FOXP1/Foxp1 exon 18b/16b, and possibly other ESC-specific AS events. Furthermore, following Dox induction and EB differentiation of independent clonal lines, I assayed expression of pluripotency factors (Oct4, Nanog, Rex1, and Nr5a2) as well as a panel of markers specific for endoderm (Gsc, Gata4, Sox17, Foxa2, and Afp), mesoderm (Brachyury, Flk1, and SMA), and ectoderm (Fgf5, Pax6, NeuroD1, and Tubb3) lineages (Figure 2-12C). The results showed that, while pluripotency marker expression were suppressed upon Mbnl1+2 overexpression, subsets of markers from each lineage display increased expression, with the largest effects seen for certain ectoderm markers.
2.4.5 Knockdown of MBNL proteins enhances somatic cell reprogramming

We next asked whether MBNL proteins impact somatic cell reprogramming (Figure 2-13A). Secondary mouse embryonic fibroblasts (MEFs) (Woltjen et al., 2009) expressing the “OKSM” factors (Oct4, Klf4, Sox2, c-Myc) (Takahashi et al., 2007) from transgenes under doxycycline (Dox)-inducible control were transfected with siRNA pools to knockdown Mbnl1 and Mbnl2 (siMbnl1+2), or with a control, non-targeting siRNA pool (siControl). At days 3 and 5 post Dox-induction, mRNA expression of endogenous pluripotency genes, including Oct4, Nanog, Sall4, and Alpl, were assayed by qRT-PCR (Figures 2-13B and A1-3A). None of these genes displayed significant changes in expression at day 3; however, at day 5, Mbnl knockdown stimulated their expression by approximately 2-fold over the siControl treatment (Figures 2-13A, 2-13B, and A1-3A). Mbnl knockdown also resulted in a ~30% increase in the colony area immunostained for SSEA1, a pluripotency-associated marker expressed early during reprogramming (Figure 2-13C). In contrast, knockdown of Oct4 (siOct4) resulted in significant reductions in endogenous pluripotency gene expression and in SSEA1-positive colonies (Figures 2-13B and 2-13C).

Successful reprogramming requires that cells transition to an OKSM transgene-independent state (Golipour et al., 2012). We therefore asked whether suppression of Mbnl proteins promotes transgene-independence. OKSM transgenes were induced for eight days, then cells were cultured for five days without Dox (Figure 2-13A). Whereas knockdown of Oct4 reduced colony
formation, knockdown of Mbnl proteins resulted in an approximate 2-fold increase in transgene-independent colonies, as detected by alkaline phosphatase (AP) staining (p = 0.0004; one-sided t-test) (Figures 2-13D and A1-3B). iPSC lines derived from transgene-independent colonies following Mbnl knockdown were pluripotent and contributed to all three germ layers in both teratoma and chimera assays (Figures 2-13E and A1-3-5). Consistent with these results, Mbnl expression is significantly reduced in secondary MEF clones (Golipour et al., 2012) cultured in the presence of Dox that are competent to achieve transgene-independence (when Dox is removed) versus those that are not (p = 0.006; one-sided t-test) (Figure 2-14, left). Moreover, the PSI levels of ESC-differential AS events, including Foxp1 exon 16b, significantly correlate with ESC/iPSC AS patterns only in those clones that transition to transgene-independence (Figure 2-14, right; Figure A1-6; SD 1-5; r= 0.80, p= 3.2e^{-12}). Strikingly, knockdown of MBNL1 in human fibroblasts expressing OKSM also resulted in an approximate 2-fold increase in the appearance of iPSC colonies (Figures 2-15 and 2-16). MBNL proteins thus have a conserved, negative regulatory role in somatic cell reprogramming.
A

2^c MEFs

Dox induction of "OKSM"  
Mbn1+2  
siRNA KD

Day 0  
Day 3  
Day 5  
Day 8  
Day 13

qRT-PCR assays  
qRT-PCR assays & SSEA1 staining  
AP-positive colony count  
Pluripotency assays

iPSCs

B

B

Oct4  
Nanog

Relative qRT-PCR units

p ≤ 0.05

siControl  
siMbn1+2  
siOct4  
MEF

siControl  
siMbn1+2  
siOct4  
MEF

C

Change in SSEA1 stained area (relative to siControl)

siMbn1+2  
siOct4

D

Number of AP-positive colonies

siControl  
siMbn1+2  
siOct4

p = 0.0034

E

Endoderm  
Mesoderm  
Ectoderm

PAS  
SafQ  
NeuN
Figure 2-13 Knockdown of Mbnl proteins enhances reprogramming efficiency and kinetics.

(A) Experimental scheme. (B) qRT-PCR quantification of mRNA expression levels of endogenous Oct4 and Nanog (data for additional genes in Figure A1-3A). Secondary MEFs were transfected with control siRNAs (siControl), siRNAs targeting Mbnl1 and Mbnl2 (siMbnl1+2) or Oct4 (siOct4) and treated with doxycycline (Dox) for 3 days (blue bars) or 5 days (red bars) before analysis. Empty bars, secondary MEFs without Dox induction. Values represent Means ± Range (n=3). (C) Top, quantification of SSEA1-stained area change relative to siControl at day 5 post Dox-induction; values represent Means ± Range (n=3); bottom, representative images of SSEA1 staining. Scale bar = 100 µm. (D) Top, quantification of Dox-independent iPSC colony formation. Secondary MEFs were treated with Dox for 8 days followed by 5 days of Dox withdrawal and counting of alkaline phosphatase (AP)-positive colonies; bottom, representative images of AP staining. (E) Teratoma assay assessing the pluripotency potential of iPSCs derived from secondary MEFs following knockdown of Mbnl proteins. Hematoxylin and eosin staining, with additional staining/immunolabeling using periodic acid-Schiff (PAS; for detection of glycogen or glycoprotein producing cells), Safranin O (Safo; for detection of cartilage), or antibody to neuronal nuclear antigen (NeuN); additional teratoma analysis and chimera testing of the pluripotency potential of siMbnl iPSCs in Figures A1-4 and A1-5. Scale bar = 100 µm. p-values of one-sided t-tests shown for all comparisons in this figure.

Figure 2-14 MBNL-regulated ESC-differential AS programs are important for transition to transgene independence.

Top, experimental scheme for clonal analysis. Upon Dox removal at day 21, clones derived from single cells either survive and form iPSCs (“transgene-independent”) or do not survive (“transgene-dependent”). Bottom, analysis of total Mbnl1/Mbnl2 mRNA expression (left) and percentage of total PSI change (right) for Foxp1 exon 16b in transgene-independent (red) and transgene-dependent (blue) clones, at day 21, where total PSI change is the PSI difference between MEFs and iPSCs during reprogramming. p-values of one-sided t-tests shown for all comparisons in this figure.
Figure 2-15 Knockdown of MBNL1 enhances human somatic cell reprogramming

(A) Quantification (by morphological examination) of human iPSC colonies formed by reprogramming BJ fibroblasts expressing shRNA targeting GFP (shGFP) or MBNL1 (shMBNL1). (B) Immunostaining of human iPSCs derived from shMBNL-expressing BJ fibroblasts for TRA1-60, NANOG, SSEA4, and OCT4 pluripotency markers. Scale bar = 50 µm. Additional characterization of human iPSCs in Figure 2-16. p-values of one-sided t-tests are shown.
Characterization of human iPSC lines generated from MBNL1-depleted fibroblasts.

(A and B) qRT-PCR analyses of iPSC lines that expressed shRNA targeting either GFP (shGFP1) or MBNL1 (shMBNL1) during reprogramming reveal gene expression patterns that are consistent with full reprogramming: compared to control fibroblasts harvested 7 days after retroviral infection (IMR90 4YF), most lines have silenced expression of viral transgenes (A); also, endogenous genes encoded by the reprogramming-factors are expressed at levels similar to those observed in human embryonic stem cells (H9 hESCs) (B). (C) iPSCs generated from fibroblasts expressing shGFP1 or shMBNL1 have normal colony morphology and express the pluripotency-associated proteins NANOG, TRA1-60, SSEA4, and OCT4 (scale bar = 50 µm). (D) Embryoid body assays show that iPSC lines generated from fibroblasts depleted of MBNL1 differentiate into cells representing all three germ layers (scale bar = 50 µm).
2.5 Discussion

2.5.1 The central role of MBNL proteins in regulating ESC-differential AS, pluripotency, and reprogramming

The results of this study reveal that MBNL proteins directly and negatively regulate an ESC-differential AS network that controls pluripotency and reprogramming (Figure 2-17). These proteins likely act in part by directly repressing the ESC-specific splicing switch in FOXP1, which promotes the expression of core pluripotency genes. However, additional genes with MBNL-regulated AS events have been linked to the control of pluripotency, suggesting a more extensive role for the AS network in ESC biology (Figure 2-17). These observations represent the first evidence that trans-acting splicing regulators play a central role in the core circuitry required for ESC pluripotency and reprogramming. Importantly, my results further offer a potential new approach for enhancing the production of quality iPSCs for research and therapeutic applications.

![Model for the role of MBNL proteins in the regulation of ESC-differential AS, pluripotency, and iPSC reprogramming.](image)

Asterisks indicate significantly enriched gene-function categories.
2.5.2 Coordinated regulation of AS networks in stem cell biology

Consistent with our initial discovery, evidence from several subsequent studies has further supported the functional significance of AS regulatory networks in pluripotency, differentiation, and reprogramming. Using high-throughput RT-PCR assays, an overlapping program of AS changes has been defined during reprogramming of human fibroblasts into iPSCs and their subsequent redifferentiation (Venables et al., 2013). Moreover, this study confirmed the important role of MBNL proteins in controlling ESC-differential AS events, and demonstrated that they function in a cooperative manner with another splicing factor, RBFOX2. In parallel, a combination of RNA-Seq and high-throughput qRT-PCR methods has been employed to survey AS changes during somatic cell reprogramming, further establishing that AS patterns progressively revert to an ESC-like state in the process of generating iPSCs (Ohta et al., 2013). Subsequently, this study also identified two potential splicing regulators, U2af1 and Srsf3, whose downregulation in MEFs inhibits efficient reprogramming. More recently, Myc and Gcn5, a subunit of the SAGA histone acetyltransferase complex, have been shown to promote the expression of splicing factor genes and thus to activate AS programs that are required for the initiation of reprogramming, highlighting interconnections between transcription, chromatin, and splicing machineries (Hirsch et al., 2015). In addition, another study has demonstrated that the spliceosome-associated factor SON is critical for human ESC survival and pluripotency by ensuring proper splicing of transcripts encoding cell cycle proteins and key pluripotency factors (Lu et al., 2013). Lastly, a positive feedback loop comprising the pluripotency transcription factor OCT4, the splicing factor SRSF2, and AS of the methyl-CpG binding domain protein MBD2, has recently been reported to play a critical part in supporting ESC self-renewal and somatic cell reprogramming (Lu et al., 2014). Interestingly, additional regulatory complexity in this subnetwork is contributed by the ESC-enriched miR-302, which specifically targets the MBD2a isoform.

Taken together, these findings underscore the functional coordination between splicing regulatory networks and other layers of gene regulation in stem cell biology. Remarkably, AS is subject to complex regulation by factors associated with distinct gene regulatory pathways. In future work, it will be important to further elucidate the regulatory networks and the underlying control mechanisms. Next chapter, I will describe a new high-throughput screen for the systematic identification of regulators for AS networks that control cell fate.
Chapter 3

This chapter is adapted from the following manuscript:

U. Braunschweig performed bioinformatic analyses with assistance from myself, R.J. Weatheritt, K.C.H, Ha, and T. Sterne-Weiler. T. Gonatopoulos-Pournatzis conducted iCLIP and in vitro splicing experiments. R.J. Weatheritt performed correlation network analyses and functional cluster identification. C.L. Hirsch and D. O’Hanlon assisted with screen hit characterization. E. Radovani performed ChIP experiments. J. Wang assisted with RT-PCR validation experiments. Q. Pan wrote the primer design script along with my input. D. Ray generated recombinant proteins. U. Braunschweig, F. Vizeacoumar, and A. Datti assisted with the siRNA screen. L. Magomedova assisted with experiments characterizing Arglu1. I developed and conducted the siRNA screen and Systematic Parallel Analysis of endogenous RNA regulation coupled to barcode Sequencing (SPAR-Seq). In addition, I performed all other experiments and data analyses presented in this chapter.
3 Systematic discovery of alternative splicing regulatory networks controlling cell fate

3.1 Abstract

Networks of coordinated alternative splicing (AS) events play critical roles in development and disease. However, a comprehensive knowledge of the factors that regulate these networks is lacking. Here, I describe a high-throughput system for systematically linking trans-acting factors to endogenous RNA regulation events. Using this system, I identify hundreds of factors associated with diverse regulatory layers that positively or negatively control AS events linked to cell fate. Remarkably, more than one third of the new regulators are transcription factors. Further analyses of the zinc finger protein Zfp871 and BTB/POZ domain transcription factor Nacc1, which regulate neural and stem cell AS programs, respectively, reveal roles in controlling the expression of specific splicing regulators. Surprisingly, these proteins also appear to regulate target AS programs via binding RNA. Our results thus uncover a large ‘missing cache’ of splicing regulators among annotated transcription factors, some of which dually regulate AS through direct and indirect mechanisms.
3.2 Introduction

Alternative splicing (AS) is the process by which different combinations of splice sites in precursor mRNA (pre-mRNA) are selected to generate structurally and functionally distinct mRNA and protein variants. It acts widely to expand the functional and regulatory capacity of metazoan genomes (Irimia and Blencowe, 2012; Kelemen et al., 2013; Licatalosi and Darnell, 2010; Nilsen and Graveley, 2010). High-throughput profiling studies have indicated that nearly all transcripts from human multi-exon genes are alternatively spliced, and a substantial fraction of these splice variants are differentially expressed in a cell- and tissue-specific manner (Pan et al., 2008; Wang et al., 2008a). AS plays critical roles in a wide range of biological processes, including cell fate determination, and misregulation of AS is associated with numerous human diseases (Daguenet et al., 2015; Jangi and Sharp, 2014; Kalsotra and Cooper, 2011). Given the prevalence and importance of AS, a major challenge that lies ahead is to understand how networks of AS events are coordinately regulated to impart their biological roles in diverse cellular contexts.

The spatiotemporal specificity of AS events is governed by the intricate interplay of combinations of cis-regulatory elements and cognate trans-acting splicing factors, which promote or inhibit spliceosome assembly at particular splice sites (Chen and Manley, 2009; Fu and Ares, 2014; Wahl et al., 2009). In addition, it is increasingly recognized that AS is controlled by complex mechanisms involving coordinated interactions with other gene regulatory layers, including transcription and chromatin (Braunschweig et al., 2013; Luco et al., 2011). Moreover, post-translational and signalling pathways have been shown to influence AS outcomes through different mechanisms, such as by altering the function and/or subcellular localization of key splicing regulators (Heyd and Lynch, 2011; Li et al., 2007). However, the full repertoires of splicing regulators and associated mechanisms that operate in different cell types are not known. This question is especially relevant to cell types with relatively complex AS patterns, such as embryonic stem cells (ESCs) and neural cells.

In recent years, considerable progress has been made in the development of screening strategies directed at the large-scale investigation of AS. Fluorescent- or luciferase-based splicing minigene reporters have been utilized in high-throughput RNA interference (RNAi), small molecule, and cDNA overexpression screens to uncover factors that control individual AS events (Moore et al.,
However, since splicing reporters often do not recapitulate important aspects of AS regulation, such as mechanisms involving cross-talk with chromatin and transcription components, high-throughput screening methodologies that monitor endogenous AS changes are invaluable for establishing full repertoires of regulators. Recent progress in this direction has included the employment of a high-throughput quantitative PCR (qPCR) assay to screen a mutant strain collection for new splicing regulators in budding yeast (Albulescu et al., 2012), automated RT-PCR platforms coupled to capillary gel electrophoresis to monitor the effects of knockdown of candidate regulators on apoptosis- and proliferation-related AS events (Papasaikas et al., 2015; Venables et al., 2008), and a related genome-wide siRNA screen for regulators of an apoptosis-associated AS event in the Fas/CD95 gene (Tejedor et al., 2015). These studies have begun to illuminate interesting and important functional inter-relationships within and between core and ancillary splicing regulators, as well as other factors that control specific AS events.

In this study, I describe "Systematic Parallel Analysis of endogenous RNA regulation coupled to barcode Sequencing" (SPAR-Seq), a highly multiplexed and quantitative functional genomics screening platform coupled to a sequencing output that is capable of comprehensively linking trans-acting factors to dozens of endogenous gene regulation events of interest. I apply SPAR-Seq to the elucidation of regulatory networks that control conserved, endogenous AS events linked to ESC pluripotency, neural differentiation, and somatic cell reprogramming (Gabut et al., 2011; Han et al., 2013; Kolle et al., 2011; Ohta et al., 2013; Rao et al., 2010; Salomonis et al., 2010; Venables et al., 2013). Our results reveal hundreds of new splicing regulators functionally associated with different regulatory layers that act in a positive or negative manner to impact distinct subsets of these AS events in ESCs and neural cells. Surprisingly, in neural cells, annotated transcription and DNA binding factors regulate AS events at a comparable frequency as defined splicing regulators. Further characterization of two examples, Zfp871 and Nacc1, demonstrates that they have dual direct and indirect regulatory roles in the control of neural and ESC-differential AS networks, respectively. Collectively, this study introduces a versatile new technology for elucidating endogenous RNA regulatory networks, and highlights its application in revealing new landscapes of trans-acting regulators and associated multilayered mechanisms that impact AS events with key roles in cell fate decisions.
3.3 Materials and Methods

3.3.1 Cell lines and cell culture

CGR8 mouse embryonic stem cells (ESCs) were cultured as described previously on gelatin-coated plates (Gabut et al., 2011; Han et al., 2013). Mouse neuroblastoma (N2A) cells were grown in DMEM supplemented with 10% FBS, sodium pyruvate, MEM non-essential amino acids, and penicillin/streptomycin. All cell lines were maintained at 37°C with 5% CO₂.

3.3.2 High-throughput siRNA knockdown and RNA purification

Knockdown and control treatments were performed in both mouse ESCs (CGR8) and neuroblastoma cells (N2A), in two biological replicates, at the Lunenfeld-Tanebaum Research Institute (LTRI) SMART robotics facility. Treatments comprised SMARTpool siRNAs (siGENOME, Dharmacon) targeting 1416 genes that function in diverse aspects of gene regulation, 32 positive controls (siRNA targeting Mbnl1 and Mbnl2, siMbnl), and 88 negative controls (non-targeting siRNA, mock transfection, and untreated cells).

An automated pipeline was developed for high-throughput cell plating, siRNA transfection, and RNA purification (Biomek FX Laboratory Automation Workstation, Beckman Coulter). Twenty-four hours prior to transfection, CGR8 and N2A cells were seeded in 96-well plates, using 3000 and 5000 cells per well, respectively. Cells were transfected with the SMARTpool siRNAs at 50 nM final concentration using DharmaFECT1 reagent (Dharmacon), as recommended by the manufacturer. Forty-eight hours post-transfection, total RNA was purified from cultured cells using the RNeasy Plus 96 Kit (Qiagen), as per the manufacturer’s instructions. In total, 6144 (sixty-four 96-well plates) RNA samples (~65 µL per well) were prepared from CGR8 and N2A cells from two replicates.

3.3.3 Systematic Parallel Analysis of endogenous RNA regulation coupled to barcode Sequencing (SPAR-Seq)

SPAR-Seq was developed for the parallel analysis of dozens of endogenous alternative splicing (AS) events in response to thousands of knockdown and control treatments. For each treatment, a multiplex RT-PCR assay was applied to simultaneously amplify 50 transcript regions that span multiple exons to assess AS and gene expression in a single reaction. In some cases, more than one AS event was monitored from the same region. Optimized event-specific primers with 5'
universal adaptor sequence were used (Figure 3-1A and 3-1B). Primers for AS events were designed to anneal near splice junctions in order to monitor different spliced variants (SD2-1 for primer sequences). The multiplex RT-PCR reaction was carried out in 96-well plates using the OneStep RT-PCR kit (Qiagen) as recommended by the manufacturer, with the following changes: reactions were performed in a volume of 20 µL with 2 µL of the purified total RNA as input, and a mixture of 50 pairs of primers was added to each reaction at a final concentration of 0.025 µM for each individual forward and reverse primer. Four identical Veriti 96-well Thermal Cyclers (Applied Biosystems) were used with the following program: 50°C for 30 minutes, 95°C for 15 minute, 30 cycles of 94°C for 40 seconds, 58°C for 1 minute (slow ramp rate), 72°C for 3 minutes, and a final extension step at 72°C for 10 minutes.

For multiplex barcode sequencing, unique, dual-index barcodes were designed, including 16 forward 8-base barcodes (minimum Hamming distance of 4) and 768 reverse 8-base barcodes, comprising a subset of previously reported barcode sequences (Hamady et al., 2008). Reverse barcodes were selected to further increase the distance between barcodes, and minimize hairpin structures and primer dimers. To multiplex 768 samples per sequencing lane in the current screen (see below), unique reverse barcodes were used for each sample, while forward barcodes were used to mark each half (48 samples) of a 96-well plate to provide additional redundancy. These two sets of barcodes were incorporated into forward and reverse primers, respectively, after the universal adaptor sequences and were added to the amplicons in the second PCR reaction, which was performed using Phusion High-Fidelity DNA Polymerase (Thermo Scientific), as per the manufacturer’s instructions. For each 20 µL of reaction, 1 µL of the multiplex RT-PCR reaction product was used as template. The thermal cycling conditions were as follows: 98°C for 30 seconds, 15 cycles of 98°C for 10 seconds, 65°C for 30 seconds, 72°C for 30 seconds, and a final extension step at 72°C for 5 minutes.

To achieve high-throughput preparation of barcode sequencing libraries for the screen, a Versette Automated Liquid Handler (Thermo Scientific) was programmed to set up both the multiplex RT-PCR reactions and the second Phusion PCR reactions in 96-well plates. The resulting libraries were pooled and sequenced at the Donnelly Sequencing Centre. The entire screen monitored 52 AS events from 6144 samples, as described above. An Illumina HiSeq 2500 flow cell with 8 lanes was used for the barcode sequencing, and 768 samples from eight 96-well plates
were multiplexed per lane. In total, approximately 2.1 billion 101-bp paired-end reads and two separate 8-bp index reads were generated.

3.3.4 Events monitored in the screen

The majority of the monitored AS events are ESC-differential AS events defined by large-scale RNA-Seq analyses (Han et al., 2013) and other studies (Gabut et al., 2011; Kolle et al., 2011; Ohta et al., 2013; Rao et al., 2010; Salomonis et al., 2010; Venables et al., 2013), some of which have been shown to be important for ESC pluripotency, differentiation, and somatic cell reprogramming. Multiple non-ESC-specific AS events (e.g. neural-regulated exons and microexons) (Irimia et al., 2014; Raj et al., 2014) and gene expression only events (e.g. splicing factor and internal control genes) were also monitored (see SD2-1 for further details).

3.3.5 Selection of candidate splicing-related genes

Candidate genes for knockdowns (SD2-2) were selected using multiple approaches. Genes that encode proteins with known or putative roles in splicing and RNA processing were curated manually from the literature and comprised all known splicing factors, including spliceosome-associated proteins as well as known and predicted RNA binding proteins (RBPs). In addition, genes encoding chromatin-related proteins and transcription factors were selected using four approaches: (1) Genes whose involvement in splicing regulation has been reported previously; (2) Genes with the GO Slim annotation 'chromosome', to capture chromatin-associated proteins; (3) Genes encoding proteins with domains that are involved in chromatin modification/binding/remodeling, based on literature and PFAM domain descriptions; (4) Genes encoding proteins with a domain composition similar to known chromatin proteins. The latter set was identified as follows: the presence of PFAM domains in the protein encoded by the transcript with most domains of each gene was tabulated. Genes were labeled as positive if associated with the GOslim term 'chromosome', and negative if not associated with it but associated with at least one term related to cytoplasm, extracellular region, plasma membrane, or translation. Two-thirds of the labeled set were used to train a support vector machine with ten-fold cross-validation using the svm() function in the R package e1071 (Meyer et al., 2015). After parameter optimization, a set of parameters that yielded a true positive rate of 0.479, false positive rate of 0.028, and accuracy of 0.945 on the training set was used to train a final model on the full labeled set of genes to predict association with 'chromosome' for unlabeled genes.
Genes positively labeled by this prediction were associated with the term 'nucleus' >3.5 times more often than negatively labeled genes. Of 1,925 pre-selected genes from all four streams, 836 were finally selected based on minimum expression cut-off, measured using RNA-Seq data) in ESCs or N2A cells and availability of SMARTpool siRNAs. Lastly, signaling components and post-translational factors that have been previously linked to splicing regulation were also included.

For all categories, a total of 1,416 genes were selected for knockdown. The four groups shown in Figure 3-1C and SD2-2 represent a post hoc categorization based on the approaches described above but prioritizing RNA binding domains, association with the spliceosome, experimental detection as part of the mRNA interactome (Baltz et al., 2012; Castello et al., 2012; Kwon et al., 2013), GO annotation “RNA binding” or “RNA processing” for the splicing factor/RBP group, and the occurrence of a C2H2 domain.

### 3.3.6 Pipeline for high-throughput screen data analyses

*De-multiplexing and mapping*

Sets of reads from each sequencing lane, consisting of forward/reverse event reads and forward/reverse barcode reads, were assigned to one of 768 samples by matching the forward and reverse barcodes to the expected combinations. Up to two mismatches were allowed if and only if these allow a match to a single barcode. ~79% of the total read sets from 8 lanes were successfully assigned to the 6,144 samples. Forward and reverse event reads were then mapped to custom junction libraries representing all expected splice variants (available on GEO: GSE80205), using bowtie with settings --best -v 3 -k 1 --trim3 26 --trim5 20 (Langmead et al., 2009). Trimming of the first 26 and last 20 bases was performed to remove lower-quality and uninformative ends and thereby increase the rate of mappable reads. Junction libraries were first constructed from NCBIm37/mm9 gene annotations, and then refined based on the results of de novo mapping of the reads from one full Illumina lane from each cell line using TopHat (default settings with -i 40 and providing Ensembl transcript annotations for NCBIm37) (Trapnell et al., 2009), after quality trimming (minimum MAPQ of 33) from the 3’-end. Initially, 64 events within 50 genes were considered. In total, ~20% of the forward and reverse reads were mapped uniquely to one splice variant. Across all 6,144 samples, the median number of reads per gene was ~300 in each direction, with ~91% reaching > 20 reads (Figure 3-3B).
**AS quantification and SSMD calculation**

Percent Spliced In (PSI) values were calculated for each alternative exon, or part thereof in the case of alternative 5' or 3' splice sites ('event'), independently from forward and reverse reads as the percentage of reads supporting inclusion divided by the total number of reads for the event. For alternative 5' or 3' splice site events that were part of an alternative exon [Foxm1 (A2), Mta1 (A1), Uspl1 (A3)], the PSI was instead calculated with reference to the total reads supporting inclusion of that exon in order to assess independent regulation of alternative splice site usage. The average of PSI values obtained from forward or reverse reads was calculated, except in cases where some forward or reverse reads were ambiguous with respect to either inclusion or exclusion due to insufficient read length, in which case only one read was used. PSI values showed low-level batch effects per 96-well plate, which were reduced by subtracting a weighted plate median in which negative controls (siNT and mock treatment) were given 20x more weight than other samples. To derive variances, and because PSI values are not nearly normally distributed but roughly follow a beta distribution, we elected to fit a beta distribution to each pair of replicate treatments, as well as to all negative controls from both replicates, using maximum-likelihood fitting as implemented in the fitdistr() function from the R package MASS (Venables and Ripley, 2002). Iterative optimization of shape parameters was initiated with settings x=PSI, shape1=1, shape2=1, method=“L-BFGS-B”, lower=0.01, and upper=mean number of reads supporting each PSI value. A modified Strictly Standardized Mean Difference (SSMD) (Zhang, 2007) was then calculated such that:

\[
\text{SSMD} = \frac{(\mu_t - \mu_c)}{\sqrt{\text{var}_t + \text{var}_c}}
\]

where \(\mu_t\) and \(\mu_c\) are the means (corresponding to the PSI), and \(\text{var}_t\) and \(\text{var}_c\) are the variances of the beta distributions fitted to the treatment replicates and negative controls, respectively. Events for which not all reads from at least one direction were informative [Mff (A2), Tead1 (A2)] were excluded from further analysis, as were events in individual treatments with less than 20 reads in one or both replicates (~9% of all events x treatment combinations). Additionally, the following events were removed from further splicing analysis: gene expression only events or constitutive exons, Fgf4, Gapdh, Sall4, Srpk2, and Srrm4; events with consistently low read counts, Atg13(A1), H2afy (A2), and Tcf7l1 (N2A only), and Dnmt3b (N2A only); or events where the measured inclusion was biased by differential length of isoforms (Fgfr1). Additionally, Dnmt3b
in CGR8 cells and Uspl1 (A3) were excluded from most analyses due to missing values in >10% of all treatments, resulting in a list of 52 filtered events. SSMD scores are provided in SD2-3.

**Differential expression analysis**

Read counts from all splice variants of each gene were used to estimate relative mRNA expression levels, reads per million reads (RPMs). Differential expression analysis based on raw read counts was performed using the generalized linear model workflow from the R package edgeR (Robinson et al., 2010). Because biases stemming from batch effects and position of the well on the plate affected the expression analysis more than the PSI (which is the ratio of two measurements), we used plate, position (edge/interior), and treatment as design factors (where all treatments of each type of control were treated as replicates.) Models were fitted using estimateDisp() to estimate the common, trended, and tag-wise dispersion for the CGR8 and N2A data separately. Subsequently, differences attributable to treatment, plate, or position contrasts were extracted with glmLRT() and represented as log2-fold changes with associated FDR. This approach showed that it was important to model plate and position explicitly. Fold-changes for treatments (SD2-4) are relative to the siNT and mock controls, while untreated controls were treated like experimental knockdowns and used to monitor the efficiency of normalization.

### 3.3.7 Correlation network analyses and functional cluster identification

To identify the appropriate SSMD cut-off, a cross-validation resampling approach was undertaken based on identification of functional groups from Enrichr (Chen et al., 2013). This approach randomly selected 30% of the events exceeding a cut-off and clustered the results using affinity propagation clustering (Bodenhofer et al., 2011; Frey and Dueck, 2007). To select number of clusters, a cut-off of h=0.4 was used for the merging objects. Each cluster was individually assessed for GO term and complex enrichment using Enrichr. For each cut-off this procedure was repeated 1000 times independently for both N2A and CGR8 samples. The reproducibility of the top functional groups at each cut-off was calculated. This identified a cut-off of 2.25 within the CGR8 samples and 4.75 within the N2A samples as providing the strongest enrichment of associated terms.

Events with SSMD scores above the identified cut-off were used. These events were clustered based on the ΔPSI values using affinity propagation clustering (Bodenhofer et al., 2011; Frey
and Dueck, 2007). A signed Pearson correlation was used to calculate pairwise correlation coefficients to dampen effects of diverging means and variances between samples. Gene enrichment for each cluster was assessed using the gene enrichment tool gprofiler (Reimand et al., 2007) using the following databases: GO, CORUM, REACTOME, and KEGG. A Benjamini-Hochberg FDR multiple correction with a p-value cut-off of 0.05 only including genes sets with maximum set size of 1000.

3.3.8 Hit frequencies of proteins with certain domains

The percentages of knockdowns that exceeded the threshold of ±3.0 SSMD for any AS event in N2A cells, among all proteins containing a certain PFAM domain was calculated. Only knockdowns with missing values for less than half of all events were considered. Separation into 'known' and 'new' AS regulators was based on association with any GO category whose name contained either of the strings 'mRNA splic' or 'spliceosom'.

3.3.9 Principal component analysis

The R function prcomp was used to derive principal components of either CGR8 or N2A SSMD values (uncentred, unscaled) from experimental knockdowns but not controls in order to avoid dominating effects of positive controls. Treatments with more than 10 missing values in each cell line were excluded. The full datasets including controls were then projected onto the obtained principal components. Outlines shown in Figure 3-13B represent the convex hull surrounding all points in the group.

3.3.10 Analysis of correlation within CORUM complexes

To score complexes for which knockdown of subunits resulted in correlated AS changes, all human and mouse complexes annotated in CORUM (Ruepp et al., 2010) were considered. For complexes isolated in both organisms, only the mouse variant was considered. When a complex was identified only in human cells, the mouse orthologs defined in InParanoid (Sonnhammer and Ostdlund, 2015) were used. Complexes in which less than three components were represented in the screen were discarded. Then, the average pairwise correlation of SSMD values between all components in a complex was calculated, and significance was assessed with a Mann-Whitney U-test between the correlations among the components in the complex and the pairwise
correlations of all non-control knockdowns in the cell line. The false discovery rate was adjusted using the Benjamini-Hochberg method.

3.3.11 RNA-Seq experiments and analysis

*siRNA knockdown*

For siRNA knockdown, cells were transfected with SMARTpool siRNAs (Dharmacon) using Lipofectamine RNAiMAX (Invitrogen), as recommended by the manufacturer. A non-targeting siRNA pool (siNT) was used as a control. Cells were harvested 48 hours post transfection, and total RNA was extracted with TRI Reagent (Sigma) or RNeasy columns (Qiagen).

*AS and gene expression analysis*

RNA-Seq data were processed using our AS and gene expression analysis pipeline, vast-tools version 1.0 (Braunschweig et al., 2014; Irimia et al., 2014), which is available on github (https://github.com/vastgroup/vast-tools). From primary output, events with poor coverage or junction balance were filtered out (vast-tools quality column score 3 other than SOK/OK/LOW for cassette exon [CE], microexon [MIC], and alternative 5' or 3' splice site [Alt5/3] events or coverage less than 15 reads for intron retention [IR] events; score 4 other than OK/B1 for CE and MIC events and score 5 of less than 0.05 for IR events). Differential AS was assessed through the vast-tools diff module available with the main pipeline. This strategy utilizes Bayesian inference, employing a Binomial likelihood function where the count of inclusion reads (K) follows \( K \sim \text{Binomial}(\Psi, N) \). \( \Psi \) represents PSI or PIR, and \( N \) is the total junction reads per-event. We apply an uninformative conjugate prior distribution (uniform Beta where \( \alpha=1, \beta=1 \)), and apply Bayes theorem to obtain the posterior distribution over \( \Psi \sim \text{Beta}(K + \alpha, (N-K) + \beta) \).

We combine biological replicates by sampling empirical posterior distributions of each replicate and fitting a new posterior Beta using maximum-likelihood (MLE) estimation with ‘fitdistr’ from the MASS package in R. The difference between two biological conditions, modelled as two posterior distributions \( X \sim \text{Beta}, \text{and } Y \sim \text{Beta}, \) follows in the form \( P(X-Y > 0) \). This probability can be estimated from the difference of empirical distributions sampled between X and Y such that \( P(X-Y > 0) = \Sigma_{i=1}^{n}(X_i - Y_i > 0) / N \). Significantly differential events were additionally required to have a PSI difference >10. Gene expression differences were calculated based on vast-tools raw read counts per gene. In cases with a single replicate (siNacc1, siMbnl, and siNT
counts were converted to read-per-million (RPM) and changes calculated as log2((1 + RPM[siSpecific]) / (1 + RPM[siNT])), and genes were required to have a vast-tools cRPKM ≥3 and a raw read count of ≥10 in at least one of the compared samples. In cases with multiple replicates (siZfp871, siSrrm4, and siNT control), differential expression was assessed with the R package edgeR.

3.3.12 Gene Ontology analyses

FuncAssociate (Berriz et al., 2009) was used to find over-represented GO terms associated with genes with changes in CE and MIC events that were significant and greater than 10 PSI. As a background, all genes with measured PSI values that survived filtering were used. Only terms with a minimum odds ratio of 4 and less than 1,000 associated genes were plotted. If two categories mutually overlapped by more than 70% of associated genes, only the category with stronger enrichment was shown. Adjusted P-values represent P-values derived by iterative simulation in FuncAssociate.

3.3.13 Analysis of features associated with Zfp871-regulated events

SVM-BPfinder (Corvelo et al., 2010) was used to assign the most likely branchpoint and associated polypyrimidine tract length of introns upstream of regulated and non-regulated exons. MaxEntScan (Yeo and Burge, 2004) was used to assess splice site strength.

3.3.14 ChIP-Seq experiments and analysis

Chromatin immunoprecipitation (ChIP) was performed as previously described (Najafabadi et al., 2015). Briefly, ~20 million N2A cells were crosslinked in 1% formaldehyde. Following sonication of DNA fragments, Nacc1 was immunoprecipitated from the lysate with 4µg of Nacc1 antibody (Abcam ab29047) followed by crosslink reversal and DNA precipitation. Libraries were sequenced on the Illumina HiSeq 2500 to a depth of 20 million 51-nucleotide single end reads.

Illumina adapter sequences were removed from 3'-ends of 51-nt reads and remaining reads were mapped to the mouse genome, NCBIIm37/mm9, using bowtie2 (Langmead and Salzberg, 2012) with default settings. After removal of duplicate reads, peaks were called jointly on immunoprecipitated and input samples with MACS 1.4 (Zhang et al., 2008a).
3.3.15 iCLIP

iCLIP experiments

Individual nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) was performed as described previously (Huppertz et al., 2014). Briefly, Nacc1 was immunoprecipitated from UV-crosslinked N2A cells using anti-Nacc1 antibody (Abcam ab29047). Zfp871 was immunoprecipitated from N2A cells engineered to express Flag-tagged Zfp871 under the control of doxycyclin using the PiggyBac system (Woltjen et al., 2009). Flag-Zfp871 expression was induced for 24 hours prior to crosslinking (0.15 J/cm²) at 254 nm with a Stratalinker 1800. For Nacc1, three independent cell pellets, collected from 1 x 150 mm dish, were used for the generation of iCLIP libraries. For Zfp871, two replicates from two independent single-cell clones expressing Flag-tagged Zfp871 were used for generating a total of four iCLIP samples.

Lysates generated from the crosslinked cells were treated with Turbo DNase (Ambion) and RNase I (1:500 for Nacc1 and 1:100 for Zfp871; Ambion) for 5 min at 37°C to digest the genomic DNA and trim the RNA to short fragments of an optimal size range. RNA-protein complexes were immunoprecipitated using 100 µl of protein G Dynabeads (Life Technologies) and 10 µg of anti-Flag (Sigma) or 10 µg of anti-Nacc1 (Abcam) antibodies. Following stringent high salt washes, the immunoprecipitated RNA was 5' end-labeled using radioactive ³²P isotopes followed by on-bead-ligation of pre-adenylated adaptors to the 3' end. The immunoprecipitated complexes were separated with SDS-PAGE and transferred to a nitrocellulose membrane (Protran). RNA was recovered by digesting proteins using proteinase K and subsequently reverse transcribed into cDNA. The reverse transcription primers include barcode sequences to enable multiplexing and a BamHI restriction enzyme site.

For Nacc1 the barcoded primers used were

Rt1clip: /5Phos/NNAACCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCCGC,
Rt10clip: /5Phos/NNGACCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC, and
Rt13clip: /5Phos/NNTCCGNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC.

For Flag-Zfp871 the barcoded primers used were

Rt1clip: /5Phos/NNAACCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCCGC,
Rt9clip: /5Phos/NNGCCANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC,
Rt10clip: \texttt{/5Phos/NNGACCNNAGATCGGAAGAGCGTCGTG}\texttt{gatcCTGAACCGC, and}
Rt13clip: \texttt{/5Phos/NNTCCGNNNAGATCGGAAGAGCGTCGTG}\texttt{gatcCTGAACCGC}.

The cDNA was size selected (low: 70 to 85 nt, middle: 85 to 110 nt, and high: 110 to 180 nt),
circularized to add the adaptor to the 5' end, digested at the internal BamHI site, and then PCR
amplified using AccuPrime SuperMix I (Life Technologies). The final PCR libraries were
purified on PCR purification columns (Qiagen) to remove residual PCR reagents, and a ratio of
1:5:5 from the low, middle, and high fractions were submitted for sequencing.

\textit{Analysis of iCLIP data}

51-nt raw reads that consisted of 3 random positions, a 4-nt multiplexing barcode, and another 2
random positions, followed by the cDNA sequence, were initially trimmed to 49 nt from the 3'-
end, and duplicates were discarded. Reads were de-multiplexed, and the random positions,
barcodes, and any 3'-bases matching Illumina adaptors were removed. Remaining reads longer
than 25 nt were mapped to the mouse genome/transcriptome (Ensembl annotation of NCBI\textit{m37})
using tophat with default settings. To prevent false assignments of reads from repetitive regions,
any reads with a mapping quality < 3 were removed from further analysis.

Plots showing average crosslinking signal of events aligned to exon borders were generated as
described previously for ChIP-seq data (Braunschweig et al., 2014), except that reads were first
reduced to their first position, which is adjacent to the crosslink position, and no normalization
against a control was performed. A 21-bp running window average was used for display only,
and average signals across replicates are shown.

For numeric analysis of intronic signal, the number of crosslinks per position and million of
sequenced reads was calculated per replicate and averaged across replicates. Mean intronic
signals were compared between groups of AS events using the one-sided Mann-Whitney U-test.

\textbf{3.3.16 \textit{In vitro} splicing assays}

Preparation of whole cell splicing extracts and purification of recombinant proteins have been
previously described in detail (Calarco et al., 2009). The Myo9b \textit{in vitro} splicing reporter was
constructed by amplifying the mouse genomic DNA region, including the specific alternative
exon, its flanking introns, as well as its constitutive exons. \textit{In vitro} splicing assays performed in a
volume of 20 µL contained 1.5 mM ATP, 5 mM creatine phosphate, 5 mM DTT, 3 mM MgCl₂, 2.6% PVA, 30 units of RiboLock RNase inhibitor (Thermo Scientific), 20 ng of splicing substrate, 50-60 µg of splicing extract, and up to 12 µL of splicing buffer (20 mM HEPES-KCl pH 7.9, 100 mM KCl, 0.2 mM EDTA, 10% glycerol, 1 mM DTT) with or without the addition of recombinant proteins. Reactions were incubated at 30°C for one hour. RNA was extracted using TRI Reagent (Sigma) and then resuspended in 10 µl of DEPC-treated water. Spliced products were amplified by RT-PCR assays using 2 µl of the recovered RNA and primers specific for Myo9b upstream and downstream constitutive exons. RT-PCR products were resolved on a 3% agarose gel.

3.3.17 Protein extraction and western blotting

Cell pellets were lysed in radioimmunoprecipitation assay (RIPA) buffer by brief sonication. 30-150 µg of protein lysate was separated on a 10% SDS-polyacrylamide gel and transferred to a PVDF membrane. The membranes were blotted with the following antibodies: anti-Flag M2 (1:1500, Sigma), anti-Nacc1 (1:7500, Abcam), and anti-α-tubulin (1:5000, Sigma). Secondary antibodies (GE Healthcare) and chemiluminescence reagents (Perkin Elmer) were used as per the manufacturer’s instructions.

3.3.18 Co-immunoprecipitation assay

293T cells were transiently transfected with Flag-Srsf2, HA-Arglu1, or both using Lipofectamine 2000 (Life Technologies). Cells were lysed in 0.5% TNTE buffer. After pre-clearing with Protein G-Dynabeads (Life Technologies), lysates were incubated with anti-Flag M2 antibody (Sigma) bound to Protein G-Dynabeads for 1 hour at 4°C. Subsequently, immunoprecipitates were washed 5X with 0.1% TNTE buffer, subjected to SDS-PAGE, transferred onto a nitrocellulose membrane, and immunoblotted with the anti-HA antibody (Roche) or anti-Flag M2 antibody (Sigma).

3.3.19 RNA extraction and (q)RT-PCR assays

Total RNA was extracted using RNeasy Mini Kit (Qiagen) or TRI Reagent (Sigma), and RT-PCR assays were performed using the OneStep RT-PCR kit (Qiagen), as per the manufacturer’s instructions. 20 ng of total RNA was used per 10 µL of reaction. The number of amplification cycles was 22 for Gapdh, and 27-32 for all other transcripts analyzed. Reaction products were
separated on 1-3% agarose gels. Quantification of isoform abundance was performed as previously described (Han et al., 2013).

For quantitative (q)RT-PCR, first-strand cDNAs were generated from 1-3 µg of total RNA using Maxima H Minus First Strand cDNA synthesis Kit (Thermo Scientific), as per the manufacturer’s recommendations, and diluted to 20 µg/µL. qPCR reactions were performed in a volume of 10 µL using 1µL of diluted cDNA and FastStart Universal SYBR Green Master (Roche Applied Science). Primers used for PCR reactions are available upon request.

3.3.20 Cloning and plasmids

Flag-tagged Zfp871 was cloned into the PiggyBac destination vector from cDNA derived from N2A cells using the Gateway system. The primers used for cloning are:

forward 5’-GGGGACAAGTTTTGTACAAAAAAGCAGGCTTCACC GAGTCAGTG CCTTTGAGGATGTG - 3’ and reverse 5’ - GGGGACCACTTTGTACAAAAAAGCAGGCTTCACC GAGTCAGTG CCTTTGAGGATGTG - 3’.
3.4 Results

3.4.1 A high-throughput screen linking trans-acting factors to endogenous AS events

To systematically discover new factors that control endogenous AS regulatory networks that control cell fate, I developed and applied SPAR-Seq to interrogate 1,536 siRNA knockdown and control treatments for effects on 52 evolutionarily conserved (i.e. between human and mouse) AS events that are associated with ESC pluripotency, neural differentiation, and transitions in somatic cell reprogramming (Figure 3-1A; SD 2-1) (Gabut et al., 2011; Han et al., 2013; Kolle et al., 2011; Ohta et al., 2013; Rao et al., 2010; Salomonis et al., 2010; Venables et al., 2013). In brief, the 52 prioritized AS events were simultaneously amplified by multiplex RT-PCR in 96-well plates, following well-specific siRNA knockdown conditions. Unique, dual-index barcodes were added to the RT-PCR amplicons to mark each well. Subsequent pooling and high-throughput sequencing analysis of the barcoded amplicons was then used to link the well-specific knockdowns to changes in AS (Figures 3-1A and 3-1B).

AS events analyzed in the screen comprise 35 single cassette alternative exons, including eight microexons (3-27 nucleotides in length), 15 alternative exons from eight genes with more complex AS patterns, and two novel alternative 3’ splice site events in Foxm1 that were identified through de novo splice junction mapping. The screen was performed in both mouse ESCs (CGR8) and neuroblastoma cells (N2A) so as to identify both positive and negative regulators of the assayed AS events. Genes subject to knockdown included a comprehensive set of 654 known and putative splicing/RNA associated factors, including all annotated spliceosomal-associated proteins, and all known and predicted RNA binding proteins (RBPs) (Figure 3-1C; SD 2-2). We also interrogated 858 genes with functional annotations linked to chromatin and transcription that are expressed in CGR8 or N2A cells (Materials and Methods). Among the annotated chromatin and transcription factor genes, 161 overlap those linked to splicing and RNA binding. Additionally, we assayed 65 factors with links to AS regulation that function in signalling and post-translational modification (Figure 3-1C; SD 2-2).
Figure 3-1 Systematic Parallel Analysis of endogenous RNA regulation coupled to barcode Sequencing (SPAR-Seq).

(A) Schematic outline of the SPAR-Seq strategy. Total RNA sample libraries from 1536 siRNA knockdown and control treatments applied to CGR8 and N2A cells were prepared in biological replicates. A two-step, PCR-based amplification and indexing protocol was used to generate multiplexed barcoded libraries for SPAR-Seq, monitoring more than 50 endogenous AS and gene expression events. (B) Schematics of SPAR-Seq library. A total of four independent reads were generated during the sequencing run, Read 1 (R1), Read 2 (R2), and two barcode sequencing reads. (C) Functional categories and Gene Ontology (GO) annotations associated with genes assayed by SPAR-Seq. Genes subject to siRNA knockdown belong to four main groups, splicing factors and RNA binding proteins (Splicing/RBP, orange), chromatin and transcription factors (Chromatin/TF, blue), factors overlapping both groups (Both), as well as signalling and post-translational factors (S, green). Number of targeted genes is indicated below.
Total RNA was harvested 48 hours post transfection of siRNAs and subject to SPAR-Seq. A custom analysis pipeline was designed to extract data, quantify AS levels by calculating percentage of transcripts with the exon spliced in (‘PSI’) values, and prioritize detected PSI changes for further analysis using the ‘Strictly Standardized Mean Difference’ (SSMD; Zhang, 2007) metric (Figures 3-2, 3-3A and 3-3B; SD 2-3; Materials and Methods). SSMD measures effect size while taking the variance between replicates into account. In total, 316,863 AS measurements were analyzed across two biological replicate screens in the two cell lines. The SPAR-Seq data were additionally used to monitor the mRNA expression levels for all genes assayed for splicing changes, as well as for a representative set of splicing factors (Figure 3-2; SD 2-4; Materials and Methods). These data indicated efficient (>60%) depletion of all monitored genes for which knockdowns were carried out (Figure 3-3C). RT-(q)PCR experiments using independent control and knockdown samples from CGR8 and N2A cells validated SPAR-Seq-detected changes for nearly all analyzed AS and mRNA expression level changes. Furthermore, the magnitude of the screen-detected PSI changes correlated well with PSI changes measured using independent RT-PCR assays (r=0.77, Pearson correlation, n=336) (Figure 3-4) and RNA-Seq.

Figure 3-2 Flowchart outlining pipeline for analysis of SPAR-Seq data.
Main steps of splicing and expression analyses are shown. See Materials and Methods for further details. PSI, percent spliced in; SSMD, strictly standardized mean difference; RPMs, reads per million reads.
Figure 3-3 Overall assessment of the SPAR-Seq strategy.

(A) Cumulative distribution of mapped reads per treatment (sum of all endogenous events monitored). Minimum of forward and reverse read counts are shown. (B) Percentage of treatments for each AS event that surpassed the filtering threshold of 20 reads in both replicates. mRNA expression of the gene harboring the AS events is indicated by colored squares. (C) Reduction of mRNA levels for twelve genes that were both knocked down and monitored by SPAR-Seq. Means are shown as dashed lines. RPM, reads per million mapped reads of monitored genes in control CGR8 and N2A cells.
AS events with SSMD scores in different ranges were validated by RT-PCR assays using independent control and knockdown samples. The Pearson correlations for the entire set and for each SSMD range are shown.

Hierarchical clustering of the data revealed that the negative control conditions (32 siNT, non-targeting siRNA; 32 mock controls; 24 untreated samples) uniformly resulted in little to no changes in splicing levels, whereas positive controls (32 siMbnl, simultaneous knockdown of Mbnl1 and Mbnl2) resulted in significant changes for many ESC-differential AS events towards an ESC-like pattern, without substantially affecting neural-enriched AS events, consistent with our previous results (Figures 3-5 and A2-1A) (Han et al., 2013). Individual knockdown of Mbnl proteins and additional splicing regulators previously linked to pluripotency and reprogramming, including Rbfox2, Son, Srsf2, Srsf3, and U2af1 (Lu et al., 2013; Lu et al., 2014; Ohta et al., 2013; Venables et al., 2013), also affected ESC-differential AS events (Figure 3-5; SD 2-3). Conversely, knockdown of the neuronal-specific splicing regulator nSR100/Srrm4 (Calarco et al., 2009) in N2A cells affected neural-enriched AS events, without substantially affecting ESC-differential AS events (Figure 3-5). Additional analyses further demonstrated the specificity and reproducibility of the screen data, both within and between replicate experiments (Figure A2-1B; Materials and Methods). In total, knockdown of 220 and 416 factors in CGR8 and N2A cells, respectively, resulted in significant changes in one or more endogenous AS events (Figure 3-5, bar graphs).
**Figure 3-5 Systematic identification of splicing regulators in CGR8 and N2A screens.**

Heatmaps of SSMD scores for representative screen results from mouse ESCs (CGR8) and neural cells (N2A). All positive (siMbnl) and negative (siNT, Mock, and Untreated) control treatments, as well as siRNA knockdown treatments of splicing factors Mbnl1, Mbnl2, and Srrm4 are shown. Endogenous AS events monitored in the screen are grouped into ESC-high, ESC-low, neural-enriched, and others. ESC-high and ESC-low refer to exons that are preferentially included and skipped in ESCs relative to other cell and tissue types, respectively. Two novel AS events detected by de novo alignment are shown at the bottom. Multiple events within the same gene are denoted as A1 up to A3. Columns (treatments) for controls are clustered in the same way for CGR8 and N2A, and rows (events) are clustered within each group. See Figure 2A-1A for full screen results. Bar graphs show numbers of knockdowns leading to a significant increase (yellow) or decrease (blue) in the inclusion level of each exon above negative controls.
3.4.2 Correlated AS changes reveal multilayered regulatory pathways and complexes

To systematically investigate functional relationships between factors that impact AS in our dataset, we determined overall correlations between the changes in PSI values involving the 52 AS events, from all pairwise comparisons of knockdowns in CGR8 and N2A cells. Clustering of the resulting data based on the degree of pairwise correlation similarity revealed discrete groups of factors that positively or negatively regulate specific subsets of AS events in a similar manner (Figure 3-6 and 3-7; SD 2-5; Materials and Methods). Consistent with the results from recent screens for splicing regulators of AS events associated with apoptosis and cell proliferation (Papasaikas et al., 2015; Tejedor et al., 2015), these groups are often significantly enriched in factors that function in distinct biological processes and pathways, and in many cases are known to physically interact within complexes. Complementing this approach, we also systematically surveyed factors in experimentally-defined complexes (from CORUM; Ruepp et al., 2010) for coordinated AS changes by scoring mean pairwise correlations between complex members (Figure 3-8). Remarkably, of 316 complexes represented by at least three factors subject to knockdown in our screen, 38 (12%) in CGR8 and 187 (59%) in N2A cells displayed significant correlations above background (FDR < 0.05, Mann-Whitney U-test), even though most of the complexes were defined in other cell types. The SPAR-Seq data thus provides a valuable means with which to link protein complexes to the regulation of specific subsets of AS events associated with cell fate, as well as to infer new physical and functional interactions between trans-acting factors that impact AS.
Figure 3-6 Correlations of AS changes in the CGR8 screen reveal protein complexes and pathways.

Symmetrical heatmap of pairwise correlation of AS changes in CGR8 cells. Knockdowns are initially clustered by information propagation, with subsequent hierarchical clustering. Sub-clusters are annotated with the most significant terms from GO, CORUM, REACTOME, and KEGG. Numbers of genes annotated with each respective term and associated enrichment p-values are indicated. Cluster numbers in parentheses are referred to in the main text. Additional details in SD 2-5 and Materials and Methods.
Figure 3-7 Correlations of AS changes in the N2A screen reveal protein complexes and pathways.

Symmetrical heatmap of pairwise correlation of AS changes in N2A cells. Knockdowns are initially clustered by information propagation, with subsequent hierarchical clustering. Sub-clusters are annotated with the most significant terms from GO, CORUM, REACTOME, and KEGG. Numbers of genes annotated with each respective term and associated enrichment p-values are indicated. Cluster numbers in parentheses are referred to in the main text. Additional details in SD 2-5 and Materials and Methods.
Figure 3-8 Complexes and pathways regulate AS in CGR8 and N2A cells.

Pairwise correlations of AS changes upon knockdown of components of CORUM complexes, where correlations represent averages for at least three components of a complex. Complexes are sorted according to the highest correlation after averaging values between CGR8 and N2A cell lines. FDR reflects significance of mean pairwise correlations when compared to a background distribution of all pairwise comparisons (see Materials and Methods for details).
As an example, one of the largest groups of correlated knockdowns in CGR8 cells involves factors that are functionally associated with the 17S U2 snRNP (Figure 3-6, cluster 8), which binds the branch site region in pre-mRNA during spliceosome formation (Wahl et al., 2009). Remarkably, the degree of correlation between factors linked to U2 snRNP function reflects physical relationships among these components (Papasaikas et al., 2015). More specifically, five members of the heptameric Sm complex form a central, highly correlated core module adjacent to additional major (i.e. SF3a/b) U2 snRNP components. In contrast, factors with auxiliary roles in facilitating U2 snRNP recruitment to pre-mRNA are less well correlated with core U2 snRNP proteins and form a distinct cluster (Figure 3-9A). These include U2af1 and U2af2, which bind to polypyrimidine tract-3’ splice site sequences to promote U2 snRNP assembly at the branch site (Wahl et al., 2009), as well as Aqr and Isy1, components of the intron binding complex (IBC), which directly interact with U2 snRNP in activated spliceosomes (De et al., 2015). An additional, distinct sub-cluster comprises the Rbm17, Cherp, U2surp, and Dhx15 factors associated with the spliceosomal A complex (Agafonov et al., 2011). Rbm17 controls splice site selection by interacting with U2 snRNP (Corsini et al., 2007) and by recognizing the 3’ splice site AG during the second catalytic step of splicing (Lallena et al., 2002), and interacts with U2surp and Dhx15 (Hegele et al., 2012). The highly correlated effects of knockdown of the latter factors suggest that they have closely related functions in the regulation of a subset of AS events linked to cell fate.

Importantly, knockdown of U2 snRNP-associated factors negatively regulated a sub-network of ESC-differential AS events in genes with diverse functions, including Numb, Map3k7, Mbnl2, Senp6, and Tcf7l1 (Figure 3-6; SD 2-3 and SD 2-5). These results are consistent with the observation of a significantly higher degree of expression of U2 snRNP-associated factors in ESCs and reprogramming cells compared to differentiated cells/tissues (Figure 3-9B) (Han et al., 2013; Hirsch et al., 2015). Collectively, these results suggest that increased levels of U2 snRNP-associated components may play an important role in establishing and maintaining ESC fate, whereas reduced levels of these components are associated with differentiated cells.
Figure 3-9 The functional implication of U2 snRNP-associated factors.

(A) Hierarchical clustering of components of 17S U2 snRNP analyzed by SPAR-Seq, based on similarity between pairwise correlations of SSMD scores. (B) Heatmap of relative expression (z-scores across cell/tissue samples) of the 17S U2 snRNP components shown in (A), in ESCs/iPSCs and diverse other differentiated cell lines/tissues. Clustering of factors on the x-axis is identical to (A). Along the y-axis, samples are first sorted into the indicated groups and then clustered within each group.

3.4.3 Novel positive and negative roles of different gene regulatory layers in controlling AS

The results described above exemplify how our screen data reveal important functional and physical relationships between factors that control distinct subsets of AS events linked to cell fate. Other clusters of significantly correlated knockdowns include factors that form additional sub-complexes involved in splicing, but also complexes and pathways associated with RNA polymerase II transcription, chromatin modification, DNA binding/transcription factor activity, mRNA surveillance, turnover, transport, and translation (Figures 3-6, 3-7, and 3-8). Many of the factors belonging to these clusters have not been previously linked to splicing regulation.

Knockdown of multiple factors associated with the exon junction complex (EJC), which functions in the control of splicing-dependent mRNA export and turnover via the nonsense
mediated mRNA decay (NMD) pathway (Tange et al., 2004), displayed significantly correlated effects on AS. Similar to the findings for U2 snRNP-associated factors, knockdown of the ‘core’ EJC components Eif4a3, Magoh, Rbm8a, and Casc3 resulted in more highly correlated effects on AS patterns compared to knockdown of ‘non-core’ EJC factors (i.e. Ddx39b, Alyref, Srrm1, and Rnps1) (Figure 3-10A). Importantly, while knockdown of these factors resulted in increased levels of premature termination codon (PTC)-containing splice variant transcripts from the Ptbp2 gene, they also strongly affected frame-preserving AS events in the Senp6, Csnk1g3, Mbnl1, and Mbnl2 genes in a correlated manner. While the peripheral EJC factors Srrm1 and Rnps1 have reported functions in splicing regulation (Blencowe et al., 1998; Eldridge et al., 1999; Sakashita et al., 2004), the present data are consistent with emerging evidence that core EJC-associated components also regulate AS (Ashton-Beaucage et al., 2010; Michelle et al., 2012; Papasaikas et al., 2015). Moreover, our results link these factors to the control of a specific subset of AS events associated with cell fate.

In addition to identifying positively-correlated effects of factor knockdowns, our data also revealed numerous and unexpected anti-correlated or opposing effects, thereby illuminating possible antagonistic physical and functional relationships between splicing regulators (Figure 3-6 and 3-7). For example, knockdown of factors associated with transcription/DNA binding activity (e.g. Nfyb, Tbx1, and Tfdp1) (Figure 3-6, cluster 10) resulted in AS changes in Foxm1 that are anti-correlated with the knockdown of a subset of splicing associated factors (e.g. Snrnp200, Bcas2, Rbfox2, and Mbnl proteins) (Figure 3-6, cluster 3) (p < 0.01, one-sided binomial tests). Antagonistic (as well as positive) effects on AS were also observed between Arg/Ser-repeat (RS) domain-containing proteins and other factors (Figure 3-6, e.g. clusters 2 and 5). As a new example, knockdown of the Arginine and Glutamate rich 1 (Arglu1) protein and Srsf2 had opposing effects on six AS events monitored in the screen (Figures 3-10B and 3-10C). Interestingly, Arglu1 has previously been associated with estrogen receptor-mediated gene activation (Zhang et al., 2011), whereas the presence of an RS domain in this protein suggested that it also has a function in AS (Boucher et al., 2001). Our data, as well as results from an independent study (L.M. and C.L.C, personal communication), confirm that Arglu1 is an AS regulator. Since RS domains function in mediating interactions between splicing factors (Lin and Fu, 2007), and because factors that associate physically often have related knockdown AS profiles in our screen, we asked whether Arglu1 and Srsf2 might antagonize each other by
interacting. Co-immunoprecipitation in the presence of nuclease treatment followed by western blotting of FLAG-Srsf2 and HA-Arglu1 proteins shows that these proteins can interact (Figure 3-10D). These results thus highlight Arglu1 as a new splicing regulator that may function by physically antagonizing Srsf2.

Figure 3-10 Novel positive and antagonistic effects on AS.

(A) Hierarchical clustering of components of EJC analyzed by SPAR-Seq, based on similarity between pairwise correlations of SSMD scores. (B) Heatmap representation of SSMD scores upon knockdown of Arglu1 and Srsf2, sorted by SSMD values in the Arglu1 knockdown. AS event groups are indicated by colors as in Figure 3-5. Asterisks indicate events tested in panel C. (C) Representative RT-PCR assays detecting AS changes in Mllt4, Spag9, Madd, Itga6, and Mbnl2 transcripts upon knockdown of Arglu1 and Srsf2. Percentage of transcripts with the exon spliced in (PSI) quantitation is indicated below each gel image. (D) Western blot analysis of inputs and immunoprecipitates (IP) from 293T cells expressing Flag-Srsf2, HA-Arglu1, or both. Co-immunoprecipitation was performed with anti-Flag antibody and blots were probed with anti-Flag or anti-HA antibody.
3.4.4 Chromatin and transcription factors function as AS regulators

Knockdown of splicing/spliceosomal-associated factors frequently resulted in AS changes in both CGR8 and N2A cells, as expected. Strikingly, in N2A cells, knockdown of chromatin/transcription factors affected AS with a similar frequency to that observed for splicing/spliceosomal-associated factors (Figures 3-11A and A2-2). This is also reflected by the observation that knockdown of factors containing chromatin and transcription factor-related domains, such as the PWWP, chromodomain, and various annotated DNA binding domains, impacted AS in N2A cells at a comparable frequency as factors containing annotated RNA binding or RNA helicase domains (Figure 3-11B).

As examples, knockdown of multiple components of RNA polymerase II (Pol II)-containing complexes had relatively pronounced effects on AS. Interestingly, knockdown of core Pol II subunits, including the largest (catalytic) subunit Polr2a strongly correlated with each other, whereas knockdown of the Pol II initiation complex factors TFIIB and TFIIID had distinct effects (Figures 3-8 and 3-12). This reveals that AS is differentially impacted by perturbing temporarily distinct Pol II complexes. Knockdown of multiple structural maintenance of chromosomes (SMC) proteins forming the cohesin complex also markedly affected AS (Figure 3-8), consistent with previous evidence of physical and functional links between SMC proteins and splicing factors (McCracken et al., 2005; Mercer et al., 2013). Furthermore, knockdown of components of complexes regulating histone modification (e.g. the PTIP-HMT, TLE1-corepressor, ING4-containing, and HBO1 complexes), as well as DNA repair (e.g. the BLM complex III and BRCA1-BARD1 complex) also displayed significantly correlated effects on AS. Interestingly, a BRCA1-containing complex has previously been shown to recruit splicing factors to gene promoters upon DNA damage (Savage et al., 2014). Our results support a role for BRCA1 complexes in AS regulation and are further consistent with links between AS and multiple different DNA repair pathways.
Figure 3-11 Chromatin and transcription factors frequently regulate AS in N2A cells.

(A) Percentages of knockdowns affecting at least one AS event with an $|\text{SSMD}| > 2.25$ in CGR8 and 3.00 in N2A. These thresholds robustly discriminated between positive and negative controls. See also Figure A2-2. (B) Percentages of knockdowns of factors containing at least one of the indicated domains that were scored as a hit in N2A cells at the same threshold as in (A). Only domains occurring in at least ten of the factors analyzed in the screen are shown (refer to SD 2-2 for domain annotations). Absolute numbers of factors identified as regulators are indicated. Bars are sorted by the percentages of newly discovered AS regulators, defined as a factor not previously annotated with a GO category related to splicing (see Materials and Methods).
Figure 3-12 The functional implication of RNA Pol II-related factors.

Hierarchical clustering of components of the RNA Pol II holoenzyme and basal transcription factors analyzed by SPAR-Seq, based on similarity between pairwise correlations of SSMD scores.

3.4.5 An extensive role for zinc finger proteins in AS regulation

Remarkably, the screen data also revealed that knockdown of 57% (137/242) of factors with zinc finger (ZnF) domains affected one or more AS events. Among this set are 42% (49/116) of analyzed C2H2 ZnF proteins, which displayed striking effects on AS in N2A cells (Figure 3-11A). This observation is of particular interest because C2H2 ZnF proteins represent the largest yet least well understood class of nucleic acid binding proteins, comprising 718 family members in the human genome and 583 members in the mouse genome (Emerson and Thomas, 2009; Tadepally et al., 2008). While the majority of analyzed C2H2 ZnF proteins bind DNA and a subset have been shown to control transcription or silence retrotransposition, the functions of the vast majority are not known (Stubbs et al., 2011). Notably, whether or not a C2H2 ZnF protein is scored as a ‘hit’ in our screen data is independent of the presence of other effector domains in these proteins (data not shown), indicating that the ZnF domains in these proteins are an important determinant of function in AS regulation.

Importantly, hierarchical clustering (Figure 3-13A) and principal component analysis (PCA; Figure 3-13B) showed that knockdown of C2H2 ZnF proteins in N2A cells affected different subsets of AS events in a positive or negative manner, with changes in magnitude that are
comparable to those following knockdown of defined splicing regulators. Interestingly, some of the most pronounced AS changes are associated with knockdown of Gtf3a (also known as TFIIIA), Yy1, Repin1, and Rest/Nrsf, all of which have previously been reported to bind RNA (Cassiday and Maher, 2002; Jeon and Lee, 2011; Lu et al., 2003; Pelham and Brown, 1980). Moreover, knockdown of Rest, a repressor of neurogenesis genes in non-neural cells (Chen et al., 1998) shifted splicing patterns of several AS events in an opposite direction to that of knockdown of Srrm4 (Figure 3-13B). This observation is consistent with our previous finding that Rest transcriptionally represses Srrm4, whereas Srrm4 promotes the inclusion of neural-enriched exons, including an exon that silences Rest activity (Calarco et al., 2009; Raj et al., 2011).
Figure 3-13 AS regulation by C2H2 ZnF proteins.

(A) Heatmap and hierarchical clustering of AS changes (represented by SSMD scores) upon knockdown of all C2H2 ZnF factors identified as hits in N2A cells, together with developmentally regulated splicing factors. Only AS events and C2H2 ZnF factors with at least one change >3 SSMD are shown. Expression changes of representative splicing factors, Mbnl1, Mbnl2, Srrm4, and Rbfox2 are shown. (B) Principal component analysis (PCA) of AS changes caused by knockdowns of C2H2 ZnF factors and several protein complexes for comparison. Outlines represent the area containing all single knockdowns in the group. PCA was done on all experimental but not control knockdowns, and the full data were then projected back onto the principal components. The percentage of total variance explained by each component is indicated in corresponding axis labels. The 55% explained by PC1 represents the coherent behaviour of multiple ESC-differential AS events.

A new and particularly interesting AS regulator from the screen is Zfp871, which contains a KRAB domain in addition to 15 C2H2 ZnF domains. Knockdown of Zfp871 affects neural-enriched exons in a manner similar to that of Srrm4 (Figure 3-13). Moreover, analysis of RNA-Seq profiling datasets from several stages of differentiation of ESCs to glutamatergic neurons (Hubbard et al., 2013) reveals that Zfp871 displays a marked increase in expression as neurons form, in a manner similar to that of Srrm4 (Raj et al., 2014) (Figure 3-14A). To investigate the function of Zfp871 as well as its relationship with Srrm4 in greater detail, I performed RNA-Seq to analyze AS and gene expression changes following knockdown of each protein in N2A cells (Figures 3-14B and 3-14C). Knockdown of Zfp871 resulted in changes (ΔPSI>10) in the inclusion of 479 exons, of which 189 are neural-differential, and 198 overlap those regulated by Srrm4 (both p<0.001, Fisher's exact test; Figure 3-14B). RT-PCR validation experiments confirmed changes for all ten analyzed Zfp871-dependent and -independent neural-differential AS events detected by SPAR-Seq and/or RNA-Seq (Figure 3-15). Similar to Srrm4 (Irimia et al., 2014), Zfp871 more frequently impacts the splicing of short neural-differential exons, including 3-27 nt microexons, and in the majority of cases promotes their inclusion. Indeed, Zfp871 and Srrm4 significantly more often regulated overlapping sets of shorter neural exons, but had disparate effects on longer exons (Figure 3-14B) (p<0.001, Fisher's exact test for exons ≤30 nt vs. longer exons). Consistent with these results, genes with AS events impacted by Zfp871 knockdown are significantly enriched in functional categories associated with neuronal morphology and function (Figure 3-16).
Figure 3-14 The C2H2 ZnF factor Zfp871 regulates neural-enriched exons and microexons.

(A) mRNA expression of Zfp871 and Srrm4 during in vitro differentiation of ESCs to glutamatergic neurons. (B) Alternative exons of different lengths affected by knockdown of Zfp871 or Srrm4, as analyzed by RNA-Seq. Bar height indicates percentage of the changing exons in the length category and absolute numbers of changing exons are also indicated. Greyscale heatmaps indicate overlap with exons in each length group that show AS changes upon knockdown of Srrm4, and with neural-differential exons. (C) Change in mRNA expression of splicing factors upon knockdown of Zfp871 or Srrm4. Factors that were knocked down are indicated in dark grey, and blue indicates significant differences (FDR < 0.05).
Figure 3-15 RT-PCR validations of AS regulation by Zfp871.

Representative RT-PCR validations of AS events identified by SPAR-Seq and/or RNA-Seq that are co-regulated by Zfp871 and Srrm4, that are regulated independently of Srrm4 (i), or that are not regulated by either protein (c-). PSI quantitation is indicated below each gel image.
GO-enrichment analysis of genes harbouring changing exons and microexons upon knockdown of Zfp871.

GO-enrichment analysis of genes with exons and microexons that show AS changes following knockdown of Zfp871. Only categories with < 1000 associated genes and an enrichment of > 4-fold are shown, and overlapping categories have been merged. Abscissa shows the enrichment odds-ratio. The list of genes containing exons for which splicing could be ascertained served as a background.

We next investigated the mechanisms by which Zfp871 regulates neural-differential and/or Srrm4-dependent exons, and considered the possibility of both indirect and direct roles. Consistent with an indirect regulatory role, knockdown of Zfp871 resulted in a ~40% reduction in Srrm4 mRNA levels, and it also reduced the levels of several other known splicing regulators linked to neural AS, including Celf4, Raver1, Nova1, and Ptbp2 (Figure 3-14C). Remarkably, using individual-nucleotide resolution cross-linking and immunoprecipitation coupled to sequencing (iCLIP) we also observed that Zfp871 binds RNA sequences proximal to its regulated target exons in N2A cells. More specifically, Zfp871 displays significant binding to exonic sequences, but it also binds preferentially to intronic sequences adjacent to its target alternative exons that it regulates, in particular those exons for which it promotes inclusion (Figures 3-17A and 2A-3). These increases in binding occupancy are not due to increased expression of the corresponding genes or retention of bound introns (data not shown). Furthermore, the binding profile of Zfp871 is similar to that of Srrm4, with increased occupancy within a ~50 nt region upstream from the 3´ splice site, coinciding with binding peaks observed for Srrm4 (Raj et al., 2014). Consistent with this observation, exons controlled by Zfp871 and Srrm4 are flanked by sequences enriched in UGC motifs, as is also observed for a larger set of Srrm4-dependent exons (Figures 3-17B and 3-17C). These results suggest that Zfp871 controls AS via both direct and indirect mechanisms that impact Srrm4 and its target exons, as well as additional neural exons that are not regulated by Srrm4.
Figure 3-17 Zfp871 binds RNA sequences proximal to its target exons.

(A) Average iCLIP signal of Flag-Zfp871 in N2A cells around exons with increasing, unchanged, or decreasing inclusion upon knockdown of Zfp871, as analyzed by RNA-Seq. Inserts show median and inter-quartile range of mean intronic CLIP signals. Asterisks indicate P-values of one-sided Mann-Whitney U-tests. See also Figure A2-3. (B) Mean coverage of the Srrm4 core motif, UGC, around groups of exons with decreased PSI upon siRNA knockdown of Zfp871, Srrm4, or both. (C) Exon properties previously associated with neural regulation by Srrm4, plotted for groups of exons with indicated behaviour upon knockdown of Zfp871 or Srrm4. PPT, polypyrimidine tract.
3.4.6 Nacc1 regulates ESC-differential AS at multiple levels

The results from analyzing Zfp871 raised the intriguing possibility that additional annotated transcription factors/DNA binding proteins with pronounced effects on AS in the screen might also have dual direct and indirect regulatory activities in AS. To investigate this, I next focused on the Pox virus and Zinc finger/Bric-a-brac Tramtrack Broad complex (POZ/BTB) family transcription factor, Nucleus accumbens associated 1 (Nacc1). From the screen data, knockdown of Nacc1 has strongly correlated effects with knockdowns of Mbnl and Rbfox2 (Figure 3-18), which, as mentioned earlier, negatively regulate ESC-differential AS events and control somatic cell reprogramming (Han et al., 2013; Venables et al., 2013). Moreover, knockdown followed by RNA-Seq analysis revealed that Nacc1 has more widespread effects on ESC-differential AS, which also correlate significantly with knockdown of Mbnl proteins (ρ=0.76, rank correlation; p < 2.2 × 10^{-16}; Figure 3-19A). The RNA-Seq data further revealed that Nacc1 knockdown reduces the expression of Mbnl1 and, to a lesser extent, other splicing regulators including Rbfox2 (Figure 3-19B). Moreover, ChIP-Seq analysis of Nacc1 occupancy in N2A cells shows that it binds proximal to the transcription start site of the Mbnl1, but not of the Mbnl2 or Rbfox2 genes, whose expression is not strongly affected by Nacc1 knockdown (Figure 3-20A and data not shown). Similarly, Nacc1 peaks are generally enriched near transcription start sites of genes with changing expression upon Nacc1 knockdown, supporting a transcription regulatory role (Figure 3-20B). Collectively, these data provide evidence that Nacc1 indirectly regulates ESC-differential AS by controlling the expression of Mbnl1 and possibly other splicing regulators.
Figure 3-18 The effect of Nacc1 on ESC-differential AS regulatory networks.

(A) Effects of N2A screen hit knockdowns (|SSMD| > 4.75) on Mbnl1 mRNA expression, and correlation of AS changes with those observed upon knockdown of Mbnl1, sorted according to changes in Mbnl1 expression. (B) Effects of N2A screen hit knockdowns (|SSMD| > 4.75) on Rbfox2 mRNA expression, and correlation of AS changes with those observed upon knockdown of Rbfox2, sorted according to changes in Rbfox2 expression.

Figure 3-19 RNA-Seq analysis reveals extensive effect of Nacc1 on ESC-differential AS.

(A) Splicing changes of alternative exons observed upon knockdown of Mbnl or Nacc1. Annotated ESC-differential events are highlighted. (B) mRNA expression changes upon knockdown of Nacc1 or Mbnl in N2A cells analyzed by RNA-Seq. Dark grey represents genes that are targeted by knockdown.
Figure 3-20 Nacc1 binds proximal to the transcription start site of Mbnl1 and other target genes.

(A) ChIP-Seq profiles of Nacc1 in N2A cells at Mbnl1 and Mbnl2 genes. FPM, ChIP fragments per million reads. (B) Numbers of genes with indicated magnitudes of mRNA expression change upon siRNA knockdown of Nacc1, as analyzed by RNA-Seq (left panel) and percentage of genes in the same categories with a promoter-proximal Nacc1 ChIP-Seq peak. Absolute numbers of genes are indicated.

Surprisingly, Nacc1 also directly regulates AS. Similar to the results for Zfp871, using iCLIP analysis in N2A cells, we observed that Nacc1 frequently crosslinks to exonic RNA, but it also displays enriched occupancy over intronic sequences adjacent to its regulated target exons, particularly those exons with decreased inclusion upon Nacc1 knockdown (Figures 3-21A and
2A-4) ($p = 1.8 \times 10^{-8}$, one-sided Mann-Whitney U-test). This observation suggests that Nacc1 may have a direct role in promoting the inclusion of ESC-low exons, whereas its role in repressing ESC-high exons may be more indirect.

To confirm whether Nacc1 has a direct role in regulating AS, we assayed bacterially-expressed recombinant Nacc1 for activity in promoting the splicing of target exons in reporter transcripts in vitro. Remarkably, increasing concentrations of Nacc1 protein stimulated the inclusion of an alternative exon from the Myo9b gene, which, based on the analysis of knockdown-RNA-Seq and iCLIP data, is regulated by Nacc1 (Figures 3-21B, 2A-5A, and 2A-5B). In contrast, addition of comparable levels of recombinant PTBP1, or of BSA, did not significantly affect splicing levels of the reporter transcript. Moreover, recombinant Nacc1 did not promote splicing in vitro of a neural-specific exon (Figure 2A-5C). These results thus provide evidence that Nacc1, like Zfp871, has dual indirect and direct roles in the regulation of AS events linked to cell fate.

![Figure 3-21](image)

**Figure 3-21 Nacc1 regulates AS through direct binding to RNA.**

(A) Average Nacc1 iCLIP signal in N2A cells around exons with increased, unchanged, or decreased inclusion upon knockdown of Nacc1, as analyzed by RNA-Seq. Inserts as in Figure 3-17A. See also Figure A2-4. (B) In vitro splicing of Myo9b minigene reporter transcripts in HeLa whole cell extracts, with or without the addition of recombinant Nacc1, or PTBP1 as a negative control. PSI quantitation is shown below. See also Figure A2-5.


3.5 Discussion

In this study, I describe SPAR-Seq, a system that affords the rapid and in-depth discovery of networks of trans-acting factors that control RNA regulatory events. A particular advantage of this system is that it simultaneously generates a highly-quantitative, sequencing-based read-out for dozens of endogenous regulatory events in response to thousands of query conditions. By applying SPAR-Seq in combination with knockdown conditions to discover regulatory networks that control cell fate-associated AS events in mouse ESCs and neural cells, I have illuminated extensive positive and negative functional inter-relationships between new and known trans-acting AS factors associated with different gene regulatory layers. A particularly striking observation in neural cells is that annotated transcription, chromatin, and DNA binding domain proteins impact AS at a similar frequency as annotated splicing factors, and that a subset of these protein factors dually control cell fate-associated AS networks through direct and indirect mechanisms.

Previous studies have described various mechanisms by which transcription and chromatin regulators can impact AS. These include the recruitment of splicing components to transcription and chromatin complexes that subsequently influence AS in nascent transcripts, as well as transcription and chromatin-associated effects on RNA Pol II elongation rate that affect AS by altering the kinetics of exposure of cis-competing splice sites (Braunschweig et al., 2013; David and Manley, 2011; Luco et al., 2011; Naftelberg et al., 2015; Shukla and Oberdoerffer, 2012). Our results extend these previous studies by providing evidence that certain transcription factors can regulate AS regulatory networks by binding RNA sequences proximal to regulated target exons, while also controlling the expression of splicing regulators that control the same target exons. These observations further add to a growing body of evidence indicating that proteins lacking canonical RNA binding domains, including those linked to transcription and chromatin regulation, can interact with RNA (Baltz et al., 2012; Castello et al., 2012; Hendrickson et al., 2016; Kwon et al., 2013).

Importantly, we observed that knockdown of more than 50% of analyzed ZnF genes resulted in pronounced effects on different subsets of AS events. Specific classes of ZnF proteins, such as those possessing CCCH-type motifs, include members with relatively well established roles in RNA binding and regulation (e.g. U2AF35 and MBNL proteins) (Konieczny et al., 2014;
Merendino et al., 1999; Wu et al., 1999; Zorio and Blumenthal, 1999). In contrast, while C2H2 ZnF proteins constitute the largest class of annotated DNA binding proteins, only a small number of these proteins have been investigated for roles in RNA binding and regulation. Notable examples include Gtf3a/TFIIIA, which controls rRNA biogenesis by activating transcription of the 5S rRNA gene and binding directly to a structured region within 5S rRNA (Layat et al., 2013), HZF/Znf385a, which controls the localization of mRNAs to synapses (Iijima et al., 2005), CTCF, a transcription and chromosome organizing factor that regulates the p53 gene via direct binding to an RNA transcript antisense to the p53 locus (Saldana-Meyer et al., 2014), and YY1, which activates gene expression by binding both to promoter/enhancer elements and RNA transcribed from the corresponding loci (Sigova et al., 2015). Our results surprisingly revealed that more than 40% of the 112 C2H2 ZnF proteins analyzed in the present study impact cell fate-associated AS events, thus suggesting a widespread role for this class of ZnF protein in splicing control.

Focused experiments on the C2H2 ZnF protein Zfp871 suggest that it regulates neural AS by stimulating the expression of the neural splicing regulator Srrm4, but also by binding RNA sequences proximal to Srrm4-dependent and other neural-regulated exons. Based on these observations, it is intriguing to consider that many additional splicing regulators exist among the hundreds of C2H2 ZnF proteins that have known or predicted DNA binding activity (Najafabadi et al., 2015), but that have not yet been investigated for roles in AS, or other aspects of RNA regulation. Consistent with this view, an analysis of 50 C2H2 ZnF proteins, including additional examples that were identified as regulators of AS in this study, revealed that 45 cross-link to RNA in cells (N. Haider Shah and J.F. Greenblatt unpublished observations). It is also interesting to consider that the expansion of this class of proteins during vertebrate evolution, which occurred in part to fulfill roles in transcriptional regulation and the suppression of rapidly evolving transposable elements (Emerson and Thomas, 2009; Stubbs et al., 2011; Tadepally et al., 2008), may also have arisen to regulate increasingly complex patterns of AS associated with the evolution of organs such as the mammalian brain (Barbosa-Morais et al., 2012; Merkin et al., 2012).

Finally, we show that the paradigm of annotated DNA binding and transcription factors multitasking at the level of RNA to perform roles in post-transcriptional regulation extends to other classes of proteins. The BTB/POZ domain protein Nacc1, in an analogous manner as
Zfp871, controls the expression of Mbnl1, a major regulator of ESC-differential AS (Han et al., 2013; Venables et al., 2013), but it also affects AS of target exons directly via RNA binding. The remarkable duality in regulatory capacities of the two factors investigated in more detail in the present study suggests that additional known or predicted transcription factors possess multitasking and coordinated regulatory functions that operate through binding DNA and RNA. The results of the SPAR-Seq screen thus highlight annotated transcription factors and DNA binding proteins as a cache of previously unknown regulators of AS networks, including those with important roles in the control of cell fate. Further exploration of the resource of new splicing regulators described in this study holds promise for the discovery of mechanisms and networks of AS regulation with critical roles in development and disease. Moreover, the flexible screening system I have described in this study opens the door to the comprehensive elucidation of RNA regulatory networks in diverse other mechanistic and biological contexts.
Chapter 4
4 Conclusions and Future Directions

4.1 Concluding remarks

The past decade has seen remarkable strides in delineating the transcriptomic landscapes of diverse cell and tissue types, developmental stages, and conditions from multiple species. In particular, the advent of high-throughput technologies, together with sophisticated computational tools, has uncovered vast repertories of AS events. These repertories represent a distinct regulatory layer that plays a major role in gene expression and biological complexity. Also emerging is that AS programs are coupled and communicated with other layers of gene regulation to significantly impact cell identity and function, both in the context of normal physiology and disease pathogenesis. A major goal for current and future studies is to systematically assign biological roles of functionally relevant splice variants, as well as to comprehensively determine the factors and mechanisms by which networks of AS events are coordinately regulated.

In this thesis, I conducted in-depth research to discover and characterize AS regulatory networks that play vital roles in cell fate control. Using integrated RNA-Seq, splicing code, and CLIP-Seq approaches, I identified the muscleblind-like RNA binding proteins, MBNL1 and MBNL2, as master negative regulators of a large program of ESC-differential AS events, which is required for pluripotency maintenance and efficient somatic cell reprogramming. This work provided among the first insights into the essential contribution of splicing regulatory circuits to fundamental properties and functions of ESCs, and led to a novel alternative method for improving the efficiency and quality of iPSC production.

Moreover, I also developed a new and versatile high-throughput screen for the systematic identification of perturbations that affect endogenous AS and gene expression events of interest. The screen in ESCs and neural cells unearthed hundreds of positive and negative factors from different gene regulatory processes that control specific subgroups of AS events associated with pluripotency and differentiation, thus shedding light on intricate and multifaceted mechanisms of AS regulation. Overall, these findings lay the groundwork for future research in the areas of splicing regulation, alternative exon function, stem cell biology, and reprogramming. The high-throughput strategies developed here in conjunction with advances in complementary methodologies are expected to ultimately generate a genome-wide, exon-resolution view of gene
function and regulation in development and disease, and to further guide the development of innovative approaches for diagnosis, prognosis, and therapy.

4.2 Future directions

4.2.1 Further characterization of biological functions of cell fate-associated AS networks and their key regulators

Results from the above studies have provided a wealth of new data on networks of cell fate-associated AS events, many of which await further functional analysis and characterization. Recent studies from our lab and others have shown that tissue-regulated alternative exons (longer than 27 nucleotides) and microexons (3-27 nucleotides in length) are highly enriched in intrinsically disordered regions and structured interaction domains, respectively, to modulate protein and other type of ligand interactions (Buljan et al., 2012; Ellis et al., 2012; Irimia et al., 2014; Li et al., 2015; Weatheritt et al., 2012). Together, these observations suggest a widespread role of regulated AS events in remodeling protein interaction networks so as to control cell fate and function. To facilitate the functional analysis of ESC-differential AS events, computational methods can be utilized to examine the sequence, structural, and functional features of these alternative exons, especially their overlap with disordered regions, modular domains, as well as additional critical motifs/sites. Subsequently, large-scale PPI assays, such as LUMIER (luminescence-based mammalian interactome mapping) (Barrios-Rodiles et al., 2005; Ellis et al., 2012), can be employed to systematically investigate the functional effect of ESC-differential AS events on PPI networks. Conserved, co-regulated alternative exons that are predicted to potentially participate in PPIs will be prioritized in this study. Notably, our initial data from a pilot LUMIER screen show that ESC-differential alternative exons regulated by MBNL proteins are capable of promoting or repressing partner interactions, including an auto- and cross-regulated alternative exon in MBNL2 (J. Ellis, H. Han, M. Barrios-Rodiles, J.L. Wrana and B.J. Blencowe, unpublished data). For top candidates predicted by computational analysis and/or identified by LUMIER, immunoaffinity purification coupled to mass spectrometry (AP-MS) can be applied for the comprehensive mapping of functions of specific ESC-differential AS events in regulating PPI networks that contribute to cell fate decisions.

Moving forward, it will be important to fully elucidate biological roles of cell fate-associated AS events in ESC pluripotency, lineage differentiation, and somatic cell reprogramming. Subsets of
AS events that have a profound impact on PPIs, as well as a few striking examples that alter other aspects of protein functions, such as DNA/RNA binding, enzymatic activities, or subcellular localization, can be selected for initial studies using cell-based assays and animal models. While custom siRNAs or shRNAs can be designed to specifically knock down each individual splice isoforms, cDNA constructs can be used to overexpress specific isoforms in corresponding cell lines. Alternatively, one can knock down or knock out endogenous genes and then, use inducible cDNA plasmids to express individual splice variants with or without the alternative exon at the levels comparable to endogenous proteins. These approaches have been successfully employed to study functional consequences of several AS events important for ESC biology, such as AS switches in FOXP1, Sall4, and Tcf3 (Gabut et al., 2011; Rao et al., 2010; Salomonis et al., 2010). Recently, the CRISPR/Cas9 system has been utilized to generate clonal mouse ESC lines that uniquely express Ptbp1 isoforms with or without the mammalian-specific exon 8, demonstrating distinct functional roles of these isoforms during in vitro differentiation of ESCs to cortical glutamatergic neurons (Gueroussov et al., 2015). With the rapid development of functional genomic technologies, it will open up the possibility for genome-scale pooled or arrayed screens to illuminate functions of the multitude of AS isoforms in the future.

Additionally, the current screen has discovered a large number of novel factors regulating subnetworks of ESC-differential and neural-enriched AS events. Their mechanisms of action and biological functions will be further determined in the context of stem cell and neural cell fate. Two of these AS regulators, Zfp871 and Nacc1, have been characterized, and mechanistic studies will also be carried out to fully delineate regulatory roles of other factors as well as interconnections between these factors. On the other hand, the biological effect of key splicing regulators will be investigated individually and in combination. Knockdown, knockout, and/or overexpression experiments can be performed in conjunction with in vitro and in vivo functional assays, such as those used to study MBNL and nSR100 proteins (Calarco et al., 2009; Han et al., 2013; Quesnel-Vallieres et al., 2015; Raj et al., 2011). Moreover, specific animal models can also be generated to further establish functional roles of master splicing regulators during development. Taken together, our increasingly expanding knowledge about global AS regulatory networks will likely offer new toolkits for modulating the cell fate of pluripotent stem cells and neural cells for basic research and clinical applications.
4.2.2 Systematic analysis of stage-specific AS regulatory networks during cell fate transitions

To gain further insights into the dynamics and complexity of AS regulatory networks, integrated high-throughput analysis can be performed across different stages during cell fate transitions. Recently, RNA-Seq profiling data have been generated at multiple time points during in vitro differentiation of mouse and human ESCs/iPSCs into neuronal cells, but landscapes of AS regulation underlying neurogenesis remain largely uncharted (Busskamp et al., 2014; Hubbard et al., 2013 for examples). Using this source of RNA-Seq data, one can initiate quantitative, global-scale surveys of stage-specific AS events, such as those associated with cell-state transitions between pluripotent stem cells, neural progenitor cells, postmitotic immature and mature neurons. In parallel, these data can also be used to systematically analyze dynamic changes of gene expression, especially detecting and classifying splicing regulator genes that are specifically expressed at different stages of neuronal differentiation. Similar analyses will be subsequently performed across multiple phases of in vivo neuronal development and the results will be compared to those from in vitro differentiation. Since the majority of previous transcriptome profiling studies were carried out at the population level, one concern is cellular heterogeneity. This problem has been partially circumvented by implementing more rapid and robust differentiation protocol to generate homogeneous neurons (Busskamp et al., 2014). Meanwhile, transgenic labeling coupled to fluorescence-activated cell sorting (FACS) as well as immunopanning using specific cell surface antibodies have been used to purify various neuronal subtypes from mouse cerebral cortex for RNA-Seq profiling (Zhang et al., 2014). In another study, FACS based on a unique combination of three nuclear markers has been applied for isolation and analysis of distinct subpopulations of neurons at multiple developmental stages (Molyneaux et al., 2015). Moreover, single-cell RNA-Seq together with a droplet-microfluidic barcoding strategy has been recently harnessed to analyze transcriptomes of individual stem and neural cells in a high-throughput manner (Klein et al., 2015; Llorens-Bobadilla et al., 2015; Macosko et al., 2015; Shin et al., 2015). Methods such as fluorescent in situ RNA sequencing (FISSEQ) have been developed for subcellular transcriptome profiling (Lee et al., 2014). By integrating tremendous data generated by the above approaches, it will allow researchers to systematically investigate AS regulatory networks during neuronal development with unprecedented spatial and temporal resolution. Likewise, transcriptome analysis will also be conducted in disease context in order to determine how misregulation of stage- and cell-specific
AS programs drives and contributes to neurodevelopmental and neurodegenerative disorders, such as autism spectrum disorder (ASD) and Alzheimer disease (AD).

In addition to RNA-Seq data, it will be important to include ribosome profiling and proteomic data in future analyses. For example, a translating ribosome affinity purification (TRAP) method has been developed to comprehensively probe translated mRNAs in genetically defined cell types in the mouse brain (Doyle et al., 2008; Heiman et al., 2008). It is noteworthy that increasing amounts of information about regulatory targets of important cell fate-associated splicing factors, including MBNL1/2, RBFOX1/2, NOVA1/2, PTBP1/2, and nSR00, has been generated using RNA-Seq, CLIP-Seq, and complementary approaches under both physiological and pathological conditions. Ultimately, using combinations of diverse forms of input datasets, integrative network analysis and splicing code modeling can be employed to infer networks of coordinated AS events that are associated with different regulatory time points of neuronal development and their relevance to various types of neurological disorders, as well as to predict cis- and trans-acting splicing regulators that potentially govern these AS networks. Collectively, the above studies will provide a framework for dissecting AS regulatory networks, which can also be applied to facilitate a systems-level understanding of other complex cell fate transitions, such as somatic cell reprogramming and cancer initiation, progression, and metastasis.

4.2.3 Broader applications of the novel high-throughput screening strategy developed in this study

4.2.3.1 High-throughput screening of trans-acting factors orchestrating stage-specific AS events linked to neurogenesis and neurological disorders

The new high-throughput screen developed in my thesis work, which employs an SPAR-Seq strategy for systematic identification of gene knockdowns or other types of perturbations that affect endogenous AS and gene expression events of interest, is a powerful system for future studies of gene regulation in normal and disease context. Besides the methodology, the current screen has generated valuable resources, including 6144 total RNA samples from siRNA knockdown and control treatments as well as forward and reverse barcode primer libraries, which are readily available for analyzing different sets of AS and gene expression events to uncover additional novel regulators in ESCs and neural cells. For example, using existing RNA samples from mouse N2A cells, high-throughput screens can be performed to systematically map
regulatory networks controlling stage- and cell-specific AS events that are linked to neuronal development and neurological disorders (see section 4.2.2). In particular, subnetworks of conserved neural AS events with distinct regulatory patterns during neurogenesis and/or misregulated in disease, especially those lacking known trans-acting regulators, can be prioritized in such screens. As mentioned earlier, a highly conserved program of neural microexons has recently been shown to play an important role in remodeling PPI networks during neurogenesis and is frequently disrupted in the brain of ASD individuals (Irimia et al., 2014). Since the neural-specific splicing factor nSR100 regulates approximately half of these microexons, it is expected that additional novel factors responsible for positively or negatively controlling neural-differential, disease-relevant microexons will be identified. Moreover, it will also be interesting to analyze neural AS events affecting cell surface proteins, which can serve as a basis for generating custom antibodies with utility in diagnosis and therapy. One of the main advantages of the SPAR-Seq strategy is its capacity to analyze groups (e.g. 50 or more) of AS and gene expression events simultaneously. Therefore, multiple previously known (e.g. nSR100, Ptbp1/2, and Nova1/2) and newly identified (e.g. Zfp871) neural-related splicing factors can be monitored in the future screen in order to define potential factors or pathways that regulate steady-state mRNA levels of these splicing factor genes.

Meanwhile, it is possible to further develop the screening system to multiplex larger number (e.g. ~100-200) of endogenous AS, gene expression, as well as other types of RNA regulation events, such as alternative polyadenylation and RNA editing. Because of the versatile and scalable nature of the system, one can easily expand the use of this screening platform to carry out genome-wide loss-of-function or gain-of-function screens in a wide variety of mouse and human cell types.

4.2.3.2 Chemical screens for small molecule modulators of AS and gene expression events relevant to development and disease

In addition to genetic screens mentioned above, high-throughput SPAR-Seq screens can be conducted using compound libraries to identify small molecules that modulate developmentally regulated, disease-relevant AS regulatory networks. Increasing evidence has shown that splicing defects underlie a broad spectrum of human diseases, such as neurological disorders, hematopoietic dysplasia, and various cancers (Daguenet et al., 2015; Singh and Cooper, 2012). As a result, small molecules targeting key AS events and their cognate splicing regulators
represent prime candidates for pharmacotherapeutic modulation. Chemical compound screens and functional genomic screens can be designed to probe the same or overlapping sets of endogenous AS and gene expression events, including but not limited to a functionally coherent program of neural-regulated microexons together with splicing factors modulating their AS that have been recently implicated in ASD. Using this highly multiplexed screening strategy, one can expect to identify small molecules that affect multiple aspects of AS regulatory networks, such as impacting RNA secondary structure and/or cis-regulatory elements through direct RNA binding and altering the expression, localization, and/or function of splicing regulators associated with different gene layers. Subsequently, by combining the chemical and genetic screen data, as well as by using mass spectrometry (Ong et al., 2009) and chemical cross-linking and isolation by pull-down (Chem-CLIP) (Guan and Disney, 2013; Yang et al., 2015) methods to detect proteins or RNA bound to small molecules of interest, it will greatly facilitate linking compounds to their targets.

To further assist and accelerate the process of potential drug discovery and characterization, complementary screening strategies can also be developed. The CRISRP/Cas9 editing system can be employed to generate multi-color fluorescent reporter cell lines to monitor several combinations of functionally significant AS events, expression events of key splicing factors and cell fate markers. For example, in neural cell lines, blue fluorescent protein (BFP) will be used to tag endogenous nSR100 to detect its expression, and a two-color splicing reporter construct will be inserted downstream of a nSR100-regulated, ASD-relevant microexon whereby the inclusion and exclusion of this microexon are associated with enhanced green fluorescent protein (EGFP) and mCherry expression, respectively. Using these stable cell lines, it is possible to screen small molecules that modulate microexon AS with or without affecting nSR100 expression, shedding light on different mechanisms. One limitation of this multi-labeling strategy is the number of AS and/or expression events monitored each time, but it will be very useful for rapid pre-screening of potentially interesting compounds, which can be further analyzed using the SPAR-Seq approach. Notably, the reporter cell lines can also be utilized for pooled genetic screens. In addition, fluorescent reporter cell lines monitoring multiple cell fate markers can be applied for functional phenotypic screens, such as pinpointing genetic (e.g. splicing regulators or AS events) and small molecule modulators that promote or inhibit cell fate transitions during neurogenesis.
Lastly, lead compounds selected can be tested in mutant primary neurons and mouse models of ASD or other related diseases for phenotypic rescue.

In summary, the development and application of all these high-throughput strategies will lead to multifaceted discoveries of functions and mechanisms of intricate AS regulatory networks in development and disease. Moreover, a powerful platform will be built for systematically identifying and characterizing small molecule compounds as well as other types of potential drugs (e.g. antisense oligonucleotides and antibodies) that impact AS regulatory networks through various mechanisms, and therefore holds promise for the discovery and development of future splicing-directed therapies.
References


proteome and its global occupancy profile on protein-coding transcripts. Molecular cell 46, 674-690.


retention in mammals functionally tunes transcriptomes. Genome research 24, 1774-1786.


interaction with its natural antisense transcript, Wrap53. Genes & development 28, 723-734.


Staknis, D., and Reed, R. (1994). SR proteins promote the first specific recognition of Pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex. Molecular and cellular biology 14, 7670-7682.


Appendices

Appendix 1 Supporting Data for Chapter 2

A

<table>
<thead>
<tr>
<th>siControl</th>
<th>1:1</th>
<th>1:2</th>
<th>1:4</th>
<th>1:8</th>
<th>siMBNL1-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>anti-MBNL1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-TUBULIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-MBNL2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-TUBULIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Human HeLa cells
(b) Human 293T cells
(c) Mouse N2A cells
(d) Human H9 cells
(e) Mouse CGR8 cells

B

FOXP1

MBNL1

MBNL2

ACTIN

C

FoxP1

Mbnl1

Mbnl2

Gapdh

D

PSI (FOXP1-ES)

E

FoxP1

PSI (FoxP1-ES)
Figure A1-1 siRNA knockdown of MBNL1 and MBNL2 promotes ESC-specific AS of FOXP1 transcripts.

(A) Western blot confirming efficient knockdown of MBNL1 and MBNL2 proteins in HeLa cells transfected with siRNA pools targeting these factors (siMBNL1+2, lane 6). Lane 5, lysate from cells transfected with a non-targeting siRNA pool (siControl). Lanes 1-4, serial dilutions (1:1, 1:2, 1:4 and 1:8) of lysate from cells transfected with siControl. (B) RT-PCR assays monitoring mRNA levels and/or AS patterns of FOXP1, MBNL1 and MBNL2 transcripts in 293T cells transfected with control siRNAs (siControl) or siRNA pools targeting MBNL1 (siMBNL1), MBNL2 (siMBNL2) or both of these factors (siMBNL1+2). Primers specific to the constitutive exons were used to amplify both FOXP1 isoforms. “PSI values” based on the ratios of expression of FOXP1 canonical (blue exon) and FOXP1-ES (red exon) isoforms analyzed by qRT-PCR using splice junction-specific primers are shown below. (C) mRNA levels and/or AS patterns of murine Foxp1, Mbnl1 and Mbnl2 were assayed as in (B) following transfections of control siRNAs (siControl) or siRNAs targeting Mbnl1 (siMbnl1), Mbnl2 (siMbnl2) or both of these factors (siMbnl1+2) in mouse N2A cells. (D and E) mRNA levels and/or AS patterns of FOXP1/Foxp1 were assayed as in (B and C) following transfections of siRNA pools in human H9 ESCs (D) mouse CGR8 ESCs (E) respectively.
Figure A1-2 Additional examples of ESC-differential AS events regulated by MBNL1 and MBNL2.

(A) Representative RT-PCR validation assays for ESC-differential AS events that have PSI changes in HeLa cells following siMBNL1+2 transfection (refer to Figure 2-1 legend); splicing patterns in human H9 ESCs are shown for comparison. (B) RT-PCR analysis of ESC-differential AS events following transfection of single siRNAs comprising the siRNA pool targeting MBNL1. Results for representative ESC-differential AS events are shown (compare also with results in Figure 2-6).
Figure A1-3 Characterization of secondary MEF reprogramming following siRNA knockdown of Mbnl proteins.

(A) qRT-PCR quantification of mRNA expression levels of endogenous Sall4 and Alpl genes during reprogramming. Secondary MEFs were transfected with control siRNAs (siControl) or siRNAs targeting Mbnl1 and Mbnl2 (siMbnl1+2) or Oct4 (siOct4) and treated with doxycycline (Dox) for 3 days (blue bars) or 5 days (red bars) before analysis. Secondary MEFs without Dox induction are shown as empty bars. Values represent Means ± Range (n=3). t-test was used to assess statistical significance of the differences between conditions. (B) Secondary MEFs were transfected with Mbnl1 and Mbnl2 siRNAs (siMbnl1+2) and treated with Dox for 8 days followed by 5 days of Dox withdraw. The resulting iPSC colonies were stained with alkaline phosphatase (AP) at day 13 and representative microscope images are shown (10X, scale bar = 400 µm). (C) Representative images of Dox-independent iPSC colonies immunostained with antibodies specific for Oct4, Nanog, Dpp4a and SSEA1 and counter stained with DAPI. Scale bar = 100 µm.
Figure A1-4 Teratoma analysis assessing the pluripotency potential of siMbnl1+2 iPSC lines derived from transgene-independent colonies during secondary MEF reprogramming.

Representative sections from teratomas generated from two independent transgene-independent siMbnl1+2 iPSC lines generated by secondary MEF reprogramming. Sections were stained with hematoxylin and eosin. Cell types representing all three germ layers – endoderm, mesoderm, and ectoderm – are indicated. Scale bar = 100 µm.
iPSCs derived from secondary MEF reprogramming following siRNA knockdown of Mbnl1+2 contribute to embryonic development. Chimeric embryos were generated by aggregating ICR (CD1 albino) morulae with GFP-positive/LacZ-positive siMbnl1+2 cell lines. Embryos at E10.5 were dissected, fixed and paraffin embedded. Whole-mount GFP detection (A), and LacZ staining (B) of chimeric embryos derived from one of the siMbnl1+2 iPSC lines. (C) Information on numbers of chimeric embryos generated from the two independent siMbnl1+2 iPSC lines. (D) Transverse sections of the whole-mount stained embryos show contribution of iPSC-derived LacZ-positive cells to derivatives of all three embryonic germ layers. hg, hindgut (endoderm); m, mesenchyme (mesoderm); n, neural tube (ectoderm).
Figure A1-6 Analyses of ESC-differential AS events in transgene-independent and transgene-dependent clones from secondary MEF reprogramming.

PSI values of ESC-differential alternative exons in fully reprogrammed iPSCs are significantly correlated with those in transgene-independent clones (a) ($r=0.8$, $p=3.2 \times 10^{-12}$), but not with PSI values of the same events in transgene-dependent clones (b) ($r=0.266$, $p=0.062$). PSI values for three transgene-independent and five transgene-dependent clones were averaged over each data point. Only ESC-differential exons with enough coverage in each of the samples and with a $\geq 25$ PSI difference between iPSCs and MEFs were included in the analysis. The 95% confidence and prediction intervals are represented by dashed and dotted lines, respectively.
Additional Supporting Data provided on the enclosed CD:

SD 1-1 Information on RNA-Seq datasets and samples.

Read length, length used in the bowtie alignments; # Reads, number of unique reads or read pairs in each sample; Source (SRA), identifiers for the NCBI Short Read Archive, where available; Reference/GEO, PubMed ID (where published available) or Gene Expression Omnibus (GEO) accession.

SD 1-2 Information on Human and mouse ESC-differential AS events.

EX LENGTH, exon length; FullCO, full coordinates of the AS event (chromosome, C1 donor, AS exon, C2 acceptor); av/min/max ESC/DIF/CL/iPS, average/minimum/maximum PSI in ESCs/differentiated tissues/cell lines/iPS cells; diffESC_DIF/CL, PSI difference between ESC and differentiated tissues/cell lines; Dif_KD_HeLa/293T/C2C12, PSI difference between siMBNL1+2 and siControl in HeLa/293T/C2C12 cells; C1/A/C2 sequence of C1, A and C2 exons of each event.

SD 1-3 DAVID (david.abcc.ncifcrf.gov/) output for functional enrichment categories for human, mouse or conserved ESC-differential AS events.

Three major functional clusters that were consistently enriched in genes harbouring the three groups of AS events are indicated: cytoskeleton (green), plasma membrane (yellow), associated to protein kinase (orange).

SD 1-4 Expression levels of the human and mouse splicing factors analyzed by RNA-Seq.

ESC-rank, rank of differential expression between ESC/iPSCs and differentiated tissues and cell lines (CL); Av_ESC+iPSCs, cRPKM average for the ESC and iPSC samples (columns H-N); Av_DIF, cRPKM average for the differentiated tissues and cell lines samples (columns M onwards); C_p-value Wilcox, Bonferroni-corrected p-value of Wilcoxon test of ESC/iPSCs versus differentiated tissues and cell lines after quantile normalization.

SD 1-5 Information on mouse ESC-differential AS events plotted in Figure A1-6.

AS events shown were represented by sufficient read coverage and ≥ 25 PSI differences between iPSCs and MEFs, as described above. TD/TI: refers to whether the clones died after Dox removal (transgene-dependent, n=5) or become stable iPSCs (transgene-independent, n=3); 21D refers to the day following Dox induction.
Appendix 2 Supporting Data for Chapter 3

A

CGR8

- Hit
- No hit
- siMbnl
- siNT
- Mock
- Untreated

ES cell-high
ES cell-low
Neural-enriched
Other
New

N2A

SSMD
-10 0 10
• Inclusion ▲

1,536 treatments (120 controls)

B

CGR8

Splicing/RBP
Both
Chromatin/TF
Signaling
siMbnl
Neg. controls

Fraction of hits
0.0 0.4 0.8
0 5 10 15 20

Fraction of hits
0.0 0.4 0.8
0 5 10 15 20

N2A

2.48
11.43

ISSMDI threshold for hit
Figure A2-1 Representation of full screen results.

(A) Heatmaps of SSMD scores for all treatments in CGR8 and N2A. Endogenous AS events monitored in the screen are grouped into ESC-high, ESC-low, neural-enriched, and others. ESC-high and ESC-low refer to exons that are preferentially included and skipped in ESCs relative to other cell and tissue types, respectively. Two novel AS events detected by de novo alignment are shown at the bottom. Multiple events within the same gene are denoted as A1 up to A3. Columns (treatments) are clustered in the same way for CGR8 and N2A, and rows (events) are clustered within each group. Greyscale heatmaps indicate PSI of each AS event in control CGR8 and N2A cells. (B) Accuracy of SPAR-Seq screen results in CGR8 and N2A. Shown are the fractions of treatments scored as hits at each value x, where a hit is a treatment that changes the SSMD of at least one AS event by more than x. The windows of 100% accuracy (all positive and negative controls are scored correctly) are indicated at the top.

Figure A2-2 Knockdowns affect AS in CGR8 and N2A cell.

Strongest effect on any AS event (absolute SSMD) for knockdowns of the indicated factor groups and negative control treatments. Red lines indicate SSMD thresholds used in Figure 3-11 (2.25 for CGR8 and 3.00 for N2A), which robustly call 100% of controls correctly. Dashed lines indicate medians, grey background the range of negative controls.
**Figure A2-3 iCLIP analysis of Flag-Zfp871 in N2A cell.**

Protein gel autoradiograph confirming the crosslinking of Zfp871-RNA $^{32}$P-labeled complexes after digestion with different concentrations of RNase I in N2A cells induced to express Flag-Zfp871. Arrow indicates the size of uncrosslinked Flag-Zfp871 protein. N2A cells expressing Flag-GFP were used as a specificity control.

**Figure A2-4 iCLIP analysis of Nacc1 in N2A cell.**

(A) Protein gel autoradiograph confirming the crosslinking of Nacc1-RNA $^{32}$P-labeled complexes after digestion with different concentrations of RNase I in N2A cells. Arrow indicates the size of uncrosslinked Nacc1 protein. IgG was used as a control. (B) Average Nacc1 iCLIP signal in N2A cells around exons with increased or decreased inclusion upon knockdown of Nacc1, as analyzed by RNA-Seq. Exons (n=134) with matched absolute PSI in control N2A cells were selected from each group with increased or decreased PSI shown in Figure 3-21A.
Figure A2-5 Nacc1 regulates *in vitro* splicing.

**(A)** Coomassie stained SDS-polyacrylamide gel of bacterially-expressed recombinant proteins and bovine serum albumin (BSA) used for *in vitro* splicing assays. 2 µg were loaded per lane.

**(B)** *In vitro* splicing of a Myo9b minigene reporter transcripts in HeLa whole cell extracts, with or without the addition of recombinant Nacc1, or BSA as a negative control. PSI quantitation is shown below.

**(C)** *In vitro* splicing of a Daam1 minigene reporter transcripts (as a specificity control) (Calarco et al., 2009) in Weri whole cell extracts, with or without the addition of recombinant Srrm4 or Nacc1. PSI quantitation is shown below.
Additional Supporting Data provided on the enclosed CD

SD 2-1 Information on endogenous AS and gene expression events monitored in the SPAR-Seq screen.

Event names, types, and coordinates (NCBI37/mm9) for alternative exons/regions are shown. The 52 AS events included in the downstream analyses are color coded correspondingly (see Figure 3-5). All forward and reverse primer sequences used for multiplex RT-PCR are provided.

SD 2-2 Information on siRNA knockdown and control treatments in the screen.

Treatment IDs/types and target gene names/IDs are given. Additional columns are: Group, 4 major functional categories of genes included in siRNA knockdown screens (see Figure 3-1C), splicing factors and RNA binding proteins (Splicing factor/RBP), chromatin and transcription factors (Chromatin/TF), factors overlapping both groups (Overlap), as well as signaling and post-translational factors (Signaling); Prediction.svm, genes predicted to be associated (chrom) or not associated (nochrom) with ‘chromosome’ using a machine learning model; mRNA.interactome, information on whether proteins were shown previously to interact with polyadenylated RNAs (Baltz et al., 2012; Castello et al., 2012; Kwon et al., 2013); RNA.binding.domain, information on whether an RNA binding domain is present; Splicosome, information on whether proteins are annotated in the Splicesome Database; C2H2.ZnF, information on whether a C2H2 ZnF domain is present; PFAM, PFAM protein domain information.

SD 2-3 Overall results for AS analysis in CGR8 and N2A cells.

Treatment IDs, target gene names, and treatment types are given. Additional columns are SSMD scores for each AS event monitored in the CGR8 and N2A screens.

SD 2-4 Overall results for gene expression analysis in CGR8 and N2A cells.

Gene names of endogenous events monitored in the screen are given. log2.RPM, log2 of reads per million reads (RPMs) of negative controls (siNT and mock controls); untreated.log2FC and untreated.FDR, log2 of fold change (FC) and false discovery rate (FDR) of untreated controls; siMbnl.log2FC and siMbnl.FDR, log2 of FC and FDR of positive controls (siMbnl). Additional columns are: log2 of FC and FDR of each siRNA knockdown treatment included in the screen.

SD 2-5 Information on functional clusters identified by CGR8 and N2A screens.

Gene names, IDs, cluster numbers, and top enriched terms from GO, CORUM, REACTOME, or KEGG are shown.

SD 2-6 List of CORUM complexes analyzed using CGR8 and N2A screen data.

List of 316 CORUM complexes identified in mouses or human cells of which at least three subunits (SU), or their orthologs, were tested in the screen. CORUM IDs and complex names, numbers of subunits in the complex, as well as numbers, gene names, and treatment IDs of subunits that were screened, are listed. Mean pairwise SSMD correlations among the subunits in CGR8 and N2A cells, p-values of two-sided Mann-Whitney U-tests against the pairwise correlations of all gene knockdowns in the screen, and Benjamini-Hochberg corrected false discovery rates (FDR) are also provided.
Appendix 3 Published Manuscripts