Multiple regulatory layers in the establishment of Rbfox2 splicing networks

by

Mohini Jangi

B.S. Biology, Brandeis University (2007)

Submitted to the Department of Biology
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2014

© 2014 Massachusetts Institute of Technology
All rights reserved

Signature of Author: ____________________________
Mohini Jangi
Department of Biology
January 15, 2014

Certified by: __________________________________
Phillip A. Sharp
Institute Professor of Biology
Thesis supervisor

Accepted by: __________________________________
Stephen P. Bell
Professor of Biology
Chair, Biology Graduate Committee
Multiple regulatory layers in the establishment of Rbfox2 splicing networks

by

Mohini Jangi

Submitted to the Department of Biology on January 15, 2014 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Biology

Abstract

Regulated alternative splicing is mediated by RNA binding proteins (RBPs) recognizing short sequence motifs in nascent transcripts. The Rbfox RBPs are highly conserved splicing factors that regulate tissue-specific exon inclusion by binding the RNA motif UGCAUG. We sought to comprehensively define the Rbfox2 splicing regulatory network and discover determinants of Rbfox2 regulation in mouse embryonic stem cells. We uncovered fundamental principles in the mechanistic aspects of Rbfox-dependent splicing and in the systems-level regulation of interconnected splicing networks.

Using high-resolution iCLIP and RNAseq, we identified many Rbfox2-mediated protein-coding splicing events and nearly 300 additional events, in particular those within RBPs, that are coupled to nonsense-mediated mRNA decay (NMD). Regulation of NMD-coupled splicing by Rbfox2 alters gene expression of autoregulated RBPs and hundreds of additional genes. These observations place Rbfox2 upstream of a large network of direct and indirect splicing changes and offer an explanation as to how autoregulated gene expression can be modulated.

We describe a validation of RNA Bind-n-Seq, a novel in vitro technique for analyzing RNA-protein interactions. We found a secondary Rbfox2 motif, GCACG, to be functional in splicing regulation in addition to the consensus UGCAUG and observed a preference for Rbfox2 binding to unstructured sequences. These findings provide a foundation for establishing the critical determinants of functional cis elements in splicing regulation.

We also investigated mechanisms of co-transcriptional splicing regulation by Rbfox2. Using chromatin immunoprecipitation, we found that Rbfox2 is recruited early in the transcription cycle to active promoters and transcriptional enhancers, likely via interaction and co-transcriptional tracking with RNA polymerase II. Modulation of chromatin structure alters Rbfox2-dependent splicing activity, supporting the emerging model that the chromatin environment influences exon choice.

Our analyses of Rbfox2 activity in mouse embryonic stem cells reveal hundreds of previously unknown splicing targets involved in diverse biological functions. In particular, we introduce a novel concept in splicing regulation whereby changes in the expression of one splicing factor induce a cascade of secondary and perhaps tertiary splicing changes through cross-regulation of RBPs. This model positions RBPs at a critical node in the establishment, reinforcement, and alteration of tissue transcriptomes during mammalian development.

Thesis supervisor:
Phillip A. Sharp, Institute Professor of Biology
Acknowledgements

There is much that is written between the lines of this thesis. The words of support, encouragement, and positivity from countless individuals have brought these years to fruition. I must first thank my Ph.D. advisor Phil Sharp, who has not just been a scientific advisor, but has at once encouraged me to think independently and confidently while providing support and insight at exactly the right times. In large part, what has made the past years such a formative and positive experience has been the unique culture of openness and respect that Phil has cultivated in the Sharp lab. The rigorous and collaborative approach to science I have learned here has been, and will continue to be, invaluable.

To my committee members, Chris Burge and Jackie Lees, I have truly appreciated the thoughtful advice, the honest criticism, and the encouragement to take ownership of my work over the years. In particular, I must thank Chris for welcoming me into the Burge lab and allowing me to expand my “splicing network” to include a really great group of people.

To my classmates, we have come a long way since the days of problem sets, ping pong, and pool(s) in the pit, but I feel lucky to be a part of one of those years that has stuck together until the end. I am sure we will all continue to intersect paths in the future, and everyone’s accomplishments, large and small, will always be a source of inspiration.

I joined the Sharp lab at a somewhat unusual time, when there was a good deal of turnover. While many of the past Sharpies with whom I only overlapped for a short time still left their indelible mark (in particular Joel, for teaching me to formulate a question before acting on it and for keeping things predictably unpredictable), the group of people who joined with me has been the most amazing group to learn from, commiserate with over our difficulties, and celebrate our successes. I must thank Paul for patiently teaching me everything I know about splicing, for an enjoyable collaboration, and for a healthy dose of good-natured realism with a little help from Allan. To the former residents of 528 and now the break room, Albert, Andrew, and Sara, thank you for keeping me grounded and always entertained. In particular, I would not have persevered through the rough patches without all of the encouragement and motivation from Albert. And to all of the Sharpies, the sum can only be as good as its parts, and I am incredibly fortunate to have been able to share this experience with each of them.

Most importantly, I owe everything over these past years to my family. I thank my brother for always reminding me that I am at the cutting edge of discovery, and that even if things do not work as expected, there is something valuable to be learned. To my parents, while some aspects of their support can be measured through countless surprise visits, loaves of banana bread, and tupperwares packed with tasty morsels, this hardly begins to scratch the surface. Their unwavering confidence, reassurance, and comfort have always been a source of stability and cannot be overstated. This thesis and everything it represents is dedicated to them.

- Mohini Jangi
Co-transcriptional splicing regulation by Rbfox2 ...........................................163

4.1 Abstract ........................................................................................................164
4.2 Introduction ...................................................................................................165
4.3 Results ..........................................................................................................167
4.4 Discussion .....................................................................................................175
4.5 Materials and Methods ..............................................................................179
4.6 References ....................................................................................................181
4.7 Figures ..........................................................................................................185

Conclusions and Future Perspectives ..................................................................201

5.1 Summary .......................................................................................................201
5.2 Mechanistic insights from global splicing analyses ....................................204
5.3 Physiological impact of splicing-coupled NMD ...........................................206
5.4 Chromatin as a marker and effector of regulated splicing ..........................208
5.5 RNA binding proteins as master regulators of cell state .............................209
5.6 References ....................................................................................................212

Biographical Note ..............................................................................................215
Chapter 1

Introduction

The regulation of gene expression is a fundamental determinant in establishing a cell’s identity and orchestrating responses to its environment. From transcription of the pre-messenger RNA, to processing that results in the mature mRNA, to translation into a protein product, many points of regulation allow for the precise tuning of gene expression networks. This thesis will focus largely on the process of pre-mRNA splicing, and in particular, the mechanisms and consequences of splicing regulation mediated by the RNA binding protein Rbfox2.

1.1 Combinatorial Control in Splicing Regulation

Cis elements are encoded in the genome to direct splicing

Messenger RNAs (mRNAs) in eukaryotes are assembled from longer precursors encoded by transcriptional units through the process of pre-messenger RNA splicing. The initial description of the removal of intervening sequences during the maturation of the adenovirus 2 hexon mRNA (Berget et al. 1977; Chow et al. 1977) led to speculation that RNA splicing is a general mechanism of mRNA biogenesis and gene regulation. Indeed, this phenomenon was soon reported in mammalian genes, where it was discovered that the rabbit β-globin gene
similarly contains an insert not present in the final mRNA (Jeffreys and Flavell 1977). Since then, our understanding of the complexity of this conserved process has increased exponentially.

The intervening sequences of those early discoveries, termed introns, contain specific sequence elements that signal their removal and the ligation of flanking exon sequences to produce the mature mRNA. These sequence elements are short and often degenerate. In metazoans, exon/intron junctions are minimally marked at the 5’ end of the intron by a GU dinucleotide (often in the context AG/GURAGU, R = purine) that constitutes the 5’ splice site (5’SS) (Mount 1982). The 3’ end of the intron terminates in an AG dinucleotide (YAG/G in metazoans, Y= pyrimidine) defining the 3’ splice site (3’SS) (Mount 1982). Two additional sequence elements near the 3’SS are also required for splicing – a branch site (Langford et al. 1984) and a polypyrimidine tract. The branch site, generally located up to ~100nt upstream of the 3’SS, bears the sequence YNYURAY and is separated from the 3’SS by the intervening polypyrimidine tract. The clear degeneracy of these sequences begs the question of how particular elements are selected to be active in splicing regulation, while others are ignored. This phenomenon highlights a key principle of splicing regulation, that of multi-step regulation and combinatorial control, which will be explored in more detail below.

It is thought that for all introns within pre-mRNAs in metazoans, the process of intron removal is carried out in two sequential trans-esterification reactions. During the first step, the pre-mRNA is cleaved at the 5’SS. This occurs via a nucleophilic attack of the guanosine in the 5’SS dinucleotide by the 2’OH of the branch site adenosine, resulting in release of the upstream exon and a non-canonical 2’-5’ phosphodiester linkage, or lariat, within the intron. The free 3’OH of the upstream exon then attacks the phosphodiester backbone at the 3’SS in the second step, bringing about the release of the lariat intron (Padgett et al. 1984; Ruskin et al. 1984),
which is rapidly debranched and degraded (Ruskin and Green 1985), and ligation of the exons in a canonical 3’-5’ phosphodiester linkage (Moore and Sharp 1993).

While the splice sites, branch point, and polypyrimidine tract are necessary for splicing, studies using artificial pre-mRNA substrates in vitro demonstrated that these elements were not always sufficient for completion of the splicing reaction (Nelson and Green 1988). Subsequent analyses found that the identity and length of exonic sequences influences the use of particular splice sites, providing an explanation for why the use of cryptic splice sites is rarely detectable in mature transcripts (Reed and Maniatis 1986). Identification of exonic splicing enhancers, or ESEs, selected for splicing activity in vitro (Tian and Kole 1995) and in vivo (Coulter et al. 1997) revealed a class of purine-rich motifs important for exon inclusion. More recently, computational dissection of gene structures has allowed for prediction and validation of additional ESEs (Fairbrother et al. 2002), as well as exonic splicing silencers, or ESSs (Wang et al. 2004; Zhang and Chasin 2004) that function to suppress nearby splice sites. Intronic splicing enhancers and silencers are also able to modulate splicing outcome and will be explored in more detail below. The combination of these and the required sequence elements provides a necessary layer of specificity in deciphering exon recognition during splicing.

The spliceosome is an RNA-protein machine

The cis elements described above are recognized by various components of the spliceosome, the multi-MegaDalton RNA-protein machine responsible for catalyzing intron removal (for reviews see Burge 1999; Wahl et al. 2009). Decades of in vitro biochemistry, supported by more recent high-throughput and in vivo analyses, have begun to elucidate the remarkable complexity and flexibility of the spliceosome. Five U-rich small nuclear RNAs
(snRNAs) U1, U2, U4, U5, and U6 (Lerner et al. 1980) assemble into their corresponding small nuclear ribonucleoprotein particles (snRNPs) with a host of >200 proteins that participate in the splicing cycle. Notably, the individual snRNPs do not catalyze splicing in isolation. Instead, the spliceosome is assembled in a stepwise manner on the nascent pre-mRNA, allowing the active site to be dynamically regulated with multiple checkpoints ensuring faithful exon ligation.

The earliest defined complex in spliceosome assembly, the E complex, forms upon recognition of the minimal cis elements required for splicing. Specifically, the U1 snRNP binds the 5’SS through base pairing interactions, while the large and small subunits of the U2 auxiliary factor (U2AF65 and U2AF35) recognize the polypyrimidine tract and the 3’SS respectively and the branch site is bound by SF1 in a series of cooperative protein-RNA interactions (Chabot et al. 1985; Jamison and Garcia-Blanco 1992). This is also referred to as the commitment complex and is the crucial first step in defining splice sites.

Next, the U2 snRNP is recruited to the branch point, and in an ATP-dependent step, base pairs with the branch point and displaces SF1 to generate the A complex. As is seen repeatedly throughout the splicing cycle, RNA-RNA interactions are coordinated and stabilized by protein partners. In this case, the arginine/serine repeats of U2AF65 help to recruit the U2 snRNP (Zamore et al. 1992), whereupon the U2 snRNP components SF3a and SF3b stabilize base pairing of the U2 snRNA (Gozani et al. 1996).

The association of U4, U5, and U6 with the A complex triggers formation of the B complex. In vitro depletion and complementation studies in both yeast and mammals indicate that these three snRNAs are recruited as the pre-assembled tri-snRNP U4/U6.U5 (Lamm et al. 1991; Seraphin et al. 1991; Utans et al. 1992). The B complex, containing all five snRNPs, is at this point catalytically inactive and must undergo major structural rearrangements to become
competent for splicing. In an ATP-dependent step mediated by the protein Prp28, the 5’SS transfers from the U1 snRNA to the U6 snRNA, dissociating U1 from the pre-mRNA and initiating spliceosome activation. Additional ATP- and GTP-dependent helicase activity mediates the unwinding of U4 and U6, resulting in the dissociation of U4 and many of the protein components of the tri-snRNP to form the B* complex (Bartels et al. 2002). In a series of conformational changes that are not precisely understood in vivo, U6 subsequently initiates pairing with U2 to liberate the catalytic residues of the U6 snRNA. These and other structural rearrangements occur during the first transesterification reaction and define the catalytically active C complex (Makarov et al. 2002). The U6 snRNA within the C complex was recently shown to catalyze both transesterifications, resolving decades of uncertainty over whether RNA or protein mediates splicing catalysis (Fica et al. 2013). Completion of the second step of the splicing reaction, concurrent with the action of the ATP-dependent helicases Prp22 and Prp16 (Schwer and Guthrie 1992; Schwer 2008), generates the posts spliceosome. The spliced product is released, and the U2, U5, and U6 snRNPs are disassembled to be recycled in additional rounds of splicing. The steps of spliceosome assembly have more recently been shown using a novel fluorescence microscopy-based assay in yeast to occur sequentially and reversibly, with each additional step contributing to commitment to splicing (Crawford et al. 2008; Hoskins et al. 2011).

Increasing evidence has supported the cooperative assembly of the U1 and U2 snRNPs on the pre-mRNA (Lamond et al. 1987; Grabowski et al. 1991; Michaud and Reed 1993). As many of the studies that initially characterized spliceosome assembly in yeast and mammals relied on short single-intron substrates, an intron-centric model was initially proposed to conceptualize cooperative assembly between closely-spaced 5’ and 3’ splice sites. For mRNAs
in higher eukaryotes, which frequently have introns that are multiple kilobases in length, the spatial conundrum posed by long-range interactions between the 5’SS and 3’SS was elegantly resolved by the exon definition model (Robberson et al. 1990; Hoffman and Grabowski 1992). Here, interactions between U2AF65 at an exon’s 3’SS and the U1 snRNP at its 5’SS initiate the splicing cycle and are rearranged from cross-exon to cross-intron interactions with the flanking splice sites prior to catalysis. Importantly, the exonic splicing enhancers outlined above are bound by the SR proteins SF2/ASF and SC35 (now termed SRSF1 and SRSF2) and play a key role in recruiting U1 and U2AF65 to the exon (Fu and Maniatis 1992; Eperon et al. 1993; Wu and Maniatis 1993; Kohtz et al. 1994; Zuo and Manley 1994). The relatively small size of vertebrate exons (on average 137 nucleotides) facilitates splice site pairing across exons (Berget 1995). The transition from exon-defined to intron-defined spliceosomes is still poorly understood, although the inability of a cross-exon E complex to fully proceed to a cross-intron A complex in the presence of the splicing repressor PTB implies that this transition is important for the completion of splicing in vivo (Sharma et al. 2008).

*Variations in cis-trans interactions lead to alternative splicing and genomic diversity*

The mechanisms outlined thus far in intron excision and exon ligation are thought to be applicable to the vast majority of splicing events. However, not all splicing events occur in all cellular conditions. The critical stage of exon definition appears to be the main distinguishing feature between constitutive splicing, in which a particular exon is always included in the mature mRNA, and alternative splicing, in which the exon is regulated to be included only in a subset of mRNAs. Owing to the degenerate nature of splice sites and the multi-step process of intron removal, regulation by trans-acting factors at the initial as well as later stages in the splicing
cycle is widespread in vertebrates. In fact, global analyses in the past decade suggest that greater than 95% of mammalian genes produce multiple isoforms through alternative splicing (Pan et al. 2008; Wang et al. 2008). This vast degree of transcriptomic diversity conferred by alternative splicing has been suggested to be essential in vertebrate evolution for increasing organismal complexity while maintaining a relatively compact genome.

Early discoveries of alternative splicing in the immunoglobulin genes (Early et al. 1980), murine αA-crystallin (King and Piatigorsky 1983), and fibronectin (Schwarzauer et al. 1983) paved the way for mechanistic investigations into splice site choice. Much of this thesis will focus on the regulation of cassette exons, or alternative exons that are included in some transcripts and skipped in others. Generally characterized by weak splice sites, cassette exons frequently require the action of RNA binding proteins not part of the core splicing machinery (alternative splicing factors) to aid, or in some cases inhibit, spliceosome assembly.

A well-studied example highlighting the various facets of regulated splicing is the tyrosine kinase c-Src exon N1. This exon encodes a 6-amino acid insert that is present in neuronal cells but is excluded in non-neuronal cells (Levy et al. 1987; Martinez et al. 1987). A number of studies have mapped sequence elements, including splicing enhancers and silencers, around and within N1 that are able to mediate exon inclusion and repression. Within the N1 exon, an ESE activates splicing by recruiting the SR protein SF2/ASF (Rooke et al. 2003). A highly-conserved ~100nt segment downstream of exon N1 was also found to be responsible for mediating exon inclusion in vitro and in neuronal cells (Black 1992). Furthermore, exogenously supplying this activating sequence was able to competitively inhibit exon N1 inclusion in vitro, arguing for the sequestering of regulatory factors by the activating sequence (Black 1992). The identity of these factors was later identified as two members of the heterogeneous nuclear
Chapter 1. Introduction

ribonucleoprotein (hnRNP family) of splicing regulators, hnRNP H and hnRNP F, which bind to G-rich elements within the conserved downstream control region (Min et al. 1995; Chou et al. 1999; Markovtsov et al. 2000). In addition, the KH-domain protein KSRP is also able to enhance N1 inclusion through a number of sequence elements downstream of N1 (Min et al. 1997), while the Rbfox proteins activate N1 through a highly specific regulatory element, UGCAUG (Nakahata and Kawamoto 2005; Underwood et al. 2005). The activity of the Rbfox proteins will be discussed in more detail below.

In non-neuronal cells, exon N1 is actively repressed by several splicing factors. Intronic sequences in the 3’SS region upstream of N1 as well as motifs downstream of the exon are responsible for conferring this repression (Chan and Black 1995). In vitro analyses demonstrated that the hnRNP protein PTB (Polypyrimidine Tract Binding protein) assembles cooperatively on CU-rich motifs across the splice sites flanking exon N1, sterically hindering interactions with the spliceosome and preventing exon definition (Chou et al. 2000). Additional repressive elements were found within the exon and were attributed to binding by hnRNP A1 in a mechanism distinct from PTB-dependent exon repression (Rooke et al. 2003).

The c-Src exon N1 exemplifies the recurring theme in alternative splicing regulation of combinatorial control. As tissue-specific expression of any given splicing regulator acting on N1 cannot fully explain the neuronal inclusion pattern, it is more likely that the precise titration of the involved splicing regulators achieves a 20-fold preference toward the inclusion of N1 in neurons (Modafferi and Black 1999). The use of multicomponent complexes to regulate individual splicing events has been suggested to increase the dynamic range as well as the robustness of the splicing reaction in a particular cellular context (Hertel et al. 1997; Smith and Valcarcel 2000). More recent systems-level and computational approaches have predicted the
cross-regulation of splicing events by multiple splicing regulators to achieve highly tissue-specific and condition-specific patterns of exon inclusion (Barash et al. 2010; Zhang et al. 2010).

1.2 Effects of co-transcriptionality of splicing regulation

Many of the mechanistic details of the splicing reaction have been characterized through extensive in vitro biochemistry. However, several observations made in the early years after the discovery of splicing hinted that splicing in the in vivo context introduced many additional variables. Further biochemical and global analyses have since solidified the idea that the majority of splicing occurs co-transcriptionally, while the nascent pre-mRNA is being actively transcribed by RNA Pol II.

Most splicing occurs co-transcriptionally

Before the direct demonstration that the splicing machinery assembled on nascent transcripts, discrepancies between the spliced products of in vitro and in vivo reactions spurred the hypothesis that co-transcriptional splicing would allow the sequential production of introns to influence splice site recognition (Aebi et al. 1986). In this “first-come, first-served” model, it was proposed that the co-transcriptional commitment of weaker upstream splice sites was kinetically favorable over the recognition of stronger downstream splice sites due to their later transcription. Examination of the transcription of chorion genes in the Drosophila early embryo by electron microscopy provided physical evidence that ribonucleoprotein (RNP) complexes assembled close to the splice sites of actively transcribing nascent messages (Beyer et al. 1981; Osheim et al. 1985). For many introns, transcription of the 3’ splice site initiated RNP assembly
and was followed by rapid intron removal within minutes, prior to the removal of downstream introns and polyadenylation (Beyer and Osheim 1988).

Later studies specifically investigated whether introns are removed co-transcriptionally in vivo. Introns from the Balbiani ring 1 and 3 genes in Chironomus tentans salivary glands were shown to be removed while transcripts were still tethered to the chromatin, and this occurred in a general, although not strict, 5’ to 3’ direction (Bauren and Wieslander 1994; Wetterberg et al. 1996). In mammals, dissecting the kinetics of transcription and splicing of the largest known gene in the human genome, dystrophin, similarly demonstrated that earlier introns are completely spliced prior to the transcription and removal of downstream introns (Tennyson et al. 1995). By comparing chromatin-associated and nucleoplasmic RNA for splicing products of the c-Src and Fibronectin 1 genes, Pandya-Jones and Black concluded that as high as 85-100% of spliced intermediates for particular introns are associated with the chromatin fraction. Notably, this was true for one or both of the introns around alternative exons, suggesting that commitment to exon inclusion may occur through rapid co-transcriptional removal of one intron, followed by the slower removal of the second intron (Pandya-Jones and Black 2009). Genome-wide analysis of chromatin-associated and nucleoplasmic RNA from activated macrophages revealed that a fraction of introns show delayed removal compared to their flanking introns, and that transcript release from the chromatin does not occur until all introns are removed (Bhatt et al. 2012).

While such evidence argues for the pervasiveness of co-transcriptional splicing in many systems in vivo, it does not necessarily imply that transcription and splicing are coupled. Multiple observations suggest that the C-terminal domain of the large subunit of RNA polymerase II (Pol II CTD) plays a key role in the coupling of splicing and transcription. In human cells in which endogenous Pol II was inhibited by α-amanitin, expression of an α-
amanitin resistant version of Pol II with a truncated CTD caused defects in splicing and polyadenylation of a reporter gene (McCracken et al. 1997). The termination defects were shown to occur through inefficient recruitment of polyadenylation machinery to the elongating polymerase (McCracken et al. 1997). Recruitment of splicing factors, and in particular SR proteins, to sites of transcription is also dependent on the CTD (Misteli and Spector 1999). SR proteins along with components of the U1 snRNP associate with the CTD and enhance splicing efficiency when present prior to initiation of transcription (Das et al. 2007).

Histone modifications and exon-intron gene structure

The understanding that transcription and splicing are closely coupled led to further investigations into the structure of chromatin and possible impacts on splicing, both constitutive and regulated. The advent of high-throughput sequencing technologies has greatly advanced these studies, revealing higher-order patterns of histone modifications that suggest extensive crosstalk between the splicing machinery and chromatin.

Despite the repeated recognition of splicing signals by multiple spliceosome components throughout the splicing cycle, the process of exon definition within transcript space remains difficult to conceptualize. Recently, a number of groups found that exonic sequences are enriched for nucleosome occupancy compared to introns (Andersson et al. 2009; Kolasinska-Zwierz et al. 2009; Schwartz et al. 2009; Spies et al. 2009; Tilgner et al. 2009). This may be explained in part by the relatively higher GC content of exons, as nucleosomes tend to be excluded from AT-rich regions (Spies et al. 2009). Increased nucleosome density within exons is conserved from plants to mammals (Nahkuri et al. 2009), and nucleosome occupancy tends to correlate with the degree of exon inclusion (Schwartz et al. 2009). The intriguing observation
Chapter 1. Introduction

that the average vertebrate exon is roughly the same length as the length of DNA that wraps around a single nucleosome further suggests an inherent splicing signal within the organization of chromatin (Schwartz et al. 2009; Tilgner et al. 2009).

A key observation made during this time was that specific histone marks, in particular histone 3 lysine 36 trimethylation (H3K36me3), are enriched on exons in actively transcribed genes even after normalizing for increased exonic nucleosome occupancy (Andersson et al. 2009; Spies et al. 2009; Huff et al. 2010; de Almeida et al. 2011). Accumulation of the H3K36me3 mark begins at the first exon of most expressed genes and continues to increase in a stepwise manner on exons across the gene body (Huff et al. 2010; de Almeida et al. 2011). Interestingly, the first intron of active genes is characterized by the histone 3 lysine 79 dimethylation (H3K79me2) mark, and a distinct border between this and the beginning of the H3K36me3 domain occurs at the first 3’SS (Huff et al. 2010). The H3K36me3 mark also preferentially marks constitutive exons and is less enriched on alternative exons (Andersson et al. 2009; Kolasinska-Zwierz et al. 2009). More recently, it was found that blocking of splicing by splice site mutagenesis in minigene constructs or globally using small-molecule inhibitors causes a redistribution or decrease in H3K36me3 marks (de Almeida et al. 2011; Kim et al. 2011b). This occurs through the splicing-dependent recruitment of the H3K36 methyltransferase, HYPB/SETD2, to exonic regions (de Almeida et al. 2011). In yeast and human, this methyltransferase interacts with serine 2-phosphorylated Pol II CTD (Pol II ser2p) or hyperphosphorylated Pol II CTD, respectively (Kohtz et al. 1994; Li et al. 2005; Sun et al. 2005; Vojnic et al. 2006). While conclusive causal links between chromatin and splicing regulation have yet to be demonstrated, these observations hint at a potential regulatory role of histone
marks in splicing and to a reciprocal interaction between the transcription and splicing machineries.

Models for co-transcriptional splicing regulation

The coordination of intron removal with transcription adds a multi-dimensional layer of regulation onto a system of cis-trans interactions that is already staggeringly complex. However, two general principles may explain much of the effect of co-transcriptional splicing, the first being the kinetic effects of transcription elongation and the second being the effect of splicing machinery recruitment to the chromatin (reviewed in Luco et al. 2011). Emerging support for both models indicates that the kinetic and recruitment effects may not be mutually exclusive \textit{in vivo} but most likely play specific roles in different contexts (Figure 1).

According to the “first come, first served” hypothesis, competition between splice sites should be affected by the length of time a splice site is available for recognition by the spliceosome. More recent evidence in support of this “kinetic” model suggests that the speed of transcription influences exon choice (reviewed in Luco et al. 2011). This can be due simply to the general rate of transcription dictated by the efficiency of elongation. Alternatively, transient pauses of the polymerase within the gene body may locally affect splice site recognition. The first suggestion that transcription kinetics play a role in splice site selection again came from differences between in vitro and in vivo splicing assays. In this case, usage of a splice site sequestered within a short-stemmed RNA hairpin depended on the length of the hairpin’s loop \textit{in vivo} but was never used in vitro. This implies that the splicing machinery can recognize the splice site only during a defined length of time that is prior to transcription of the complementary arm of the hairpin (Eperon et al. 1988). An alternative exon in fibronectin, exon 33, became the
focus of a number of studies aiming to further dissect co-transcriptional splicing regulation. Treatment of cells with histone deacetylase inhibitors decreases exon 33 inclusion, presumably by increasing Pol II processivity (Nogues et al. 2002). Similarly, in the presence of the transcriptional activator VP16, inclusion of exon 33 is dramatically decreased, while the co-localization of transcripts with SR proteins increases (Kadener et al. 2001). These results argue for an inverse correlation between Pol II processivity and splicing recognition and demonstrate that splicing factor recruitment is insufficient for exon inclusion. Perhaps the most convincing evidence for elongation kinetics impacting splice site regulation is increased exon 33 inclusion observed in human cells expressing a slow mutant of Pol II (de la Mata et al. 2003). An analogous effect was also seen in vivo for the Ultrabithorax gene in Drosophila harboring a slow Pol II elongation mutation (de la Mata et al. 2003).

While general transcription rates affect exon inclusion for several splicing events, the specific induction of Pol II pause sites near alternative exons seems to be sufficient to cause a splicing change in certain contexts. Insertion of MAZ binding sites, which promote Pol II pausing by recruitment of the MAZ zinc finger protein, near an alternative exon in the tropomyosin gene leads to increased inclusion of the exon (Roberts et al. 1998). The human CD44 gene, which undergoes complex alternative exon choice in its central variable domain, displays enhanced exon inclusion mediated by the chromatin remodeling protein Brm. This is concurrent with a specific increase in Pol II ser5p over the variable region, suggestive of Pol II pausing and efficient splice site recognition (Batsche et al. 2006). In yeast, Pol II pausing at 3’ splice sites is accompanied by a phosphorylation switch from Ser2p to Ser5p (Alexander et al. 2010). Although this occurs at a typical yeast single-intron gene, it raises intriguing parallels to
mammalian systems for kinetic checkpoints in alternative splicing mediated by local changes in Pol II phosphorylation.

Several studies investigating the coupling of transcription and splicing reported that changing the identity of the promoter driving the fibronectin exon 33 minigene influences inclusion levels (Cramer et al. 1997; Cramer et al. 1999). Somewhat unexpectedly, this is independent of global or local elongation rates and is attributed to differential recruitment of splicing factors (Cramer et al. 1999). Observations such as these led to the “recruitment” model of co-transcriptional splicing regulation, whereby association of splicing machinery to the chromatin directly impacts splice site choice independent of changes in transcription rates.

Chromatin structure has become increasingly relevant to the recruitment model and is thought to associate with splicing machinery via various adaptor proteins. In yeast, the histone acetyltransferase activity of Gcn5, a member of the SAGA complex, is required for co-transcriptional recruitment of components of the U2 snRNP (Gunderson and Johnson 2009). Another SAGA component in mammals, CHD1, recognizes the H3K4me3 mark and increases U2 snRNP recruitment and splicing efficiency (Sims et al. 2007). The related mammalian STAGA complex similarly co-purifies with U2 snRNP components (Martinez et al. 2001). These cross-species findings suggest co-transcriptional recruitment of splicing machinery may be evolutionarily conserved. In further support of a conserved mechanism, the HP1 proteins have been shown to associate with the H3K9me3 mark in *Drosophila*, where they serve to recruit hnRNPs (Piacentini et al. 2009), as well as in human, where they interact with SR proteins (Loomis et al. 2009). Most recently, it was found that the H3K36me3 mark in human recruits the alternative splicing regulator PTB through the adaptor protein MRG15, resulting in cell type-dependent splicing regulation (Luco et al. 2010). Together, these examples represent emerging
insights into the diversity and complexity of chromatin-dependent splicing. How the kinetic and recruitment models interact to establish tissue-specific, heritable splicing patterns is an area of active investigation.

1.3 Functional consequences of coordinated splicing decisions

Since the seminal discoveries of its existence, many efforts have focused on deciphering the biological impact of pervasive alternative splicing. Alternative exon usage generally has one of two effects on transcriptional output. First, regulated exon inclusion or skipping can change the protein coding sequence and thus alter protein stability or function. Second, alternative splicing can affect gene expression by impinging on downstream regulatory pathways, such as nonsense-mediated decay (NMD). The coordinated impact of both of these regulatory modes defines the splicing patterns in a particular cellular context.

Alternative splicing changes protein function

Although nearly all mammalian genes undergo alternative splicing in some context, the physiological consequences of most regulated splicing events have been under much scrutiny. Most tissues express multiple isoforms of a given gene, and only a small fraction of exons undergo tissue-restricted, “switch-like” alternative splicing changes in which an exon is predominantly included in one tissue and skipped in another (Wang et al. 2008). However, this subset of exons tends to preserve reading frame and is highly conserved (Xing and Lee 2005; Wang et al. 2008). This is especially true of mutually exclusive splicing events, in which only one of two neighboring alternative exons is included in the final transcript (Wang et al. 2008).
Switch-like exons may thus be more likely to produce functionally distinct, full-length protein products than their less tissue-specific counterparts.

Individual examples of splicing switches that change the protein-coding sequence have been shown to play roles in mammalian differentiation, metabolism, and stress response. Embryonic stem cell (ESC) pluripotency is maintained by a core set of transcription factors, one of which is Oct4 (Boyer et al. 2005). Two predominant isoforms of Oct4 are expressed during development, Oct4A and Oct4B, of which Oct4A is highly expressed in ESCs (Takeda et al. 1992; Cauffman et al. 2006). Oct4B has a distinct N-terminal domain and a shorter POU DNA-binding domain when compared to Oct4A that arises from usage of an alternative promoter and alternative splicing (Cauffman et al. 2006). While Oct4A is required for pluripotency, Oct4B fails to sustain ESC survival due to inhibited DNA binding by the unique N-terminal domain (Lee et al. 2006).

Another splicing switch pertinent during development as well as in tumorigenesis is a mutually exclusive splicing event resulting in the M1 and M2 isoforms of pyruvate kinase, the enzyme responsible for catalyzing the formation of pyruvate and ATP during glycolysis. These isoforms differ by 22 nucleotides and possess distinct metabolic activities. While M1 forms a constitutively-active tetramer, M2 tetramerization and activation is dependent on the binding of upstream intermediates in glycolysis (Christofk et al. 2008; Anastasiou et al. 2012). PKM1 is generally expressed in adult tissues with high ATP requirements, such as brain and muscle, while PKM2 is expressed in embryonic and several adult tissues as well as in cancer cell lines (Clower et al. 2010; Mazurek 2011). Splicing of PKM2 in glioma, for example, has been attributed to the c-Myc-driven activation of the splicing repressors PTB and hnRNPA1/A2 (David et al. 2010). Importantly, the switch to the regulatable activity of PKM2 in tumors seems to be important in
achieving the dynamic metabolic requirements of non-proliferating and proliferating tumor cells (Israelsen et al. 2013).

Other protein-coding splicing changes function within autoregulatory feedback loops to propagate or repress the response to a signal. The E3 ubiquitin ligase MDM2 is a critical regulator of the tumor suppressor p53. An alternative splicing event within the MDM2 transcript produces a dominant negative isoform of MDM2 that sequesters MDM2 in the cytoplasm, resulting in sustained p53 activation (Evans et al. 2001). This splicing switch is triggered by genotoxic stress such as UV irradiation (Chandler et al. 2006) and can be overcome by the transcriptional upregulation of MDM2, a direct target of p53 (Haupt et al. 1997). Such self-limiting feedback loops demonstrate the ability of splicing switches to achieve rapid functional outputs on short timescales.

The broader effects of changing the protein-coding capacity of splicing networks has recently been made clearer by high-throughput sequencing assays. An enrichment in alternative splicing events has been reported in transcripts encoding transcription factors, splicing factors, cytoskeletal proteins, and kinases (Talavera et al. 2009; Kosti et al. 2012; Merkin et al. 2012). While most of these splicing events may not lead to distinct changes in protein domains, an investigation of neural-specific alternative exons suggested that they are enriched in putative unstructured protein-protein interaction domains (Ellis et al. 2012). This is also true of many other tissue-specific exons, particularly those within developmental or signaling regulators, suggesting that concerted splicing changes may lead to rewiring of protein interaction networks (Buljan et al. 2012; Ellis et al. 2012). Along similar lines, highly tissue-restricted splicing events are enriched in genes encoding phosphoproteins. Tissues with high kinase activity tend to have more variable splicing of their target phosphoproteins, indicating that alternative exon inclusion
can effectively modulate kinase activity (Merkin et al. 2012). These observations imply that splicing changes can function in concert to regulate protein networks.

*Alternative splicing alters gene expression*

An increasingly appreciated consequence of alternative exon inclusion is a change in gene expression. When the inclusion or skipping of an alternative exon introduces a premature termination codon (PTC) upstream of the canonical termination codon, the resulting transcript can become a target of nonsense-mediated mRNA decay (NMD). The NMD pathway is highly conserved; however, it differs in a number of mechanistic details between yeast and mammals, and this introduction will focus mainly on mammalian NMD.

Broadly speaking, the NMD pathway is a surveillance mechanism that degrades transcripts containing a PTC, thus preventing the translation of potentially deleterious truncated proteins (reviewed in Kervestin and Jacobson 2012). The obvious difficulty in this process lies in how to faithfully distinguish a premature from a normal termination codon. It is now becoming clear that several events in the nucleus and the cytoplasm work in concert to ensure this specificity. Early discoveries that a PTC results in low mRNA levels only when introduced in internal exons of spliced genes argued that the cytoplasmic decay mechanism is coupled to nuclear splicing (Urlaub et al. 1989; Cheng et al. 1994; Thermann et al. 1998). The identity of this splicing signature was later discovered to be a large SR protein-associated complex deposited 20-24 nt upstream of the majority of splice junctions while still in the nucleus (Le Hir et al. 2000a; Le Hir et al. 2000b; Singh et al. 2012). The components of this “exon-junction complex” (EJC) include the splicing factor SRm160, the shuttling factor Y14, and the mRNA export factor REF, and they remain at least partially associated with the mRNA upon transport.
into the cytoplasm (Le Hir et al. 2000a; Le Hir et al. 2000b). Additionally, the NMD factors Upf2 and Upf3 were found to associate with the cytoplasmic EJC (Le Hir et al. 2000b). Four mammalian proteins with homologs in yeast, Upf1, Upf2, Upf3a, and Upf3b, comprise the core NMD machinery and play various roles in detecting and degrading PTC-containing transcripts. The Upf3 proteins shuttle between the nucleus and cytoplasm, presumably through their interaction with the EJC (Le Hir et al. 2000b; Lykke-Andersen et al. 2000). In contrast, the NMD adaptor protein Upf2 displays perinuclear localization, while the ATP-dependent helicase Upf1 is cytoplasmic and is a critical mediator of NMD (Perlick et al. 1996; Applequist et al. 1997; Sun et al. 1998; Lykke-Andersen et al. 2000). Upon export from the nucleus, EJC's on the mRNA acquire Upf2 through an interaction with Upf3; additionally, Upf2 can interact with Upf1 (Lykke-Andersen et al. 2000; Mendell et al. 2000; Serin et al. 2001).

NMD is a translation-dependent process, and several lines of evidence suggest that it is the first, or “pioneer”, round in particular during which transcripts are susceptible to NMD. Recognition of a termination codon, either premature or normal, by the eukaryotic release factors eRF1 and eRF3 triggers translational termination. These release factors recruit Upf1 and its kinase, SMG-1, to the stalled ribosome to form the SURF complex (Czaplinski et al. 1998; Kashima et al. 2006). The key distinction between premature and normal termination is the presence of a downstream EJC, and the interaction of SMG-1 and the other components of the SURF complex with the EJC is critical for SMG-1-dependent phosphorylation of Upf1 (Kashima et al. 2006). These interactions follow a “distance rule” whereby a termination codon is determined to be premature if it is >50-55 nucleotides upstream of an exon-exon junction complex (Carter et al. 1996; Nagy and Maquat 1998). Once phosphorylated, Upf1 recruits a number of downstream effectors of NMD, including the terminal effector SMG-7, which directs
the transcript to processing bodies (P-bodies) for degradation (Ohnishi et al. 2003; Unterholzner and Izaurralde 2004; Fukuara et al. 2005). Destruction of PTC-containing transcripts in P-bodies proceeds through normal mRNA decay mechanisms, including decapping, deadenylation, and exonucleolytic decay (Chen and Shyu 2003; Lejeune et al. 2003; Couttet and Grange 2004).

Although NMD was found to degrade transcripts containing PTCs arising from mutation or errors in transcription or splicing, the close coupling to splicing offered the tantalizing possibility that the alternative inclusion of PTCs could allow for NMD-dependent gene regulation. This would depend on the introduction of a PTC upon inclusion or exclusion of an alternative exon (reviewed in McGliney and Smith 2008). Recent estimates from gene expression profiling suggest that in fact, as high as 10-30% of mammalian genes may be regulated by alternative splicing-coupled NMD (AS-NMD) in particular contexts (Lewis et al. 2003; Mendell et al. 2004; Weischenfeldt et al. 2012). Several studies implicate AS-NMD networks as being especially relevant during developmental transitions or stress conditions, perhaps due to the rapid changes in protein expression afforded by regulation at the level of splicing as opposed to transcription (Thoren et al. 2010; Wong et al. 2013).

While AS-NMD may thus regulate splicing networks through various means, one particular feature of AS-NMD deserves particular attention. This is the ability of AS-NMD to function in negative autoregulatory feedback loops. A key observation leading to this discovery was that AS-NMD events tend to be enriched in genes encoding splicing factors and are often ultraconserved, meaning they fall within regions of particularly high conservation (Lareau et al. 2007; Ni et al. 2007). The model proposed by these studies was that SR and hnRNP proteins are able to bind their own transcripts, cause splicing of the NMD variant, and downregulate protein levels to maintain homeostatic protein expression (Lareau et al. 2007; Ni et al. 2007; Saltzman et
Chapter 1. Introduction

al. 2008) (Figure 2). How variations in steady-state expression can be achieved in the presence of tight autoregulation remains to be determined. One possibility is that cross-regulation of AS-NMD by other RNA binding proteins modulates the efficiency of autoregulation, thus leading to differences in steady-state expression. While such a mechanism has been reported for several individual splicing factors (Boutz et al. 2007; Calarco et al. 2009; Rossbach et al. 2009), the more widespread impact of AS-NMD in the regulation of interconnected splicing networks is still largely unknown.

Establishing cell type-specific splicing networks

Upstream of functional splicing changes, the coordinated expression of constitutive and alternative splicing factors activates broad splicing regulatory networks. Most recent studies, aided by advances in high-throughput sequencing technologies, have focused on the effect of individual splicing regulators in maintaining the splicing patterns of a certain cellular condition. Importantly, changes in individual core or alternative splicing regulators frequently have pleiotropic effects that can lead to dramatic changes in cell state.

Despite their classification as core splicing regulators, specific basal splicing factors may have differential effects on different introns. For example, knockdown of the core snRNP protein SmB/B’ results in altered splicing of a number of RNA processing factors, including both NMD and non-NMD splicing events (Saltzman et al. 2011). In a more physiological context, mutations in the core splicing machinery have been implicated in the etiology of myelodysplastic syndromes, presumably by altering proliferation-specific splicing networks. Specifically, factors involved in branch point and 3’ splice site recognition, including Sf3b1 and U2AF35, are frequently mutated in such syndromes (Yoshida et al. 2011; Graubert et al. 2012). The specific
effect on cell proliferation may be explained by splicing changes seen after treatment of cells in culture with spliceostatin A, an inhibitor of Sf3b155 and anti-tumor drug, which results in cell cycle defects and decreased proliferation (Corrionero et al. 2011). The sensitivity of proliferation-specific splicing events to spliceostatin A treatment or mutations in branch point recognition components defines a coherent, regulatable splicing network.

More commonly, the activation of splicing networks can be traced to changes in the expression of alternative splicing regulators. One of the best-characterized splicing networks is regulated by the Nova proteins, Nova1 and Nova2, in various cell types of the brain. Nova1 and Nova2 exhibit largely non-overlapping patterns of expression in different regions of the brain across development (Yang et al. 1998). As with many splicing regulators, Nova proteins recognize a short sequence motif (in this case, YCAY) to regulate either exon inclusion or skipping, depending on the location of binding with respect to the regulated exon (Ule et al. 2006). Integrating this RNA map with splicing changes observed in Nova knockout mouse brains identified ~700 Nova-dependent splicing targets (Zhang et al. 2010). Many of these targets are involved in regulating synapse activity and axonal guidance (Ule et al. 2005). Regulation of a component of the reelin signaling pathway by Nova2, for example, results in defects in neuronal cell migration (Yano et al. 2010). Notably, Nova1 and Nova2 may regulate distinct splicing events with distinct functional outcomes. The functional redundancy of these related factors thus warrants further investigation.

An important characteristic of tissue-specific splicing networks is that although one or two key regulators may largely define a cell’s splicing patterns, co-regulation by multiple factors is most likely necessary for their stability. In support of this concept, up to 15% of Nova-dependent splicing events also have binding sites for the Rbfox proteins (Zhang et al. 2010).
Precise temporal regulation of such splicing factors may thus lead to developmentally-timed shifts in alternative splicing networks.

1.4 The Rbfox family of splicing regulators

Over the last decade, Rbfox proteins have emerged as important regulators of tissue-specific splicing in numerous contexts. First characterized as splicing factors in zebrafish, Rbfox proteins have conserved RNA binding activity to (U)GCAUG motifs (reviewed in Kuroyanagi 2009). Recent computational approaches have identified hundreds of target exons, particularly in brain and muscle. Despite these mechanistic and global analyses, the breadth of Rbfox splicing regulation and the means by which (U)GCAUG motifs are defined as functional remain unclear. This thesis investigates the mechanisms involved in defining the Rbfox2 splicing regulatory network in mouse embryonic stem cells.

Rbfox proteins in development and disease

Fox-1 (Feminizing locus On X) was first described as an initiating element of the sex determination pathway in C. elegans (reviewed in Kuroyanagi 2009). These studies suggested that fox-1 is an RNA binding protein that most likely functions in dosage compensation through post-transcriptional regulation of the critical sex determination gene xol-1 (Hodgkin et al. 1994; Skipper et al. 1999). Following these reports, the RNA recognition motif (RRM)-containing protein zFox-1 in zebrafish was identified as a homolog of nematode fox-1 and the mammalian ataxin2-binding protein A2BP1 (Hodgkin et al. 1994; Skipper et al. 1999). In vitro selection revealed the pentamer GCAUG to be the primary sequence recognized by zFox-1 (Jin et al. 2003). In previous experimental and computational analyses, this pentamer had been found to
regulate the inclusion or skipping of several conserved alternative exons (Huh and Hynes 1994; Kawamoto 1996; Hedjran et al. 1997; Modafferi and Black 1997; Lim and Sharp 1998; Deguillien et al. 2001) and was enriched around brain- and muscle-specific exons (Brudno et al. 2001), although the factor responsible for mediating the regulation was not known. Combining these observations, Jin et al. demonstrated that recognition of the (U)GCAUG motif by Fox-1 homologs in zebrafish and mammals regulates the alternative splicing of several exons (Jin et al. 2003). The NMR structure of the Rbfox1 RRM, in complex with the sequence 5’-UGCAUGU-3’, subsequently revealed a strong affinity for this motif due to a non-canonical interaction with the β-sheet of the RRM (Auweter et al. 2006).

Three mammalian homologs of fox-1 have since been characterized as splicing regulators and recently renamed as Rbfox1 (Fox-1 or A2BP1), Rbfox2 (Fox-2 or RBM9), and Rbfox3 (Fox-3, HRNBP3, or NeuN). Rbfox1 and Rbfox2 share identical RRM domains, while Rbfox3 is 94% identical in its RRM and also regulates splicing through recognition of (U)GCAUG motifs (Kim et al. 2011a; Gehman et al. 2012). All three proteins are highly expressed in neurons, although each shows somewhat distinct expression patterns in subsets of neurons (Underwood et al. 2005; Kim et al. 2009). Rbfox1 is also present in muscle precursors and increases in expression during postnatal heart development (Underwood et al. 2005; Kalsotra et al. 2008). Rbfox2 expression, on the other hand, has been reported in a wider variety of tissues and cell lines, including muscle precursors and neuronal cells as well as embryonic stem cells and a number of epithelial and mesenchymal tumors and cell lines (Nakahata and Kawamoto 2005; Underwood et al. 2005; Yeo et al. 2007b; Lee et al. 2009; Venables et al. 2009; Shapiro et al. 2011; Venables et al. 2013a; Venables et al. 2013b). These diverse expression patterns are also reflected in the various phenotypes associated with altered Rbfox2 expression or activity. In
human embryonic stem cells, knockdown of RBFOX2 results in increased levels of apoptosis (Yeo et al. 2009). Deletion of Rbfox2 in the murine central nervous system causes severe defects in cerebellar development and also disrupts normal motor function in mature animals (Gehman et al. 2012). In a disease context, altered splicing of Rbfox2 and its targets has been detected in a number of breast carcinomas (Venables et al. 2009), and Rbfox2 has been implicated in regulating a mesenchymal-specific splicing network in cell culture models of epithelial-mesenchymal transition (Shapiro et al. 2011; Venables et al. 2013a). Taken together, unlike Rbfox1 and Rbfox3 which show tissue-restricted expression often in terminally-differentiated cell types, Rbfox2 may play critical roles earlier in development and in adult progenitor cell types, and its normal functions may also be co-opted by cancer cells to enhance tumorigenicity.

Mechanisms of Rbfox activity

For many splicing regulators including the Rbfox proteins, binding within the intron upstream of a cassette exon tends to repress the exon’s inclusion, whereas binding downstream generally enhances its inclusion. Many mechanistic studies have addressed how the positional effects of Rbfox binding modulate recruitment of the splicing machinery. Upstream binding is thought to sterically hinder splicing machinery responsible for recognizing the branch point and 3’ splice site. For example, splicing of calcitonin/CGRP exon 4 is antagonized via two UGCAUG motifs located in the upstream intron and within the exon. Binding of Rbfox1/2 to the upstream motif occludes SF1 recognition of the branch point, thus preventing formation of the pre-spliceosomal E’ complex. Subsequent Rbfox1/2 binding within the exon blocks usage of an exonic splicing enhancer and prevents U2AF65 recruitment to the 3’ splice site, ensuring that
any E’ complexes that may have bypassed the first inhibitory step will not proceed to the E complex (Zhou and Lou 2008). Similarly, Rbfox2 binds the intronic region upstream of Fgfr2 exon IIIc and cooperates with hnRNP H1 to block SR protein recognition of exonic splicing enhancers, thereby repressing exon IIIc inclusion (Baraniak et al. 2006; Mauger et al. 2008). One possible mechanism has been suggested for exon activation by Rbfox proteins. The C-terminal domain of Rbfox2 interacts with the zinc finger domain of the U1 component U1C, thus enhancing U1 recruitment to the weak 5’SS of protein 4.1 exon 16 and causing exon inclusion. This is mediated by a UGCAUG motif located downstream of the 5’ splice site (Huang et al. 2012). Thus, cooperation between Rbfox proteins and other splicing factors works to inhibit or enhance spliceosome assembly on alternative exons.

Genome-wide computational approaches have corroborated this position-dependent regulation by Rbfox proteins. Early studies found UGCAUG and GCAUG to be highly significantly enriched downstream of phylogenetically conserved brain- and muscle-specific alternative exons (Brudno et al. 2001; Minovitsky et al. 2005). A more comprehensive analysis of motifs enriched around all internal exons across several mammalian genomes showed that the UGCAUG motif is strongly enriched both upstream and downstream of alternative exons (Yeo et al. 2007a). Coupling this analysis to splice-junction arrays and transcriptome sequencing, several studies correlated proximal upstream and downstream motif enrichment to exon repression and inclusion, respectively (Das et al. 2007; Kalsotra et al. 2008; Wang et al. 2008; Zhang et al. 2008). While motifs close to alternative exons correlate particularly well in these analyses, distal UGCAUG motifs located hundreds of nucleotides away are also functional in regulating splicing (Lim and Sharp 1998; Parra et al. 2012; Lovci et al. 2013).
Chapter 1. Introduction

Defining the Rbfox2 regulatory network

In spite of these ongoing experimental and computational dissections of Rbfox function, a number of puzzling questions remain. First, the minimal Rbfox pentamer, GCAUG, should occur within transcripts with a frequency of approximately one every kilobase. At this fairly high frequency, it is unlikely that every occurrence of the motif is bound to regulate splicing, yet it is unclear how the selection for functional motifs is made. Indeed, a CLIPseq study of human RBFOX2 in embryonic stem cells revealed that many UGCAUG elements within expressed introns show no evidence of binding, even while other sites in the same transcript are bound (Yeo et al. 2009). Conversely, this same study suggested that only one-third of RBFOX2 binding sites contain the minimal pentamer, arguing for additional and as-yet unidentified determinants of binding (Yeo et al. 2009). Moreover, further investigation is needed to separate the Rbfox-specific effects from co-regulated splicing events within tissue-specific splicing networks. This thesis describes mechanisms and consequences of Rbfox2 activity in mouse embryonic stem cells (mESC), in which Rbfox2 is the sole expressed Rbfox paralog. We determine binding sites of Rbfox2 across the mESC transcriptome and uncover a widespread network of splicing regulation mediated through splicing-coupled decay pathways. Using a novel technique to assay RBP-RNA interactions, we deconvolute the structural and sequence preferences for Rbfox2 binding. We further explore the role of chromatin in determining functional Rbfox2 binding sites and propose kinetic effects of chromatin modifications on Rbfox2-dependent splicing events. Together, these studies reveal significant insights into the complexities of defining and maintaining cell type-specific splicing networks.
1.5 REFERENCES


Chapter 1. Introduction


Chapter 1. Introduction


Fukuhara N, Ebert J, Unterholzner L, Lindner D, Izaurralde E, Conti E. 2005. SMG7 is a 14-3-3-like adaptor in the nonsense-mediated mRNA decay pathway. Molecular cell 17: 537-547.


Chapter 1. Introduction


Chapter 1. Introduction


Chapter 1. Introduction


Chapter 1. Introduction


Chapter 1. Introduction


Chapter 1. Introduction


Chapter 1. Introduction


Wetterberg I, Bauren G, Wieslander L. 1996. The intranuclear site of excision of each intron in Balbiani ring 3 pre-mRNA is influenced by the time remaining to transcription termination and different excision efficiencies for the various introns. *Rna* **2**: 641-651.


Figure 1. Models for co-transcriptional splicing regulation
The kinetic model of splicing posits that the speed of transcription elongation dictates splice site choice (top panel). If transcription rates are slow (RNA Pol II shown in green), weak splice sites around alternative exons (shown in red) are preferentially recognized by alternative regulators.
(orange) and are included in the final transcript. If transcription rates are fast, weak splice sites are not detected, and exon skipping occurs. The recruitment model suggests that alternative regulators (orange) interact with histone modifications proximal to alternative exons, resulting in regulation of the exon by the alternative splicing factor. In this case, for a splicing activator, recruitment to the modified histone leads to exon inclusion.
Figure 2. Autoregulatory splicing
NMD splicing events within the transcripts of splicing activators (green) and splicing repressors (red) are ultraconserved (blue schematic conservation track). Splicing activators tend to produce NMD substrates through the introduction of a premature termination codon upon exon inclusion, while splicing repressors tend to produce NMD substrates upon exon skipping. In both cases, negative autoregulation results in homeostatic control of steady-state expression levels.
Chapter 2

Rbfox2 controls autoregulation in RNA binding protein networks

This chapter is adapted from the following manuscript:


Experimental contributions:

Mohini Jangi performed Rbfox2 iCLIP, NMD and autoregulation experiments, and computational analyses, with input from Paul Boutz. Paul Boutz performed Rbfox2 knockdown RNaseq experiment. Prakriti Paul performed RNAseq validation splicing assays.
Chapter 2. *Rbfox2 controls autoregulation in RNA binding protein networks*

### 2.1 ABSTRACT

The tight regulation of splicing networks is critical for organismal development. To maintain robust splicing patterns, many splicing factors autoregulate their expression through alternative splicing-coupled nonsense-mediated decay (AS-NMD). However, as negative autoregulation results in a self-limiting window of splicing factor expression, it is unknown how variations in steady-state protein levels can arise in different physiological contexts. Here, we demonstrate that Rbfox2 cross-regulates AS-NMD events within RNA binding proteins to alter their expression. Using individual nucleotide-resolution crosslinking immunoprecipitation coupled to high-throughput sequencing (iCLIP) and mRNA sequencing, we identify more than 200 AS-NMD splicing events that are bound by Rbfox2 in mouse embryonic stem cells. These “silent” events are characterized by minimal apparent splicing changes but appreciable changes in gene expression upon Rbfox2 knockdown due to degradation of the NMD-inducing isoform. Nearly 70 of these AS-NMD events fall within genes encoding RNA binding proteins, many of which are autoregulated. As with the coding splicing events we find to be regulated by Rbfox2, silent splicing events are evolutionarily conserved and frequently contain the Rbfox2 consensus, UGCAUG. Our findings uncover an unexpectedly broad regulatory network controlled by Rbfox2 and offer an explanation for how autoregulatory splicing networks are tuned.
2.2 INTRODUCTION

Alternative pre-mRNA splicing is essential for the context-specific diversity of metazoan transcriptomes (Nilsen and Graveley 2010). Frequently, alternative exon usage leads to variations in protein coding sequence and protein function. Alternative splicing can also impact gene expression post-transcriptionally through the nonsense-mediated mRNA decay (NMD) pathway (Lewis et al. 2003; Weischenfeldt et al. 2012). In a given cellular condition, splicing patterns are dictated by the precise balance of expressed trans-acting splicing factors. Most of these splicing factors recognize short sequence motifs, allowing them to regulate large, interconnected splicing networks (Irimia and Blencowe 2012). In spite of this apparent redundancy, altering the levels of an individual splicing factor can drastically reprogram splicing patterns both in normal development and disease (Yoshida et al. 2011; Han et al. 2013; Xue et al. 2013). It remains unclear how RNA binding protein expression is tightly regulated to direct these splicing programs in some cases and to prevent their aberrant activation in others.

Alternative splicing is often coupled to NMD, a post-transcriptional regulatory pathway that prevents the production of truncated proteins from transcripts containing premature termination codons (PTC) (reviewed in Kervestin and Jacobson 2012). Both during the pioneering round of translation and in subsequent rounds (Durand and Lykke-Andersen 2013; Rufener and Muhlemann 2013), the presence of a termination codon located $>50$ nt upstream of the last exon-exon junction triggers transcript degradation by the NMD machinery (Le Hir et al. 2001). In addition to working as a quality control mechanism to avert errors in splicing and transcription, NMD can regulate gene expression when coupled to alternative splicing (AS-NMD) (Lewis et al. 2003). This can occur, for example, through the regulated inclusion of a “poison” alternative exon containing a PTC. Intriguingly, AS-NMD events are enriched among
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks

members of the SR and hnRNP families of splicing regulators, where they occur within regions of high phylogenetic conservation (Lareau et al. 2007; Ni et al. 2007). Many of these proteins and other core spliceosomal components bind their own transcripts and catalyze splicing of the nonproductive NMD isoform, resulting in an autoregulatory negative feedback loop that limits protein expression (Lareau et al. 2007; Ni et al. 2007; Saltzman et al. 2008; Saltzman et al. 2011). This homeostatic mechanism also accommodates cross-regulation by other splicing factors, as has been demonstrated for several hnRNPs that activate AS-NMD of related family members (Wollerton et al. 2004; Boul et al. 2007; Rossbach et al. 2009). Direct negative feedback of a gene product on its coding transcript defines a self-limiting range of gene expression. How this level is modulated to allow for different steady-state protein concentrations is unknown. In addition, as global detection of AS-NMD events is hindered by the inherent instability of PTC-containing transcripts, the extent to which AS-NMD regulates gene expression has not been determined.

The Rbfox RNA binding proteins are important tissue-specific and signal-responsive alternative splicing regulators that function through binding (U)GCAUG motifs (Jin et al. 2003; Nakahata and Kawamoto 2005; Underwood et al. 2005; Auweter et al. 2006; Ponthier et al. 2006). Unlike its paralogs Rbfox1 (Fox-1 or A2BP1) and Rbfox3 (Fox-3, HRNBP3, or NeuN), which are predominantly expressed in muscle and neuronal tissues, Rbfox2 (Fox-2 or RBM9) is the sole expressed Rbfox family member in embryonic stem cells and a number of epithelial and mesenchymal cancer cell lines (Jin et al. 2003; Nakahata and Kawamoto 2005; Lee et al. 2009; Venables et al. 2009; Warzecha et al. 2009). Rbfox2 has been implicated in coordinately regulating a mesenchymal-specific splicing network in cell culture models of tumorigenesis and mesodermal differentiation (Venables et al. 2009; Shapiro et al. 2011; Venables et al. 2013), is
required for viability of human embryonic stem cells (Yeo et al. 2009), and is essential for normal cerebellar development in mouse (Gehman et al. 2011; Gehman et al. 2012). These observations suggest a broad role for Rbfox proteins, and in particular the more ubiquitously-expressed Rbfox2, throughout many stages of development.

Bioinformatic analyses using the high-information consensus motif recognized by the Rbfox proteins have predicted many target splicing events (Auweter et al. 2006; Wang et al. 2008; Zhang et al. 2008). Rbfox motifs are enriched upstream of tissue-specific repressed exons and downstream of enhanced exons (Wang et al. 2008; Zhang et al. 2008; Barash et al. 2010; Merkin et al. 2012; Ray et al. 2013), and RBFOX2 crosslinking immunoprecipitation-sequencing (CLIPseq) in human embryonic stem cells supports this context-dependent regulation (Yeo et al. 2009). While Rbfox2 recognizes its motif with high fidelity in vitro (Auweter et al. 2006), only one-third of in vivo binding events have been attributed to recognition of the minimal GCAUG motif (Yeo et al. 2009). Equally puzzling, many UGCAUG elements in transcribed regions show no evidence of binding. Despite these computational and experimental investigations, the explanation for why only a subset of motifs appears to be selected to regulate splicing in a given cellular context remains unknown. The lack of comprehensive Rbfox2 binding sites and Rbfox2-dependent gene expression and splicing changes within the same cell type has precluded a thorough examination of these questions.

Here, we generated a global analysis of Rbfox2 splicing regulation combined with a highly specific, single nucleotide-resolution Rbfox2 RNA binding map in mouse embryonic stem cells (mESC), where Rbfox2 is the only expressed Rbfox paralog. We find that Rbfox2 regulates the splicing and expression of many previously unknown targets, and particularly a number of RNA binding proteins (RBPs), by modulating AS-NMD. This demonstrates a novel
advantage of CLIPseq data in uncovering “silent” alternative splicing events, for which the instability of NMD target transcripts dampens true changes in isoform ratios. Based on our observations of RBP-Rbfox2 co-regulation with a polarity predicted by Rbfox2 binding, we propose a model whereby Rbfox2 tunes autoregulatory splicing events to control RBP expression levels. Our results position Rbfox2 at a critical node in a broader splicing regulatory network with roles in both normal development and disease settings.

2.3 RESULTS

Rbfox2 iCLIP is enriched in introns and alternatively spliced regions

To determine the binding sites of Rbfox2 across the mouse embryonic stem cell (mESC) transcriptome, we performed a modified iCLIP protocol optimized to provide nucleotide-level resolution and high specificity. We stably transduced V6.5 mESC with a lentiviral vector carrying FLAG-HA tagged or untagged human RBFOX2 (FHFOX2 or UntagFOX2) expressed from a doxycycline-inducible promoter. Titration of doxycycline allowed for expression of the transgene at levels similar to endogenous Rbfox2, and both constructs were functional in regulating the splicing of candidate Rbfox2 targets (Fig.S1A,B). After sequencing of iCLIP libraries prepared from cells expressing either FHFOX2 or UntagFOX2 (Fig.S1C,D), reads were mapped to the genome and single-nucleotide crosslink sites were inferred. From this we obtained 5.64 million tagged and 1.21 million untagged crosslink sites (Fig.S2A,B), which we then classified by genomic region. Nearly 80% of FHFOX2 crosslinks mapped to the sense strand of genes, and 10% mapped to intergenic regions greater than 1 kilobase from genes. Only 3% of crosslinks aligned antisense to genes, suggesting that our protocol maintained a high degree of strand specificity. Notably, FHFOX2 crosslink sites were significantly enriched in
introns over UntagFOX2 crosslink sites, consistent with the known role of Rbfox2 in pre-mRNA splicing (Fig.1A).

Statistically significant genic clusters of FHFOX2 binding were determined next, using the total crosslink sites extended on either side by 12nt as input into the CLIPper algorithm (available at https://github.com/YeoLab/clipper; Lovci et al. 2013). Allowing for the calling of clusters within pre-mRNA regions, we identified 40,243 FHFOX2 clusters and only 2,961 UntagFOX2 clusters with a gene-level FDR <= 0.05 and a p-value <= 0.05. We additionally required that each FHFOX2 cluster be enriched over UntagFOX2 crosslinks within the bounds of the cluster using a binomial test with a q-value threshold of 0.05, reasoning that the crosslink site distribution in the UntagFOX2 iCLIP was representative of background signal (Fig.S2A). This process resulted in 35,639 high-confidence clusters. Similar to the distribution of crosslink sites, FHFOX2 iCLIP clusters were heavily weighted toward intronic regions (Fig. S3A).

Due to this intronic enrichment and Rbfox2’s known role in regulating alternative splicing through recognition of intronic sequence elements, we next asked whether FHFOX2 iCLIP signal was enriched around alternatively-spliced regions using an intron-centric approach. Each mouse intron was associated with unique categories of splicing events. Correcting for the difference in library size, FHFOX2 crosslink sites were dramatically enriched compared to UntagFOX2 in introns flanking cassette exons and mutually exclusive exons, while the number of crosslink sites in constitutively-spliced introns was slightly but significantly depleted (Fig.1B). All introns were next subdivided into those without FHFOX2 iCLIP clusters and those with clusters <200nt from either the 5’ or 3’ splice site. For both splice sites, the average splice site strength was weaker for introns marked by a splice site-proximal iCLIP cluster compared to unbound introns. The subset of introns flanking cassette exons, which as a class had
significantly weaker splice sites than constitutive introns, did not have significantly different splice site strengths between FHFOX2-bound and –unbound events (not shown). We observed that the subset of FHFOX2-bound introns flanking constitutive exons had significantly weaker splice sites than unbound constitutive introns (Fig.1C), suggesting that splice sites bound by FHFOX2 in mESCs may be more dependent on trans-acting regulators, in this case Rbfox2. Here, Rbfox2 may function to increase the efficiency of constitutive splicing events rather than influence the choice between alternative splice sites, as is the canonical role of splicing regulators.

Regions flanking the splice sites of cassette exons with nearby FHFOX2 clusters were more highly conserved across placental mammals. Exonic conservation for this subset was also significantly higher, indicative of preserved regulation and perhaps function of genes encoding Rbfox2-bound splicing events (Fig.1D). These data together confirmed that globally, FHFOX2 iCLIP signal was specific and bore hallmarks of the binding patterns of a bona fide regulator of conserved splicing events.

**RBFOX2 iCLIP is strongly enriched around (U)GCAUG motifs**

One defining characteristic of the Rbfox family of splicing regulators that distinguishes it from most other splicing factors is its recognition of a highly specific motif, (U)GCAUG (Jin et al. 2003; Underwood et al. 2005). While several individual Rbfox2-dependent splicing events have been specifically demonstrated to depend on this motif (Huh and Hynes 1994; Modafferi and Black 1997; Lim and Sharp 1998; Underwood et al. 2005; Baraniak et al. 2006), a previous RBFOX2 CLIP study in human suggested that only one-third of binding sites contained the minimal pentameric motif (Yeo et al. 2009). Due to the enhanced specificity and spatial
resolution of our iCLIP protocol, we found several lines of evidence arguing that most Rbfox2 binding was driven by the canonical motif. First, 51% of intronic FHFOX2 iCLIP clusters contained at least one GCAUG occurrence, and 6.1% contained a conserved GCAUG (Phastcons placental mammals > 0.4). We also performed an unbiased motif analysis on the top 5000 clusters ranked by crosslink frequency, and as expected, UGCAUG was the top-scoring hexamer (Fig. 2A). We found that 71% of clusters contained a motif significantly similar to the UGCAUG position-specific probability matrix generated from this analysis (FDR < 0.1, p < 0.001). Conversely, 16% of all conserved exact UGCAUG matches in expressed introns were within 100nt of a cluster. To determine the fraction of motifs bound at high resolution, distance plots of FHFOX2 crosslink sites to conserved UGCAUG motifs in all introns were generated. Twenty-eight percent of UGCAUG occurrences were within 5 nt of a crosslink site and 53% were within 300 nt, arguing that a subset of conserved UGCAUG motifs are specifically bound and detectable at high resolution (Fig.2B, left panel). Correspondingly, crosslinks peaked sharply at positions 2 and 6 in the motif, with a ~5-fold drop in signal beyond positions 1 and 7 (Fig.2B, right panel). This is likely due to direct interaction and crosslinking at the two uridine residues, as UV-induced crosslinks occur more frequently at uridine residues (Sugimoto et al. 2012), resulting in reverse transcription terminating one nucleotide upstream. Finally, in support of specific iCLIP signal reflecting functional binding, intronic UGCAUG motifs within 100nt of a FHFOX2 cluster were significantly more conserved than an expression-matched set of unbound UGCAUG motifs (Fig.2C). Taken together, Rbfox2 binds to a sizeable fraction of conserved UGCAUG elements across the mESC transcriptome and is largely dependent on the motif for binding. The strong preference for UGCAUG-containing binding sites in vivo is also in
agreement with a recent investigation of in vitro binding preferences of Rbfox2 (Lambert et al. 2014).

Two other motifs reached significance in the hexamer motif analysis, [CAG]C[AU]CAC (found in 6.8% of all clusters, FDR < 0.3, p < 0.0002) and UGUGUG (found in 18% of all clusters, FDR < 0.1, p < 0.0003), although iCLIP signal around these motifs was less pronounced than around UGCAUG motifs (Fig.2D-G). These motifs resemble the consensus motifs for hnRNPL and CELF/Cugbp1, respectively (Teplova et al. 2010; Ray et al. 2013). Motif analysis of clusters lacking the UGCAUG-related motif failed to produce significant or high information content motifs. Consistent with this observation, 39% of CCUCAC and 72% of UGUGUG motif matches co-occurred with a UGCAUG motif within 10nt flanking iCLIP cluster summits. To specifically examine their co-occurrence within clusters at high resolution, the distance of each secondary motif to the top motif was plotted (Fig.2H). In accordance with the crosslink coverage plots, the UGUGUG motif was present on either side of the UGCAUG motif, with a periodicity implying a longer stretch of UG dinucleotides. The CCUCAC motif was similarly enriched upstream and downstream of the consensus motif, although with lower overall frequency. Motifs recognized by other RNA binding proteins can thus occur in the proximity of Rbfox2 binding events but are unlikely to lead to Rbfox2 binding in the absence of a UGCAUG-related motif. Interestingly, splicing analysis of cardiac differentiation revealed enrichments of Rbfox, hnRNPL, and CELF motifs around differentially spliced exons (Kalsotra et al. 2008). This raises the possibility that Rbfox2 may compete with other splicing regulators recognizing proximal motifs in different cellular contexts.
Hundreds of splicing and gene expression changes in response to Rbfox2 loss

Our iCLIP data suggested that many Rbfox2 binding events were candidates for regulation of splicing due, for example, to enrichment within alternatively-spliced regions. To test this hypothesis and determine global patterns of Rbfox2-dependent splicing changes in mESC, we performed deep sequencing of mRNA (RNAseq) upon shRNA-mediated depletion of Rbfox2. Two control hairpins targeting GFP and luciferase (shLuc, shGFP) and two treatment hairpins targeting the 3’ UTR of Rbfox2 (shFox2-1, shFox2-2) were lentivirally transduced into V6.5 mESC in biological duplicate, and poly(A)-selected mRNA was prepared for sequencing on the Illumina platform. Analysis of annotated splicing changes between control and Rbfox2 knockdown using the MISO algorithm (Katz et al. 2010) revealed hundreds of putative Rbfox2-regulated splicing events, a number of which were successfully validated by RT-PCR (Fig. S4A,B). Splicing events were quantified using percent spliced in, or ψ, to determine the percent of total isoforms represented by one of two isoforms. High-confidence Rbfox2-dependent alternative splicing changes in each splicing category were defined as those with ∆ψ > = |5%| and a Bayes’ factor exceeding 5 in either of two comparisons: shLuc vs shFox2-1 or shGFP vs shFox2-2. We further required events that were significant in a given comparison to change in the same direction in the other comparison and termed these “concordant” events. In this manner, we detected concordant splicing changes in all annotated categories ranging from 4-13% of expressed events (Fig.S4A,C).

Rbfox2 RNA map reveals position-dependent regulation of cassette exons

Meta-exon analyses of CLIP for a number of factors including Nova, PTB, and Mbnl have drawn correlations between the location of the binding event with respect to the cassette exon and the
direction of splicing regulation (Ule et al. 2006; Licatalosi et al. 2008; Xue et al. 2009; Wang et al. 2012). Previously, motif analyses and minigene assays using Rbfox2 target cassette exons argued that binding within the ~200nt downstream of a cassette exon resulted in Rbfox2-dependent exon inclusion, whereas binding upstream of the cassette exon correlated with repression. To determine whether upstream and downstream binding of FHFOX2 globally correlated respectively with exon repression or activation, we overlayed the high-resolution FHFOX2 iCLIP binding data with the concordant events defined above. Crosslink sites were mapped in the introns flanking exons activated by Rbfox2 ($\Delta \psi \geq 0.05$, $BF \geq 5$) and exons repressed by Rbfox2 ($\Delta \psi \leq -0.05$, $BF \geq 5$) (Fig.3A). The resulting distribution showed that enriched crosslinking within the ~200nt downstream of the cassette exon as well as at more distant sites proximal to the downstream constitutive 3’ splice site correlated strongly with exon activation, whereas crosslinking within the ~300nt-400nt upstream of the cassette exon correlated weakly with exon repression. iCLIP signal around non-regulated and expressed cassette exons ($|\Delta \psi| \leq 0.02$, $BF \geq 3$) did not show enrichment in either the upstream or downstream introns.

We next asked whether the distribution of both conserved and all canonical Rbfox2 motifs explained the crosslinking pattern we observed. The distribution indeed showed an increased frequency of motifs within the first 200nt downstream of Rbfox2-enhanced cassette exons, with weaker motif enrichment upstream of exons repressed by Rbfox2 (Fig.3B). Interestingly, a subset of enhanced exons also contained UGCAUG motifs proximal to the downstream constitutive 3’ splice site, perhaps implicating distant motifs in Rbfox2-dependent exon inclusion (Lim and Sharp 1998; Parra et al. 2012; Lovci et al. 2013). Most of the motif
occurrences near regulated exons were well-conserved. This is consistent with stronger evolutionary selection maintaining robust and presumably functional splicing changes.

The motif and crosslink distributions defined a critical window +/- 300 nt around regulated cassette exons within which Rbfox2 binding was most potent. We further wanted to address whether the strength of Rbfox2 binding within these critical regions flanking cassette exons impacted the degree of regulation. To this end, all detected cassette exons with BF > 3 were separated into those with a two-fold or greater difference in expression-normalized crosslink density between the upstream or downstream 300 nt. Exons with primarily upstream or downstream binding were further subdivided into quartiles of normalized crosslink density and average Δψ was plotted within each of these quartiles. For both upstream- and downstream-bound cassette exons, higher crosslink density correlated with stronger repression and activation, respectively. Additionally, cassette exon repression by Rbfox2 tended to be less potent than exon activation (Fig.3C). This trend persisted upon decreasing the threshold of normalized upstream versus downstream crosslink density to 1.5-fold, suggesting that small relative differences in upstream versus downstream binding were sufficient to confer exon repression or activation.

**Silent FHFOX2 binding events are associated with nonsense-mediated mRNA decay**

The high-resolution RNA map defined above strongly correlated context-dependent binding and regulation for direct targets of Rbfox2. We subsequently asked whether binding within the critical 300 nt of an intron boundary was necessary or sufficient to confer splicing regulation by intersecting bound and regulated splicing events. Focusing first on direct targets of Rbfox2, we observed a large and statistically significant overlap between Rbfox2-regulated
concordant events (“regulated cassettes”) and events containing a FHFOX2 iCLIP cluster within 300nt of the cassette exon (“bound cassettes”) (Fig. 3D). Previous work has reported an involvement of Rbfox2 in regulating splicing of transcripts associated with epithelial-mesenchymal transition and cell migration (Venables et al. 2009; Shapiro et al. 2011). Gene ontology analysis using the functional annotation tool DAVID (Huang da et al. 2009b; Huang da et al. 2009a) revealed that genes containing the 61 splicing events bound and regulated by Rbfox2 were enriched for cytoskeletal and cytoplasmic terms after correction for multiple hypothesis testing and requiring a minimum FDR cutoff < 0.1 (Fig. 3E). Many of the detected targets in this category were in fact shown to be differentially regulated in epithelial and mesenchymal cell systems (Venables et al. 2009; Warzecha et al. 2009; Shapiro et al. 2011). The 220 splicing changes that were not bound by FHFOX2 did not show significant enrichment in any particular gene ontology category, most likely owing to the heterogeneity of indirect effects caused by Rbfox2 loss.

Although the subset of events both bound and regulated by Rbfox2 in our study was significant, the question remained as to why some cassette exon-proximal FHFOX2 binding events were not coupled with a splicing change. While it could not be excluded that these were countered by the action of other splicing factors or were simply not in the correct context to elicit a splicing change, we considered the possibility that these 87 splicing events were in fact Rbfox2-regulated but the changes were rendered silent through another mechanism. In contrast to the bound and regulated set, the splicing events that were significantly unchanged upon Rbfox2 knockdown but were bound by FHFOX2 iCLIP were most strongly enriched for the RNA binding protein gene ontology term (Fig. 3F). Upon closer inspection of the bound and unchanging loci, we were surprised to find that in 37 out of 87 instances, one of the two isoforms
appeared to be a substrate for degradation by nonsense-mediated mRNA decay (NMD), defined by the introduction of a premature termination codon (PTC) >50 nt upstream of the last splice junction. Isoforms that harbor a PTC arising from exon inclusion or skipping would not be detected in RNAseq data due to the instability of the transcript, and consequently, the change in splicing due to Rbfox2 knockdown would be underestimated for these variants. Alternative splicing-coupled nonsense-mediated mRNA decay (AS-NMD) has been documented as a regulatory mechanism for gene expression. In particular, the expression of several RNA binding proteins (RBPs) is autoregulated or cross-regulated through AS-NMD-dependent mechanisms (Wollerton et al. 2004; Boutz et al. 2007; Lareau et al. 2007; Ni et al. 2007). The striking observation that a large fraction of bound and apparently unregulated events were likely to generate NMD substrates motivated the hypothesis that Rbfox2 was controlling gene expression levels by regulating AS-NMD.

**Rbfox2 modulates AS-NMD events in RNA binding proteins**

Due to the documented autoregulatory capacity of many RBPs, we first focused on verifying whether Rbfox2 was regulating AS-NMD within RBPs. Manual curation of the cassette exons in RBPs with proximal FHFOX2 iCLIP clusters revealed that 45 were putative NMD substrates by generating a PTC upon exon inclusion (NMD-INC) and 20 upon exon skipping (NMD-SK). iCLIP signal was distributed both upstream and downstream of these exons, suggesting Rbfox2 could either promote or inhibit the NMD isoform depending on the context. Based on the positional regulation criteria we determined for cassette exons, we hypothesized that NMD-INC exons with greater downstream iCLIP density and NMD-SK exons with greater upstream iCLIP density would represent genes for which NMD is enhanced by
Rbfox2 ("NMD enhanced"). Conversely, NMD would be suppressed by Rbfox2 for genes in which NMD-INC exons are predominantly bound upstream or NMD-SK exons are predominantly bound downstream ("NMD suppressed") (Fig. 4A).

To specifically assay whether the bound exons within RBPs were in fact NMD-coupled splicing events regulated by Rbfox2, we chose six candidate splicing events within iCLIP crosslinked transcripts encoding RNA binding proteins to examine more closely. The six genes, Snrnp70 (U170K), Srsf7, Tra2a, Sf1, Ptbp2, and Tia1, contained silent NMD splicing events with 0-5% splicing change upon Rbfox2 knockdown and distinct predicted effects of Rbfox2 binding. Cells stably expressing hairpins against Luciferase or Rbfox2 were treated with cycloheximide for 6h to block translation, which is essential for NMD. The type of splicing event putatively responsible for NMD (exon inclusion or exon skipping) was determined and primers were designed to assay relative levels of the NMD isoform, normalized to total transcript levels, by RT-qPCR for each condition. All six predicted NMD isoforms increased in abundance upon cycloheximide treatment, suggesting that these transcripts were indeed normally degraded by NMD (Fig. 4B,C). Two of the candidates, Tra2a and Sf1, showed significantly decreased levels of the NMD isoform upon Rbfox2 knockdown and cycloheximide treatment (Fig. 4B), indicating that Rbfox2 promoted the NMD isoform of these genes, while the NMD isoforms of Snrnp70, Srsf7, Ptbp2, and Tia1 were upregulated upon Rbfox2 knockdown in the absence of NMD (Fig. 4C), suggesting Rbfox2 suppressed the NMD isoform. The responses of Sf1, Srsf7, and Snrnp70 were not specifically predictable due to iCLIP signal both upstream and downstream of the cassette exons. However, the direction of the effect on splicing is that of exon repression by Rbfox2 in all three cases, indicating that binding both upstream and downstream of the exon may lead to repression. As the iCLIP signal was predominantly
downstream of the NMD-SK exon in Ptbp2 and upstream of the NMD-INC exon in Tia1, the observed increases in NMD isoforms by qRT-PCR were consistent with antagonism of the NMD isoform by Rbfox2. Strong binding downstream of the NMD-INC exon in Tra2a, indicating enhancement of the NMD isoform by Rbfox2, was similarly consistent with the relative decrease in the NMD isoform by qRT-PCR upon Rbfox2 knockdown. This argued for Rbfox2-dependent modulation of AS-NMD events within RBPs in a manner predicted by the pattern of Rbfox2 binding. Notably, 50% of the Rbfox2 dependent splicing changes in this set of exons were undetectable in the absence of NMD inhibition, indicating that genome wide, a large number of Rbfox2 dependent splicing events go undetected under standard experimental conditions.

**Autoregulation thresholds are partially set by Rbfox2**

Negative autoregulation feedback loops have been proposed to maintain cellular homeostasis by buffering gene expression against stochastic variation. It has been previously demonstrated that RNA binding proteins are frequently negatively autoregulated by AS-NMD (Lareau et al. 2007; Ni et al. 2007), resulting, for example in the case of SFRS10/Tra2b, in low degrees of cell-to-cell variability of expression (Sigal et al. 2006). The corollary to this model is that another mechanism must exist to allow for changes in expression of these autoregulated genes, for example between cell types. We hypothesized that for a subset of RBPs, the threshold of negative autoregulation, and ultimately the steady-state level of gene expression, is established by cross-regulation of the AS-NMD event by Rbfox2. First, we sought to specifically verify whether the Rbfox2-regulated NMD events were themselves autoregulated. We focused on Ptbp2 and Tia1, both of which had strong motif-dependent FHFOX2 iCLIP signal in well-conserved flanking introns (Fig. 5A) and showed 1.3-fold and 1.9-fold downregulation,
respectively, in the Rbfox2 knockdown RNA-seq. These have been shown or suggested in the literature to be candidates for autoregulation (Le Guiner et al. 2001; Boutz et al. 2007; Wang et al. 2010). Since our previous observations indicated that Rbfox2 inhibited the NMD isoform, it was possible that higher steady-state RBP expression could be achieved in the presence of Rbfox2 due effectively to a higher threshold for autoregulation. To test this model, epitope-tagged Ptbp2 or Tia1 were expressed from an exogenous plasmid in cells stably expressing shLuc or shFox2-1 (Fig.5B). Upon inhibition of NMD by transient knockdown of Upf1, the relative fraction of NMD isoform increased for both Ptbp2 and Tia1 when each of these factors was overexpressed, consistent with negative autoregulation (Fig.5C). The fraction of NMD isoform was further increased upon Rbfox2 knockdown, supporting the hypothesis that Rbfox2 was modulating the autoregulation of these RNA binding proteins. Most strikingly in the case of Tia1, autoregulation by AS-NMD under steady-state conditions was not significant and only became apparent upon Tia1 overexpression and Rbfox2 knockdown. This argues strongly for tight regulation of Tia1 NMD only above a specific threshold of gene expression.

Widespread regulation of AS-NMD by Rbfox2 controls gene expression

Our results suggest that a subset of Rbfox2-dependent binding events is rendered silent through AS-NMD, in particular within autoregulated RBPs. We next tested the prediction that this mode of regulation impacts the steady-state gene expression of these RBPs. The sixty-five Rbfox2-bound AS-NMD splicing events identified above were filtered for those with greater than twofold differential iCLIP density between the 300 nt upstream and downstream of the regulated exon. These were next separated into predicted NMD suppressed or NMD enhanced categories (24 and 20 genes, respectively), and the gene expression upon Rbfox2 knockdown
was plotted for bins of increasing normalized iCLIP density. Consistent with our predictions, RBP genes containing strongly bound NMD suppressed events decreased significantly in expression upon Rbfox2 loss, while genes with strongly bound NMD enhanced events increased significantly (Fig. 6A). These results lend compelling support to the model that Rbfox2 cross-regulation of AS-NMD events within RBPs impacts their gene expression patterns.

To expand our observations beyond RBPs, we next sought to define and characterize Rbfox2-bound AS-NMD events genome-wide. First, to computationally identify known and novel NMD-associated alternative splicing events, RNA-seq reads from control and Rbfox2 knockdown were combined with reads from control and Upf-1 knockdown and control and cycloheximide-treated mESC (data from Hurt et al. 2013) for transcript assembly using Cufflinks (Trapnell et al. 2012). We next determined whether each individual expressed exon, when included or skipped, would result in the production of a predicted NMD isoform. To do this, each internal exon was filtered based on the presence of transcripts in the annotation supporting both the inclusion and skipping event, generating an assembly of transcripts containing cassette exons. Each of these transcripts was then translated in silico, assigning reading frames using Ensembl start codon annotations. If the putative alternative splicing event resulted in a premature termination codon greater than 50 nt upstream of the last exon-exon junction, the event was categorized as either NMD-INC or NMD-SK, depending on whether the NMD isoform resulted from exon inclusion or skipping, respectively. Isoform pairs that were coding and not NMD substrates regardless of whether the alternative exon was included or skipped were categorized as alternative coding sequence, or AS-CDS.

To verify that the algorithm was in fact identifying true AS-NMD events, we measured splicing changes in NMD-INC, NMD-SK, and AS-CDS categories upon Upf1 knockdown using
MISO. $\Delta \psi$ values (shGFP-shUpf1) were significantly negative for the 69 NMD-INC events and significantly positive for the 30 NMD-SK events with a Bayes’ factor $> 3$, supporting the stabilization of AS-NMD isoforms upon inhibition of NMD (Fig.6B). This was further validated at the gene level using RNA-seq by measuring fold change in expression upon Upf1 knockdown. Genes containing NMD splicing events were separated into non-NMD or NMD isoforms, and fold changes in expression were compared with isoform expression from genes containing AS-CDS events. Unstable NMD isoforms were significantly increased upon Upf1 knockdown compared to AS-CDS isoforms (Fig.6C). Stable non-NMD isoforms were significantly decreased; similar observations that an increase in NMD isoforms was coupled with a decrease in non-NMD isoforms have been reported in HeLa cells upon Upf1 knockdown (Pan et al. 2006).

We applied these NMD predictions to the Rbfox2 binding and regulation data to identify evidence of global AS-NMD regulation by Rbfox2. Applying the bucketing approach described previously, FHFOX2 iCLIP signal was significantly enriched in NMD-INC and NMD-SK loci in addition to being enriched in AS-CDS loci (Fig.6D). This argued that Rbfox2 regulation of AS-NMD is likely to be a more widespread phenomenon than would be predicted based solely on transcriptome data from Rbfox2 depletion. As seen previously for all cassette exons, the average conservation of intronic regions flanking AS-CDS, NMD-INC, and NMD-SK exons was significantly higher for those events containing a FHFOX2 iCLIP cluster within 300nt of the exon (Fig.6E). In particular, bound NMD-INC events showed high levels of intronic and exonic conservation; this may be partially attributable to the presence of ultraconserved elements (Fig.6E). The specific evolutionary maintenance of Rbfox2-regulated AS-NMD casettes argues for a widespread functional role of this mode of regulation in development or physiology.
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks

The effect of Rbfox2-dependent splicing regulation on the expression of all genes containing bound AS-NMD events was considered next. As for RBPs, genes with greater than twofold differential iCLIP density around AS-NMD exons were separated into NMD suppressed (124 genes) and NMD enhanced (123 genes) categories based on the location of Rbfox2 binding. In support of pervasive AS-NMD regulation by Rbfox2, NMD suppressed genes with higher iCLIP density tended to decrease in expression upon Rbfox2 knockdown, although not significantly. As predicted, NMD enhanced genes with strong Rbfox2 binding were significantly upregulated upon Rbfox2 knockdown (Fig.6F). This trend persisted when RBPs were removed from these gene sets. Our analysis provides strong evidence that Rbfox2 regulation of AS-NMD impacts steady-state gene expression patterns of RBPs as well as a host of other potentially autoregulated genes in mESC.

2.4 DISCUSSION

Here, we performed a genome-wide analysis of Rbfox2 activity in mouse embryonic stem cells by mapping high-resolution Rbfox2 binding sites to transcriptome changes conferred by Rbfox2 loss. The use of a negative control iCLIP library to filter non-specific signal allowed us to detect high-fidelity binding to the consensus UGCAUG motif. More than 70% of binding sites contained a similar motif, a much larger percentage than was reported in the previous RBFOX2 RNA map from human embryonic stem cells (Yeo et al. 2009). Investigation of a large subset of “silent” Rbfox2 binding events around cassette exons that were not coupled with measurable Rbfox2-dependent splicing regulation revealed an unexpected enrichment in genes for which one splice variant was predicted to be a substrate for alternative splicing-coupled nonsense mediated decay (AS-NMD). Many of the Rbfox2-bound candidate AS-NMD events
Chapter 2. *Rbfox2 controls autoregulation in RNA binding protein networks*

were within transcripts encoding core components of the splicing machinery, alternative splicing regulators, and other RNA binding proteins. We demonstrated that Rbfox2 determines a threshold for the ratio of NMD to non-NMD isoforms for several of these factors. Globally, regulation of AS-NMD by Rbfox2 resulted in changes in gene expression consistent with the polarity of Rbfox2 binding and regulation, thus greatly expanding the breadth of the Rbfox2 regulatory network beyond the previously understood canonical splicing regulation. In fact, this novel general model whereby one splicing regulator controls the auto-regulatory splicing of another protein likely explains how levels of autoregulated proteins vary between cell states.

The observation that hundreds of silent splicing events bound by Rbfox2 are putative AS-NMD cassette exons suggests a functional splicing activity can be attributed to the majority of Rbfox2 binding events when in the correct context, i.e. within 300nt of an exon. This provides an interpretation for the extensive Rbfox2 binding in this region around constitutive exons. Although no evidence of alternative splicing was detected upon Rbfox2 loss or NMD inhibition for the large number of constitutive exons with proximal iCLIP clusters, it is likely that in these cases Rbfox2 actively reinforces constitutive exons with weak splice sites whose missplicing would only be evident upon simultaneous NMD inhibition and Rbfox2 knockdown. Both the observation that Rbfox2 bound constitutive introns have statistically weak splices sites and that Rbfox2-bound loci, as well as the motifs directly bound by Rbfox2, show a high degree of evolutionary conservation supports this conclusion. Conversely, a large number of splicing changes within genes of heterogeneous functional categories observed upon Rbfox2 knockdown cannot be explained by detectable direct binding. As many targets of Rbfox2 are in fact RBPs that change in expression upon Rbfox2 knockdown, we posit that these are the products of
Chapter 2. *Rbfox2 controls autoregulation in RNA binding protein networks*

indirect splicing changes arising from differential expression of components of the splicing machinery.

The NMD pathway is required for organismal development as well as in several tissue- and condition-specific contexts, suggesting a regulatory role extending beyond a general surveillance mechanism (Medghalchi et al. 2001; Weischenfeldt et al. 2008; McIlwain et al. 2010; Weischenfeldt et al. 2012). While it has been demonstrated that transcripts with upstream open reading frames, long 3’ untranslated regions (UTRs), or strong Upf1 binding within the 3’ UTR are targeted by NMD (Mendell et al. 2004; Hurt et al. 2013), the contribution of AS-NMD to the regulation of gene expression has been conflicting. Analysis of AS-NMD splicing upon inhibition of NMD in actively-growing HeLa cells revealed disappointingly few genes with significant increases in expression by microarray (Pan et al. 2006). In contrast, between 15% and 30% of genes within specific mouse tissues were differentially expressed due to loss of AS-NMD upon genetic ablation of Upf2 (Weischenfeldt et al. 2012). Concerted shifts in amino acid metabolism pathway components have been attributed to NMD inhibition induced by amino acid starvation (Mendell et al. 2004). During developmental transitions, gene expression changes of individual factors, such as the PTB/nPTB switch in neuronal development, or entire networks of factors, as with myelocyte-specific proteins in granulopoiesis and regenerative pathways in the adult liver, also depend on splicing-coupled NMD (Boutz et al. 2007; Makeyev et al. 2007; Thoren et al. 2010; Wong et al. 2013). These observations suggest that the scope of gene expression regulation by AS-NMD may be particularly relevant in physiological context such as stress or differentiation. Furthermore, our results imply that the activation of entire AS-NMD networks could be initiated by altering the expression of specific splicing factors within such networks.
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks

AS-NMD is frequently associated with negative autoregulatory feedback loops, which generate tightly-controlled systems of gene expression. Best characterized at the level of transcription, the ability of negative autoregulation to modulate gene expression has been suggested to vary based on the strength of feedback. The paradigm for strong negative feedback is the production of oscillations in transcriptional output (Staiger et al. 2003), while weak or intermediate negative feedback serves to dampen noise in gene expression (Becskei and Serrano 2000; Kaern et al. 2005; Austin et al. 2006). In the case of splicing, deviations in the precise levels of individual components of the splicing machinery could be expected to have far-reaching consequences in maintaining cellular homeostasis. As a result, negative autoregulation by the direct feedback of RBPs on the splicing of their own transcripts could have evolved to safeguard against fluctuations in expression. Several examples of conserved cross-regulation of an autoregulatory AS-NMD event within one RBP by another RBP exist (Le Guiner et al. 2001; Izquierdo and Valcarcel 2007; Spellman et al. 2007; Rossbach et al. 2009; Saltzman et al. 2011; Anko et al. 2012). Our results hint that this may be a more general characteristic of autoregulatory splicing designed to tune gene expression while maintaining the noise buffering inherent to negative autoregulation. We propose that Rbfox2 is capable of functioning in this role in a positive or negative manner, depending on the context. For the set of genes for which Rbfox2 enhances the NMD isoform, as in the case of Tra2a, Rbfox2 potentiates negative autoregulation by the RBP, thus lowering the threshold amount of RBP needed for autoregulation and decreasing its steady-state levels. Here, decreased expression of Rbfox2 results in a concomitant increase in the levels of the coding variant. Conversely, for those genes for which Rbfox2 represses the NMD isoform, such as Ptbp2 or Tia1, Rbfox2 antagonizes the negative autoregulation, and higher expression of the RBP is necessary to compete out Rbfox2...
and trigger splicing of the NMD isoform. Decreased expression of Rbfox2 correlates with decreased expression of the coding variant in this scenario. In this manner, changes in Rbfox2 expression during normal physiological processes would be expected to alter splicing networks not only due to changes in direct targeting by Rbfox2, but also through expression shifts of other splicing regulators. This effect of autoregulation tuning on gene expression is diagrammed in Fig.7.

The tuning of autoregulation is not necessarily limited to RBPs and other splicing factors. While it is simplest to envision direct competition or cooperativity between Rbfox2 and a splicing factor at an autoregulated splicing event, several provocative examples indicate that many AS-NMD cassettes in genes with no reported RNA binding activity may autoregulate through non-canonical mechanisms. In particular, one NMD-INC event predicted to be enhanced by Rbfox2 occurs in the gene Sumo3, a small ubiquitin-related modifier reported to modify hnRNPs, zinc finger proteins, and nuclear pore components (Li et al. 2004) and induce changes in subcellular localization (Pelisch et al. 2010). Sumo modification of RBPs has also been suggested to be relevant during stress responses, such as heat shock (Pelisch et al. 2010). This raises the possibility that downregulation of Sumo3 by Rbfox2 may feed back on its own splicing through decreased sumoylation of hnRNPs. In a similar vein, phosphorylation of SR proteins by the Clk kinases, of which NMD events in Clk1, 2, and 4 are bound by Rbfox2, may lead to SR protein-mediated regulation of Clk RNA splicing accomplished through the kinase activity of the encoded proteins (Colwill et al. 1996; Duncan et al. 1997).

The model we propose has the potential to alter splicing regulation on both short and long timescales. Rbfox2 and other Rbfox family members undergo changes in expression during development, with Rbfox1 expressed in brain and muscle, Rbfox3 restricted to neurons, and
Rbfox2 showing more diverse patterns of expression both in brain and muscle as well as in cells in the mammary ducts and skin fibroblasts (Uhlen et al. 2005; Kuroyanagi 2009; Uhlen et al. 2010). Sustained increases in Rbfox2 protein levels would thus shift expression of the network of Rbfox2 AS-NMD targets by resetting steady-state expression as outlined in Figure 7. However, this may also be regulated transiently, for example in response to extracellular cues or stresses. In neuronal cells, Rbfox1 undergoes a depolarization-induced splicing switch within its own transcript, resulting in an alternate C-terminal reading frame that relocates Rbfox1 from the cytoplasm to the nucleus (Lee et al. 2009). Rbfox2 contains an analogous splicing event, raising the possibility that it could undergo similar signal-dependent relocalization to transiently activate or repress the Rbfox2 splicing network.

Our integrative genome-wide analysis has uncovered a novel network of RNA binding proteins and other autoregulated factors whose expression is modulated through AS-NMD by Rbfox2. Beyond confirming the context- and motif-dependent effects of Rbfox2 binding on splicing regulation, our results bring to light the simultaneous tunability and robustness of splicing regulatory networks arising from the cross-regulation of negative feedback loops. Importantly, this is likely to be generalizable to other splicing regulators in various physiological contexts.

2.5 MATERIALS AND METHODS

iCLIP library preparation

Mouse V6.5 ES cells were cultured in DME/HEPES supplemented with 2mM L-glutamine, 100 Units Penicillin, 100 ug Streptomycin, 1X Non-essential amino-acids (Invitrogen), 15% Defined FBS (HyClone-ES screened), 1000U/ml ESGRO (Chemicon) on gelatinized flasks. Cells were
infected with lentivirus expressing doxycycline-inducible untagged or FLAG-HA-tagged human RBFOX2 (UntagFOX2, FHFOX2) and selected in 150ug/ml hygromycin for 6 days. FHFOX2 and UntagFOX2 cells were induced for 24 hours with 1ug/ml doxycycline and crosslinked at 150mJ/cm², collected, and flash frozen. Cell pellets were resuspended in lysis buffer, partially digested with RNase I, and serially immunoprecipitated with FLAG and HA antibodies. Radiolabeled protein-RNA complexes were resolved on SDS-PAGE gels and transferred to nitrocellulose. RNA corresponding to ~20-100nt fragments crosslinked to Untag/FHFOX2 were eluted by proteinase K digestion, ligated to a 3’ linker, and reverse transcribed. The resulting cDNA was purified on a denaturing polyacrylamide gel, circularized, and amplified for 16 cycles using Illumina-compatible barcoded primers. PCR product was gel purified on a denaturing polyacrylamide gel. UntagFOX2 library was amplified for an additional 7 cycles and re-purified. Samples were pooled and sequenced on the Illumina HiSeq platform.

**iCLIP mapping and clustering**

Adapter-trimmed reads generated with the fastx toolkit were mapped to the UCSC mm9 build using Bowtie1 (Langmead et al. 2009), allowing 2 mismatches and unique alignment. Mapped reads were collapsed on unique 5’ and 3’ ends to correct for overamplification. 5’ terminal nucleotides of reads that were contiguously mapping (no deletions) were classified as truncation crosslink sites. Unmapped reads were uniquely aligned using Bowtie2 (Langmead and Salzberg 2012). Aligned reads with identical 5’ and 3’ ends were collapsed, and single-and multi-nucleotide deletions were extracted. Multi-nucleotide deletions were averaged and rounded down to obtain single nucleotides, and these combined with the single-nucleotide deletions were classified as readthrough crosslink sites. Truncation and readthrough crosslink sites were combined and extended on 12nt in either direction to obtain 25nt crosslink-centered “reads.”
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks

These were used to generate clusters using the Clipper algorithm (available at https://github.com/YeoLab/clipper; Lovci et al. 2013), with pre-mRNA lengths used to determine background read distributions. Resulting clusters were further filtered using a binomial test for FHFOX2 vs UntagFOX2 CLIP signal. Reads in UntagFOX2 were scaled to match the FHFOX2 library size by performing linear regression of crosslink counts within genes. Clusters with FHFOX2 enrichment above a q-value threshold of 0.05 over UntagFOX2 were considered for further analysis.

**Rbfox2 knockdown and RNAseq**

Mouse V6.5 ES cells were infected in biological duplicate with lentivirus expressing control (shLuciferase, shGFP) or Rbfox2 (shFox2-1, shFox2-2) shRNAs. Cells were harvested after 3 days of selection in 1ug/ml puromycin followed by 24 hours in medium containing no puromycin and RNA was collected using Trizol. Poly(A)+ RNA was prepared for sequencing on the Illumina platform.

**RNAseq analysis**

Paired-end reads were aligned to mm9 using Tophat 2.0 (Trapnell et al. 2012). For gene expression analysis, mapped reads were aligned using Cufflinks 2.1.1 (Trapnell et al. 2012) to a custom annotation file generated from annotated and novel junctions called by Tophat from the RNAseq performed in this study as well as that from Hurt et al. 2013. For splicing analysis, reads were aligned to annotated mouse splicing events using the MISO algorithm (Katz et al. 2010). Significant splicing events were determined by requiring a Bayes’ factor > 5 and delta PSI > 0.05 for at least one of the two comparisons of shLuc vs shFox2-1 and shGFP vs shFox2-
2, with the direction of splicing change being consistent in the two comparisons. Each event was required to pass the default MISO minimum read coverage thresholds.

**Metagene analysis**

BedTools (Quinlan and Hall 2010) and custom Python scripts were used to quantify crosslink coverage or conservation with respect to regulated cassette exons. For conservation analysis, mean Phastcons scores across placental mammals were plotted per nucleotide per cassette exon. P-values represent rank sum test per nucleotide between bound (iCLIP cluster within exon or upstream or downstream intron) and expression-matched unbound (no iCLIP cluster within exon or upstream or downstream intron) cassette exons. Average crosslink coverage or motif coverage was computed across a 100nt sliding window, normalized to the total number of exons in each category.

**PTC prediction**

Custom gene annotations were generated using the RNA-seq libraries generated in this study and in (Hurt et al. 2013). Reading frames were annotated by intersecting the custom annotation with Ensembl start codons. Internal exons were categorized as cassette exons based on the presence of annotated transcripts supporting both exon inclusion and exon skipping. Inclusion and exclusion transcript pairs were translated in silico and characterized as always coding (AS-CDS), or NMD upon skipping (NMD-SK) or NMD upon inclusion (NMD-INC) if a premature stop codon was introduced >50nt upstream of the last exon-exon junction. NMD events were validated by further filtering by gene expression > 0.1 FPKM in both control and shUpf1
conditions. MISO annotations were generated for each cassette exon and analyzed using a Bayes’ factor > 3 and default minimum read coverage for either control or Upf1 knockdown.

**siRNA transfection and NMD inhibition**

25nM AllStars Control siRNA 1 ("siCtrl") or siRent1-6 ("siUpf1") (Qiagen) were transfected into mESC expressing shLuc or shFox2-1 48h prior to harvesting, followed by cell expansion and repeat transfection in suspension with LacZ, Ptbp2, or Tia1 expression vectors 24 prior to harvesting protein and RNA.

**AS-NMD gene expression analysis**

Predicted AS-NMD splicing events within RNA binding proteins or within all genes were filtered for those with greater than 10 reads and greater than twofold differential iCLIP crosslink density between the 300nt upstream and downstream of the alternative exon. From this point, only the intron with greater iCLIP density was considered. If the intron with greater density was upstream of an NMD-INC exon or downstream of an NMD-SK exon, the event was categorized as NMD suppressed, while if the intron with greater iCLIP density was downstream of an NMD-INC exon or upstream of an NMD-SK exon, the event was categorized as NMD enhanced. Log2 fold change in gene expression after Rbfox2 knockdown was plotted for bins of increasing FPKM-normalized iCLIP density.

**ACKNOWLEDGEMENTS**

We thank members of the Sharp laboratory for insightful discussion and experimental assistance, in particular Andrew Bosson for help with the iCLIP protocol and analysis. We thank Michael
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks

Lovci and Gene Yeo for assistance with the Clipper algorithm. This work was supported by United States Public Health Service RO1-GM34277 and by an Integrative Cancer Biology Program Grant U54 CA112967 from the National Institutes of Health to PAS and partially by the Koch Institute Support (core) grant P30-CA14051 from the National Cancer Institute. MJ acknowledges support for the year 2011-2012 from the Koch Institute Graduate Fellowship Program. The authors acknowledge the service to the MIT community of the late Sean Collier.

SUPPLEMENTAL INFORMATION includes four supplemental figures detailing Rbfox2 expression system and iCLIP schematic (Supplementary Figure S1), flow chart of iCLIP analysis pipeline (Supplementary Figure S2), genomic distribution of iCLIP crosslink sites and clusters (Supplementary Figure S3), and validation of Rbfox2-dependent cassette exons (Supplementary Figure S4).

2.6 REFERENCES


Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks


Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks


Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks

2.7 FIGURES

Figure 1. FHFOX2 iCLIP is enriched in conserved, alternatively-spliced loci.
Fold enrichment of FHFOX2 crosslinks over UntagFOX2 crosslinks in genomic regions (A) or intronic categories (B) shows statistically significant enrichment within genes and in introns flanking cassette or mutually exclusive events. P-value calculated using Chi-square test. C) Splice sites within 200nt of FHFOX2 iCLIP clusters are significantly weaker than splice sites of unbound introns, both for all introns and for constitutive introns. Splice site strength calculated using MaxEnt algorithm. Per-nucleotide p-value calculated using Wilcoxon test. D) Average Phastcons conservation score across placental mammals around expressed annotated cassette exons shows cassette loci with iCLIP clusters are more highly conserved than expression-matched unbound loci.
Chapter 2. *Rbfox2 controls autoregulation in RNA binding protein networks*
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks

Figure 2. The majority of FHFOX2 iCLIPped regions contain consensus Rbfox2 motifs.
A,D,F) Motif analysis using MEME (Bailey et al. 2009) for the 20 nt flanking the summits of the top 5000 iCLIP clusters shows the Rbfox2 consensus motif, UGCAUG, is the most strongly enriched motif and is present in 71.4% of clusters. Two additional hexamers are significantly enriched and present in a minority of clusters. B,E,G) Total crosslink coverage across intronic occurrences of the top three enriched hexamers conserved across placental mammals (average PhastCons > 0.4) shows nucleotide-level resolution of binding within 300nt (left panels) and 30nt (right panels). Position 0 represents the 5’ most nucleotide of the hexamer. C) Intronic UGCAUG motifs < 100nt from an iCLIP cluster are more highly conserved than unbound motifs. H) Distribution of secondary motif matches around UGCAUG matches within clusters. Matches determined using position-specific probability matrices (pspm) from Figure 3A. Coverage represents the 5’ most nucleotide of each motif occurrence, with position 0 being the first nucleotide of UGCAUG pspm match.
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks

Figure 3. RNA map of Rbfox2 regulation.
Distribution of average FHFOX2 cross link sites mapped by iCLIP (A) or motif frequency (B) around exons enhanced (green) or repressed (red) by Rbfox2 in mESC as determined by RNA-seq upon Rbfox2 knockdown shows that motif-dependent binding downstream correlates strongly with exon activation, while upstream binding correlates with exon repression. C) Average $\Delta \psi$ of cassette exons within each quartile of crosslink density in a 300nt window upstream or downstream of the cassette exon. D) Intersection of Rbfox2-regulated exons defined by MISO (Katz et al. 2010) (|$\Delta \psi$| >= 0.05, Bayes’ factor >= 5) and exons within 300nt of a FHFOX2 iCLIP cluster show a statistically significant overlap. Unregulated, bound exons defined by |$\Delta \psi$| <= 0.02, Bayes’ factor >= 5. Gene ontology analysis using DAVID (Huang et al., 2009a, Huang et al., 2009b) of (E) bound and regulated (corresponding to Venn diagram overlap outlined in green) or (F) bound and unregulated (corresponding to Venn diagram region outlined in yellow) cassette exons over the background of all cassette exons expressed in mESC shows enrichment in RNA binding protein-encoding genes in the bound and unregulated category.
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks
Figure 4. Rbfox2 regulates AS-NMD events in RNA binding proteins
A) Schematic of modes of AS-NMD regulation by Rbfox2. Binding downstream of an NMD-INC exon or upstream of an NMD-SK exon is predicted to result in Rbfox2-dependent NMD (“NMD enhanced”), while binding upstream of an NMD-INC or downstream of an NMD-SK exon will result in more stable transcripts (“NMD suppressed”). Gray regions represent coding sequence, and hexagons represent PTCs. Quantification of percent of transcripts encoding the PTC-containing isoform in Rbfox2 destabilized (B) and stabilized (C) genes after Rbfox2 knockdown and cycloheximide treatment (6h, 10ug/ml) from qRT-PCR using primers spanning exon junctions. Error bars represent standard error across biological triplicates; asterisk represents p < 0.05, paired Student’s t test. Density plots below qRT-PCR plots represent FHFOX2 iCLIP crosslink density +/- 300 nt of NMD-INC (“INC”) or NMD-SK (“SK”) exons.
Figure 5. Rbfox2 inhibits autoregulation of FHFOX2-bound splicing factors.

A) IGV genome browser image of iCLIP signal around an NMD-skipping event in Ptbp2 and an NMD-inclusion event in Tia1. mESC stably expressing shLuc or shFox2-1 were transiently transfected 48h before harvesting with 25nM AllStars control siRNA or siUpf1. 24h prior to harvesting, cells were re-transfected with control or siUpf1 and plasmids expressing LacZ, Ptbp2, or Tia1. B) Western blot showing Upf1 and Rbfox2 knockdown and Ptbp2 and Tia1 exogenous expression. C) Quantification of percent of transcripts encoding the PTC-containing isoform as in Fig.4B. Error bars represent standard error across biological triplicates; p-values determined using paired Student’s t test.
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks
Figure 6. Rbfox2 regulation of AS-NMD has widespread effects on gene expression

A) Gene expression changes upon Rbfox2 knockdown of RBPs with increasing amounts of iCLIP density around AS-NMD exons shows NMD suppressed genes decrease significantly in Rbfox2 knockdown, while NMD enhanced genes increase significantly. Significance calculated using one-sided one sample Student’s t test (*p<0.05). B) Predicted NMD-INC transcripts show increased exon inclusion upon Upf1 knockdown, while predicted NMD-SK events show increased skipping, consistent with expectations. Splicing changes calculated using MISO. P-values represent one-tailed one sample Student’s t test. C) Cdf plot of gene expression changes for each class of isoform in Upf1 knockdown vs. control. Predicted PTC-containing isoforms are...
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks

significantly upregulated upon Upf1 knockdown, while non-NMD isoforms are significantly downregulated. D) FHFOX2 iCLIP crosslinks are enriched in introns flanking NMD-INC and NMD-SK splicing events over UntagFOX2 crosslinks. Significance calculated using Chi-square test. E) Average Phastcons conservation score across placental mammals around expressed NMD-INC and NMD-SK cassette exons shows loci with iCLIP clusters are significantly more highly conserved than unbound loci. F) As in 6A, but using all predicted AS-NMD events with greater than twofold differential binding between upstream and downstream introns. NMD enhanced genes increase significantly in expression upon Rbfox2 knockdown.
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks

Figure 7. Rbfox2 regulates silent NMD events to tune RNA binding protein networks
A) Schematic Venn diagram of Rbfox2-regulated and Rbfox2-bound cassette exons shows how the direct regulation of NMD splicing events within RNA binding proteins by Rbfox2 can mediate the indirect splicing changes seen upon Rbfox2 loss. B) Model of Rbfox2-mediated tuning of RNA binding protein expression patterns. As Rbfox2 expression increases (blue line), steady-state expression of RNA binding proteins (RBPs) for which Rbfox2 suppresses NMD also increases (green line). Simultaneously, steady-state expression of RBPs for which Rbfox2 enhances NMD decreases (red line). This results in a shift from Splicing network A to Splicing network B, defined by both direct and indirect targets of Rbfox2.
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks
Supplemental Figure S1. Manipulation of Rbfox2 expression for RNA-seq and iCLIP.

A) Doxycycline-inducible FLAG-HA tagged (FHFOX2) or untagged (UntagFOX2) were stably transduced into mESC. Cells were infected with shRNA targeting GFP or endogenous Rbfox2. Titration of doxycycline results in a range of Rbfox2 expression from ~0 to 2 fold. B) Rbfox2-dependent splicing events show a dose-dependent change in splicing by RT-PCR using primers placed in constitutive exons flanking Rbfox2-regulated cassette exons. C) Schematic of iCLIP cloning and analysis protocol used to generate FHFOX2/UntagFOX2 iCLIP libraries (modified from Sugimoto et al. 2012). D) SDS-PAGE gel of RNA-protein complexes after limited RNase I and complete RNase A digestion. Samples in lanes 1, 2, 4, and 5 are 32P-labeled samples, while lanes 3 and 6 are not 32P-labeled. Red boxes indicate fragment sizes selected for cloning.
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks

Supplemental Figure S2. Schematic for iCLIP mapping and clustering.
A) Flow diagram of analysis performed to generate iCLIP crosslink sites and clusters. Numbers of reads corresponding to each category are indicated in millions (m) or thousands (k). B) Sequenced read length for reads mapping with deletions (top panel) or contiguously (bottom panel) demonstrate that contiguously-mapping reads are shorter on average and are most likely due to crosslink-induced inhibition of reverse transcription.
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks
Supplemental Figure S3. Genomic distribution of FHFOX2 iCLIP crosslinks and clusters.
Distributions of crosslink sites and clusters across genomic regions (A) and intronic categories (B). Percentages reflect unnormalized proportions of crosslink sites or clusters in each bin, regardless of genomic lengths covered by each category.
Chapter 2. Rbfox2 controls autoregulation in RNA binding protein networks
Supplemental Figure S4. Identification of cassette exons regulated by Rbfox2.
A) Plot of delta PSI of each cassette exon between shLuc/shFox2-1 (y-axis) and shGFP/shFox2-2 (x-axis). Concordant splicing events were defined as those events showing a significant splicing change in one set of comparisons (shLuc vs shFox2-1 or shGFP vs shFox2-2) and changing in the same direction in the other set (green-bordered quadrants). B) Validation of candidate Rbfox2-regulated cassette exons by semiquantitative RT-PCR shows strong correlation with splicing changes estimated by MISO. Error bars represent standard deviation of
biological triplicates. C) Table of significant splicing changes upon Rbfox2 knockdown for different categories of splicing event.
Chapter 3

Quantitative analysis of protein-RNA binding reveals novel functional motifs and impact of RNA structure

The work in this chapter is presented in the context of an ongoing collaboration with the laboratory of Christopher Burge and is adapted with permission from the following manuscript:


*Contributed equally

Experimental contributions:

Nicole Lambert optimized and generated RBNS libraries and performed computational analyses. Alex Robertson performed computational analyses, with input from Mohini Jangi. Mohini Jangi performed Rbfox2 titration experiment and Rbfox2 iCLIP. Sean McGeary performed RBNS modeling. Nicole Lambert, Alex Robertson, and Christopher Burge wrote the manuscript, with input from Mohini Jangi and Phillip Sharp.
Chapter 3. Comprehensive, quantitative protein-RNA binding

3.1 ABSTRACT

Specific interactions between protein and RNA guide most post-transcriptional gene regulation. Here, we describe RNA Bind-n-Seq (RBNS), which comprehensively characterizes the sequence and structural specificity of RNA binding proteins (RBPs), and its application to the developmentally-regulated splicing factors RBFOX2, MBNL1 and CELF1/CUGBP1. For each factor, we recovered canonical motifs and additional near-optimal binding motifs. We found that RNA secondary structure inhibits binding of RBFOX2 and CELF1, while MBNL1 favors unpaired Us but tolerates C/G pairing in UGC-containing motifs. Relative dissociation constants calculated from RBNS data using a novel algorithm correlated highly with surface plasmon resonance $K_d$ values. Motifs identified by RBNS were almost invariably associated with in vivo binding, sequence conservation and with splicing regulatory activity. However, a subset of motifs identified by CLIP-Seq but not RBNS lacked detectable splicing regulatory activity or conservation. Together our data suggest that the binding and activity of these splicing factors in vivo is driven largely by their intrinsic RNA affinities and that RBNS clarifies interpretation of CLIP-Seq data.
3.2 INTRODUCTION

RBPs bind sequence and/or structural motifs in pre-mRNAs to direct their processing, and in mature mRNAs to control their translation, localization, and stability. Proteins of the RBFOX, CUG-BP, Elav-like (CELF) and muscleblind (MBNL) families are among the most important and highly conserved RBPs that regulate developmental and tissue-specific alternative splicing. These factors also play additional regulatory roles, with MBNL proteins contributing to mRNA localization (Adereth et al. 2005; Wang et al. 2012) and CELF proteins targeting mRNAs for destabilization (Moraes et al. 2006; Vlasova et al. 2008).

RBFOX2, a close homolog of RBFOX1 (Underwood et al. 2005), is required for neural development (Gehman et al. 2012), regulates epithelial-mesenchymal transition (EMT) (Baraniak et al. 2006), and is required for human embryonic stem cell (ESC) survival (Yeo et al. 2009). The consensus binding motif for RBFOX proteins – UGCAUG or simply GCAUG – has been determined by systematic evolution of ligands by exponential enrichment (SELEX) and is conserved from nematodes through vertebrates (Jin et al. 2003; Ponthier et al. 2006). However, the iterative selection steps used in SELEX favor recovery of just the strongest binding motifs and may not detect moderate and lower affinity motifs. About one-third to one-half of in vivo identified RBFOX2 binding sites contain these canonical motifs (Yeo et al. 2009; Jangi et al. in review), but it has remained unclear whether RBFOX2 can recognize other sequence motifs. In general, motifs recognized by RBPs with lower affinity are more challenging to characterize, but such motifs may play biological roles that are as important as those of higher affinity. For RBPs that accumulate during development (like MBNLs), higher affinity motifs may be bound at earlier time points, while lower affinity motifs may specify regulation at later developmental time points or only in specific cell types where the RBP accumulates to high levels. By
extension, we might expect that disease states that result in perturbation but not complete loss of RBP activity may preferentially alter targets with moderate affinities that may be more sensitive to RBP levels, though this has not yet been clearly demonstrated.

CELF1 and MBNL1 proteins are functionally linked by their roles in development and disease, often regulating the same splicing targets in an antagonistic fashion. In heart development, during which CELF protein levels decrease and MBNL proteins increase over time, this antagonism may help to confer sharper developmental transitions (Kalsotra et al. 2008). This heart development expression pattern is the reverse of that seen in the muscle wasting disease myotonic dystrophy type 1 (DM1), in which expanded CUG repeats in the 3' UTR of DMPK mRNAs reduce available cellular levels of MBNL1 by sequestration (Taneja et al. 1995; Mankodi et al. 2005), and CELF1 is stabilized by hyperphosphorylation (Kuyumcu-Martinez et al. 2007). CELF1 has three RNA recognition motifs (RRMs) that bind repeats of UGU (Ladd et al. 2001; Marquis et al. 2006). MBNL1 has two pairs of zinc fingers that bind preferentially to YGCY (Y = C or U) motifs (Ho et al. 2004). To date, it has remained unclear whether MBNL1 primarily recognizes secondary structural motifs or single stranded RNA elements. CUG repeat RNA crystallizes as an A-form helix (Mooers et al. 2005b) with C and G bases paired and Us unpaired, and additional biochemical studies have shown that a mismatched RNA hairpin structure is important for recognition by MBNL1 (Warf and Berglund 2007). However, structures of MBNL1 co-crystallized with CGCUGU RNA suggest that MBNL1 recognizes single-stranded RNA (Teplova and Patel 2008b). MBNL and CELF each bind strongly to sequences containing repeats of their preferred motifs, but the presence of multiple RNA binding domains, each potentially capable of binding individual motifs, complicates the identification of the minimal requirements for binding. Additionally, the context between
tandem motifs, including degeneracy of any intervening base(s) and motif spacing, remains largely uncharacterized.

Widely used methods for mapping protein-RNA interactions \textit{in vivo} based on ultraviolet cross-linking and immunoprecipitation (CLIP) (Ule et al. 2003; Underwood et al. 2005) have generated novel insights into mechanisms of post-transcriptional regulation. These techniques, however, are laborious and require many selection steps that likely introduce various types of bias. Motif analysis from CLIP data is complicated by the fact that it does not distinguish binding directed by a single protein from that originating from a protein complex, and it may preferentially detect uridine-rich sequences (Sugimoto et al. 2012b). Existing \textit{in vitro} methods for measuring RNA-protein interactions have limitations in the depth and quantitative capacity of motif discovery. Iterative binding approaches like SELEX can identify RNA motifs bound with high affinity in vitro, but are not quantitative and may miss lower affinity motifs. A newer method, RNAcompete, uses \textit{in vitro} RNA-protein binding followed by microarray analysis, allowing the high-throughput identification of RNA binding motifs (Ray et al. 2009; Ray et al. 2013). However, RNAcompete is constrained by the number of probes assayed, limiting the information that can be obtained about RNA secondary structure or other contextual effects on RNA binding, and does not yield $K_d$ values. Quantitative biophysical measurements can be obtained from methods such as electrophoretic mobility shift assays (EMSA) or surface plasmon resonance (SPR), but the throughput of these techniques is quite low.

To better characterize the functions of RBPs with important roles in development, we sought to develop a method that would measure affinities to the full spectrum of bound RNAs in a quantitative and high-throughput manner. Methods for characterizing protein/DNA interactions that are both high-throughput and quantitative have been developed, including HT-
SELEX and Bind-n-Seq, both of which use one-step binding to a pool of randomized DNA in vitro followed by deep sequencing (Jolma et al.; Zykovich et al. 2009), and HiTS-FLIP, which directly measures protein bound to dsDNA on a flow cell (Nutiu et al. 2011). We adapted the general approach used by HT-SELEX and Bind-n-Seq to the study of protein-RNA interactions in vitro in a method we call RNA Bind-n-Seq (RBNS). Our method extends the protein/DNA interaction assays in two important ways. First, we use multiple RBP concentrations to optimize analysis of different ranges of affinity. Second, we have expanded the analytical framework to more accurately estimate relative dissociation constants, and to assess the effects of RNA secondary structure on binding. RBNS analyses of RBFOX2, CELF1 and MBNL1 yielded comprehensive portraits of the sequence and RNA secondary structural determinants of RNA recognition by these factors. Comparison of RBNS data with iCLIP/CLIP data for these factors shows that in vitro binding affinities largely drive binding in vivo. Analysis of data from systems in which these RBPs are depleted or inducibly over-expressed in mouse cells or tissues provides evidence of function for both non-canonical and canonical binding motifs identified in vitro. We also find that some motifs enriched by CLIP but not by RBNS are not associated with regulatory activity. Therefore, RBNS provides information that is complementary to CLIP and can aid in the identification of high-confidence functional binding sites.

3.3 RESULTS

Design considerations for RNA Bind-n-Seq experiments

RNA Bind-n-Seq is a technique designed to dissect the sequence and structural preferences of RBPs. A recombinantly expressed and purified RBP is incubated with a pool of randomized RNAs of a given size $\lambda$ at several different protein concentrations, typically ranging from low
nanomolar to low micromolar (Figure 1A). The RNA pool typically consists of random RNAs of length $\lambda = 40$ nt flanked by short primers used to add the adapters used for deep sequencing. This RNA pool design simplifies library preparation, avoids biases that can result from RNA ligation, and ensures that any bacterial RNA carried over from protein expression will not contaminate the sequenced library. In the rare case where the RBP has significant affinity to the primers, different primer sequences must be substituted. In each experiment, the RBP is captured via a streptavidin binding peptide (SBP) tag. RBP-bound RNA is reverse-transcribed into cDNA and multiplex sequencing adapters are added by PCR to produce libraries for deep sequencing. Libraries corresponding to the input RNA pool and to 7-10 different RBP concentrations (including zero RBP concentration as an additional control), are sequenced in a single Illumina HiSeq 2000 lane, typically yielding 15-20 million reads (or more) per library.

Most RBPs bind single-stranded RNA sequence motifs 3-8 bases in length (Stefl et al. 2005). Here, we performed one experiment using the RBFOX2 RRM with short oligonucleotides ($\lambda = 10$ nt). However, we soon realized that use of longer sequences ($\lambda = 40$ nt) provided comparable affinity measures to short linear motifs of size $k$ (kmers) in the range of interest (about 3 to 10 nt, Fig. S1) while also enabling assessment of contextual effects such as RNA secondary structure on binding that cannot be assessed using 10mers (see below). Use of $\lambda = 40$ nt is closer to the in vivo situation where RBPs typically bind long RNAs, while remaining in the length range where structure can be most accurately predicted by thermodynamic RNA folding algorithms (Hofacker 2003).

**RNA Bind-n-Seq comprehensively identifies known and novel motifs of RBPs**
RBNS was performed using recombinant RBFOX2, MBNL1 and CELF1 proteins incubated with randomized RNA 40mers flanked by short primers (Methods). For each protein, at each of several concentrations, motif read enrichment (R) values were calculated for each kmer as the ratio of the frequency of the kmer in the selected pool to the frequency in the input RNA library. In the zero concentration experiment for RBFOX2, 99.9% of 6mers had R values less than 1.19, and the highest value was 1.21, indicating little if any sequence bias from the apparatus, and the other zero concentration experiments yielded similar results.

For RBFOX2, at all concentrations ≥14 nM the 6mer UGCAUG had the highest R value (Figure 1B and below), confirming this well known motif as the highest affinity 6mer (Auweter et al. 2006; Ponthier et al. 2006). The enrichment of UGCAUG reached a maximum R of 22 at a protein concentration of 365 nM (Figure 1B). We derived an equation relating the observed R value to the relative affinity (ratio of dissociation constants) between nonspecific and specific binding under idealized conditions (Robertson 2013). With R = 22, k = 6 and λ = 40, this equation implies at least ~900-fold higher binding affinity to UGCAUG than to nonspecific 6mers. In total, forty-two 6mers had R values at least three standard deviations above the mean, including all 8 of the 6mers that contain GCAUG (Figure 1B), consistent with the known affinity of RBFOX proteins for this 5mer (Jin et al. 2003). Several 6mers containing GCACG were also highly enriched above background, from which we infer that this 5mer represents an alternate RBFOX2 binding motif. Extensive analysis of the in vivo binding and activity of this and other motifs are presented below. Certain other 6mers not containing GCAUG or GCACG, but often containing GCAU, also had significant R values, suggesting that RBFOX2 has some affinity for other RNA motifs as well.
Proteins of the CELF family are known to preferentially bind to UG- and UGU-containing motifs (Timchenko et al. 1996; Marquis et al. 2006). For CELF1, a large number of 6mer and 7mer motifs had significant R values (7mer analysis shown in Figure 1C). Inspection of these motifs showed that the highest R values were observed for 7mers containing two UGU triplets. In fact, all 7mers containing two UGUs were significantly enriched, suggesting that presence of two UGUs is sufficient for strong binding and that CELF1 tolerates presence or absence of a 1 nt spacer between UGUs (Figure 1C). The highest R value observed for any 7mer in the CELF1 analysis, R = ~8 for UGUUUGU implies > ~250-fold binding affinity over background (Robertson 2013), somewhat below the affinity observed for RBFOX2 relative to its top motif. This observation and the fatter tail of the CELF1 R value distribution emphasize that this factor binds a broader spectrum of motifs with lower specificity than RBFOX2. Of the top fifty 7mers, all contained at least one UGU. However, not every motif containing a single UGU was significant, and some 7mers lacking UGU were significantly enriched, indicating that RNA recognition by CELF1 is more complex. Inspection of the top 50 CELF1 7mers suggested that they can be clustered into ~4 groups depending on the spacing of UGU and GU motifs, with classes matching GU_NGU for x = 1,2,3 and a fourth class matching UDUGU (D = G, A or U). Each class had a different preferences for U (or occasionally G or A) at the remaining positions (Figure 1E). This representation emphasizes the complexity of CELF1’s recognition of RNA, which likely derives from the presence of multiple RRM domains.

MBNL1 is known to favor binding to YGCY motifs both in vitro (by SELEX) and in vivo (by CLIP-Seq) (Goers et al. 2010b; Wang et al. 2012). The most enriched 7mers for MBNL1 contained either the 4mer YGCU or GCUU, often supplemented by a second GC, and the 7mer with highest R value, GCUUGCU, contained both of these 4mers (Figure 1D), and had
an R value near 9, slightly above that of CELF1’s top motif. Overall, 54% of 7mers containing YGCU, 61% of those containing GCUU, and only 9% of those containing YGCC had significant R values, suggesting that MBNL’s specificity is best described as YGCU + GCUU with some affinity for other motifs, rather than by alternative descriptions such as YGCU. Focusing on the core GC dinucleotide, MBNL1 prefers 7mers containing two GC 2mers, with a slight preference for spacing of 2 or 3 Us (Figure 1F). MBNL1’s observed preference for multiple GC’s separated by variable spacing is consistent with previous studies (Cass et al. 2011). Similar to our observations for CELF1, 7mers bound by MBNL1 could be clustered into 4 groups depending on the spacing of GC motifs, with classes matching GCN_xGC for x = 1,2,3 and a fourth class matching YGCU, with each class having different preferences for U (or occasionally C) at the remaining positions (Figure 1F).

**Relative dissociation constants are accurately estimated from RBNS**

To better understand the dependence of R values on RBP concentration and to assess the extent and effects of experimental noise, we modeled Bind-n-Seq experiments and predicted the output under various assumptions. In an idealized setting in which an RBP binds a high affinity motif X with $K_d = 5 \text{ nM}$ and several moderate affinity motifs Y each with $K_d = 30 \text{ nM}$ (assuming 1:1 stoichiometry and a Hill coefficient of 1), the fraction of each motif bound is expected to follow essentially a sigmoidal function of RBP concentration, with 50% binding occurring at a concentration near the $K_d$ value of each motif (Figure 2A). From the predicted binding fraction, assuming complete recovery of protein, the expected R value at each concentration can be determined under various assumptions about the affinity of the protein for non-specific RNA and
the amount of non-specific RNA bound to the apparatus (e.g., the beads), both of which reduce R values.

The modeled enrichment profiles (Figure 2B) show that at higher RBP concentrations, R values of high affinity motifs decrease under all conditions tested. This effect is readily understood by considering that high RBP concentrations will tend to drive binding of lower affinity RNAs (and high affinity motifs may be saturated), resulting in a lower fraction of RNA-bound RBP associated with high affinity motifs. These simulations also show that even a small amount of nonspecific binding to the apparatus greatly reduces R values at very low RBP concentrations because nonspecifically-recovered RNA dilutes the small amount of specifically-recovered RNA. Together, these two effects produce a characteristic unimodal curve that peaks at intermediate RBP concentrations under most reasonable assumptions. Presence of secondary motifs of lower binding affinity also reduces the fraction of reads containing the primary motif at higher RBP concentrations (Figure 2C).

Experimental enrichment profiles for highly enriched kmers were all unimodal as a function of protein concentration for RBFOX2, CELF1 and MBNL1, in general agreement with our model under the assumption of moderate levels of nonspecific background (Figure 2D). In all cases, R values near 1 were observed at RBP concentrations of 0 nM and began to climb above 1 in the low (4-40) nanomolar range, decreasing to 1 at the highest (micromolar) protein concentrations. For each factor, the relative rankings of kmers obtained at different protein concentrations were highly correlated, supporting the assay’s robustness (Robertson 2013).

In order to assess the extent to which quantitative binding constants could be inferred using our approach, we estimated $K_d$ values from RBNS data. To estimate binding constants, the initial quantity of each kmer present was estimated as the frequency (per oligo) of the kmer in
sequence reads obtained from the input library multiplied by the total concentration of RNA oligonucleotides (1 µM). The concentration of kmer in complex with RBP was then calculated from the total concentration of RBP-RNA complex as quantified by Bioanalyzer analysis (Methods). The fraction of bound RNA attributable to binding at each specific kmer was then estimated using a novel “streaming kmer assignment” (SKA) algorithm (Methods), which generalizes the analytical approach described in Robertson 2013 in that it can account for arbitrarily complex combinations of affinities to different kmers. The SKA algorithm works by processing sequence reads sequentially and assigning the binding to specific kmers in the sequence probabilistically, based on continually updated estimates of relative binding preferences in multiple passes through the sequence read data (details provided in Methods). The algorithm is formally analogous to the streaming assignment of ambiguously mapping sequence reads to a genome by the recently described eXpress algorithm (Roberts and Pachter 2013). Using simulated read data, we have found that SKA more accurately assigns binding locations within reads than do raw R values or B values inferred using a simplified model of RBNS (Robertson 2013).

SKA is particularly helpful in distinguishing bound motifs from those that are enriched merely through frequent overlap with bound motifs. For example, binding of RBFOX2 to GCAUG motifs will cause overlapping motifs of the form CAUGN (N = A, C, G or T) to be enriched in bound reads even if these sequences have no intrinsic affinity to RBFOX except when preceded by a G. Given data in which presence of an authentic bound motif gives rise to enrichment of overlapping unbound motifs, the SKA algorithm uses the greater enrichment of the bound motif to assign binding preferentially to this motif. In successive iterations, SKA assigns higher binding probability to the more enriched kmers and lower probability (usually
near background levels) to overlapping motifs, effectively “learning” to assign binding to the bound motif rather than overlapping motifs when the enrichments can be explained by a single motif.

The concentration of each unbound kmer in the binding reactions is estimated from the difference between the total concentration of that kmer and the estimated concentration that is in complex. Using these estimates of bound and free kmer concentrations, we then calculated “relative $K_d$ values” to kmers, defined as the ratio of a motif’s absolute dissociation constant to that of the highest affinity kmer (Methods). We emphasize relative rather than absolute $K_d$ values here to avoid consideration of technical factors that may systematically affect absolute $K_d$s and may vary between experiments (e.g., proportion of properly folded protein). The kmers for which SKA predicts binding (those with absolute $K_d < \sim 2000$ nM) have relative $K_d$ estimates spanning several orders of magnitude and are highly correlated to SPR measurements ($r = 0.94$, $P < 0.001$) (Figure 2E). Similarly high correlations were observed relative to previously measured SPR data for RBFOX1, a close paralog of RBFOX2 with identical RNA binding domain (Figure S2). Together, these observations demonstrate that RBNS yields quantitative measures of protein-RNA affinity.

**Secondary structure inhibits binding of RBFOX and CELF proteins to RNA**

In addition to characterizing the sequence preferences of RBPs, RBNS can detect effects of RNA structure on binding. We applied the thermodynamically-based Vienna RNAfold algorithm (Hofacker 2003) to sequence reads to assess the contribution of RNA structure to RBP:RNA interactions. In a motif-centric analysis, we analyzed folding of all RNAs harboring high affinity UGCAUG, UGUUU, or UGCUGC motifs in RBFOX2, CELF1 or MBNL1 RBNS
datasets, respectively (as well as other motifs), and in control libraries. The probability of intramolecular base pairing at each base in the motif was calculated from the energy-weighted ensemble of structures and averaged across the bases in the motif. Sequences were then binned by this “average base-pairing probability” (ABP), and R values were calculated separately for each combination of motif, protein concentration and ABP bin. In these analyses, the bin with lowest ABP (0.0 – 0.2) was invariably the most enriched for both RBFOX2 and CELF1 at all non-zero RBP concentrations (Figure 3A). As the ABP increased, R values decreased. This decrease in R values appeared somewhat less pronounced at the highest RBP concentrations, where the increased RBP levels may more effectively compete with intramolecular interactions (which are expected to be independent of RBP concentration). Similar results were obtained when analyzing other motifs for these two factors. Together, these data suggest that RBFOX2 and CELF1 preferentially recognize single-stranded RNA motifs and that intramolecular base-pairing directly competes with RBP recognition of these motifs (Auweter et al. 2006; Edwards et al. 2013).

**MBNL1 binding tolerates pairing of GCs but favors unpaired Us**

The RNA structure analysis for MBNL1 yielded a different pattern, with the highest R values observed for motifs with moderate ABP (in the range 0.2 – 0.6). To determine the impacts of base-specific RNA structure on MBNL1 binding, the base-pairing probability was calculated for each individual base in protein-bound sequences containing UGCUGC, and normalized to the ABP of UGCUGC-containing RNAs in the input library, matched for C+G% content (Methods). This analysis showed no preference for lower base-pairing probabilities at GC positions, but showed substantially reduced base-pairing of Us in bound sequences (Figure
A similar tolerance for pairing of the central GC dinucleotide and preference for unstructured flanking pyrimidines was observed for all high affinity MBNL1 motifs tested, including UGCUU, GCUUGC, CGCUU and GCUGCU. This apparent tolerance of GC pairing does not result simply from MBNL1’s preference for binding multiple GpC dinucleotides in tandem (which might base pair with one another), as similar structural preferences remained when GpC content was controlled for. Similar RNA folding analyses of data for RBFOX2 and CELF1 showed a relatively uniform preference for absence of structure at every position across the binding motif, again consistent with binding to single-stranded RNA (Figure 3B).

**MBNL motifs adjacent to ancient alternative exons have unpaired Us**

In a recent comparative study, we classified conserved exons by their pattern of alternative or constitutive splicing across four mammals and one bird and observed that introns adjacent to exons alternatively spliced in all of the studied mammals (“ancient alternative exons”) are enriched for certain motifs, including those associated with the MBNL and RBFOX families of splicing factors (Merkin et al. 2012). Curiously, we found that MBNL binding to these exons (as measured by CLIP-Seq) far exceeded that expected based on motif enrichment, suggesting that these exons possess additional contextual features that favor MBNL binding. To ask whether the RNA structural preferences observed above might explain this preferential binding, we performed RNA folding analysis of these ancient alternative exons, and of more lineage restricted alternative and constitutive exons. We observed that Us occurring in MBNL motifs such as GCUU which occur in ancient alternative exons have lower base-pairing probability than similar motifs occurring in constitutive exons or more lineage-restricted alternative exons (Figure 3C). These observations suggest that these exons have evolved
structures in which MBNL motifs have unpaired Us in order to facilitate binding of MBNL proteins.

Motifs identified in vitro are almost invariably bound in vivo

RBNS data resolve a spectrum of high, moderate and low affinity motifs that are bound in vitro. To assess the extent to which these affinities guide binding in vivo, we compared to CLIP-Seq data for each factor. A modified version of the high-resolution iCLIP procedure (Konig et al. 2010) was performed using tagged RBFOX2 in mouse embryonic stem cells (mESCs) in a study exploring RBFOX2’s role in the regulation of splicing factor networks via nonsense mediated decay (Jangi et al. in review). These data enabled mapping of precise sites of crosslinking in the transcriptome at nucleotide resolution (Methods).

Sites of crosslinking corresponded in many cases to canonical UGCAUG motifs or to the alternate motif, GCACG, identified above. For example, an iCLIP cluster overlapping a GCACG motif was observed in intron 2 of Dyrk1a (Figure 4A). To systematically assess RBFOX’s in vivo binding specificity, the number of crosslinking sites overlapping occurrences of UGCAUG and other motifs in introns and 3’ UTRs were compiled and visualized in a meta-motif representation (Figure 4B). Sharp peaks of crosslinking density directly over UGCAUG sites were present in both introns and 3’ UTRs, illustrating the high specificity of RBFOX2 binding and the high precision of the iCLIP method (Figure 4B; upper). The doublet shape of the peak appears to result from preferential crosslinking to the two Us in the motif. We also observed distinct peaks of crosslink density overlapping occurrences of the alternate motif, GCACG, in both introns and 3’ UTRs (Figure 4B; middle). These peaks were apparent despite the somewhat higher “noise” (background fluctuations) in these plots that results from the lower
abundance of GCACG sites in the transcriptome. This lower abundance is likely caused by the presence of a (mutation-prone) CpG dinucleotide in the motif. These peaks were RBFOX2-specific: CLIP-Seq data from an unrelated RBP showed no significant enrichment near canonical RBFOX2 motifs (Figure 4B, bottom).

Similar analyses of MBNL1 motifs using MBNL1 CLIP-Seq data from our previously published study using mouse C2C12 myoblasts (Wang et al. 2012) yielded a pronounced peak over MBNL motifs such as GCUUGC in introns and 3' UTRs (Figure 4C; upper). Analysis of CELF1 CLIP-Seq data from a study of this factor’s role in splicing and mRNA stability, also using mouse myoblasts (Wang et al., in preparation) also yielded a pronounced peak in the vicinity of canonical CELF motifs such as UGUUGU (Figure 4C; lower). The peaks observed in the MBNL1 and CELF1 CLIP data were not as sharp as those observed for RBFOX2, likely reflecting the lower resolution of the standard CLIP-Seq protocol used for these factors relative to iCLIP. Again, these peaks were RBP-specific.

We next compared in vitro and in vivo binding across a broader spectrum of motifs. For this purpose, we defined a CLIP “signal:background” (S/B) ratio for each motif as the total CLIP-Seq read coverage overlapping occurrences of the motif (“signal”) divided by the average of the CLIP coverage in 40 nt regions located 40 bases upstream and 40 bases downstream of the motif, representing the “background” level of CLIP coverage in motif-containing transcripts. Comparing CLIP S/B values to RBFOX2 RBNS R values across all 6mers, we observed a strong correlation of these values for the set of motifs with significant R values, but not for other 6mers (Figure 4D; left). In fact, virtually every motif with significant R value had a CLIP-Seq S/B > 1, including not only all 6mers containing the canonical 5mer GCAUG but also all of those containing the alternate 5mer GCACG, and most other 6mers with significant R values. Similar
correlations were observed in the data for CELF1 and MBNL1 (intronic sites in Figure 4D; 3' UTR sites in Figure S3). These observations suggest that the intrinsic binding preferences identified by RBNS determine in vivo binding locations of these proteins to a surprisingly large extent. They also suggest that RBNS has a very low false negative rate in that virtually every motif identified as significantly bound in vitro was also preferentially bound in vivo. However, this relationship was not reciprocal: many motifs with high CLIP S/B were bound in vitro, but many others lacked significant in vitro binding, a phenomenon that we explore below.

Alternate and canonical motifs are associated with alternative splicing regulation

Binding of RBFOX2, CELF1 and MBNL1 proteins to specific locations near alternative exons is frequently associated with splicing regulation (Yeo et al. 2009; Orengo et al. 2011; Wang et al. 2012). To explore the splicing regulatory activity of the RBFOX2 motifs identified by RBNS, mESCs with a range of RBFOX2 expression levels were generated. Over-expression of RBFOX2 to different extents was achieved by administration of various concentrations of doxycycline to a mESC line containing a tetracycline inducible version of RBFOX2 (Jangi et al. in review). Inhibition of RBFOX2 expression was achieved by stably introducing vectors expressing short hairpin RNAs (shRNAs) targeting the 3' UTR of the endogenous gene (or shRNAs targeting GFP as a control). RNA-Seq analysis of cell lines expressing 8 different levels of RBFOX2 was then performed to assess changes in alternative splicing.

Expression of RBFOX2 increased from an FPKM of 12 in the lowest RBFOX2 condition (shRBFOX2, 0 µg/mL DOX) to 32 at the highest induced level (shGFP, 1 µg/mL Dox), a ~3-fold dynamic range. Western analysis confirmed knockdown of endogenous RBFOX2 in shRBFOX2 conditions and monotonically increasing levels of exogenous RBFOX2 protein with increasing
doxycycline concentrations (Figure 5A). The percent spliced in (PSI) values of RBFOX2-sensitive regulated alternative exons are expected to consistently increase or consistently decrease as RBFOX2 expression is increased. To systematically address the consistency of changes in splicing, we defined a “monotonicity Z-score” (MZ) for each exon whose PSI value changed significantly. MZ captures the extent to which the exon’s PSI consistently increases (MZ > 0) or consistently decreases (MZ < 0) in a set of conditions with increasing levels of a regulatory factor (Methods).

Applying this approach to a set of mouse alternative exons, the exons with the highest MZ scores were exon 9 of the UAP1 gene (MZ = 2.98) and the EIIIB exon of Fibronectin1 (MZ = 2.81). The latter is a well-established RBFOX2 target whose downstream intron contains six canonical UGCAUG motifs (Huh and Hynes 1993; Lim and Sharp 1998; Jin et al. 2003). RNA-Seq data for the regulated UAP1 exon are displayed in Figure 5B, showing that the PSI value increases from below 10% in conditions where RBFOX2 is depleted to 61% at the highest overexpression condition. Interestingly, inclusion of the alternative exon in UAP1 changes substrate specificity and leads to production of UDP-GlcNAc as opposed to UDP-GalNAc (REFS - Wang-Gillam et al., 1998). As this is the precursor to the essential O-GlcNAc modification, which is thought to be a nutrient and stress sensor (Zachara and Hart, 2004), precise titration of UDP-GlcNAc production may be accomplished by strongly monotonic splicing regulation. To assess the extent to which particular sequence motifs were associated with splicing regulation, we defined an MZ score for each 6mer as the average MZ value of those alternative exons which have the 6mer present in the first 200 bases of the downstream intron, a region in which RBFOX2 binding is associated with activation of exon inclusion (Ponthier et al. 2006; Yeo et al. 2009). Comparing motif MZ scores with RBNS R values of 6mers, we observed that >80% of
6mers with significant R values had positive MZ scores, consistent with a role in enhancement of splicing in response to increased RBFOX2 levels (Figure 5C). Positive MZ scores were observed not only for all 6mers containing the canonical GCAUG 5mer, but also for all 6mers containing the GCACG alternate motif, providing strong evidence that this motif confers RBFOX-dependent splicing regulation.

To assess the relationship between in vitro binding and regulatory activity for CELF1, we took advantage of data from an RNA-Seq time course analysis of splicing changes in heart following induction of CELF1 expression using a doxycycline-inducible transgenic mouse model (Wang et al, in preparation). CELF1 binding to upstream introns has been associated with repression of exon inclusion (Kalsotra et al. 2008; Dasgupta and Ladd 2012). Consistent with this activity, we observed that presence of a 6mer with significant R value in the upstream intron was associated with negative MZ scores (Figure 5D). A similar but slightly weaker bias for negative MZ scores was observed for exons that contained 6mers with significant R values, consistent with CELF1 conferring splicing repression when binding to exonic locations (Figure S4).

**RBNS identifies sequence biases in CLIP data**

CLIP-Seq is a widely used and highly effective technique for mapping RBP binding sites in vivo (Licatalosi et al. 2008; Sugimoto et al. 2012a). However, the lack of alternative methods has made it challenging to critically assess the quality of CLIP data for systematic biases or sources of false positives/false negatives. Previous studies have shown that CLIP favors U-rich sequences, because uridines form RNA-protein crosslinks more readily than other bases {Sugimoto, 2012 #282}. To explore potential compositional effects, we colored 6mers
according to the number of Us that they contained in the plot of RBFOX2 CLIP S/B against RBNS R values for all 6mers (Figure 6A). This simple visual aid revealed a group of 6mers with high U content (≥ 4 U out of 6) at the top center of the distribution that has no significant RBNS enrichment but high CLIP S/B values. In contrast, the remainder of high CLIP S/B 6mers also had significant RBNS R values and contained moderate numbers of Us (usually 1 or 2). This observation and the systematic trend for higher iCLIP S/B values to be associated with higher U content (Figure 6A; right) suggested that U-richness systematically and substantially enhances detection by CLIP, to an extent that it may detect protein-RNA interactions of low specificity in contexts that are sufficiently U-rich. An alternative interpretation would be that, in addition to UGCAUG and the related motifs identified by RBNS, RBFOX2 also has a high affinity for very U-rich sequences that was somehow missed by the RBNS assay and by previous SELEX analyses (Ponthier et al. 2006).

To determine the extent to which CLIP+/RBNS– motifs result from binding to U-rich sequences near authentic RBFOX motifs, we analyzed the sequences surrounding crosslinked CLIP+/RBNS– motifs (Figure 6C). We observed a ~2-fold increase in GCAUG motifs near these sites (within 40 nt) relative to uncrosslinked occurrences of these motifs, suggesting that some of these crosslink sites result from nearby RBFOX2 binding to canonical sites. The presence and magnitude of this effect can also be inferred from the observation that nearby bases are also enriched in UGCAUG’s meta motif plot (Figure 4B, top). While clustering of RBFOX sites may contribute, this effect also likely reflects crosslinking to other parts of the protein. Overall, presence of a nearby GCAUG motif was observed for only ~15% of crosslinked sites associated with CLIP+/RBNS– motifs (Figure 6C), suggesting that most of the CLIP signal for
such motifs instead derives from crosslinking of protein that is associated with RNA non-specifically or via interaction with other RBPs.

To assess the potential functional importance of motifs detected exclusively by CLIP, we compared the splicing regulatory activity of three sets of motifs: (i) 6mers with high CLIP S/B, but low RBNS R values (the CLIP+/RBNS– set); (ii) 6mers with significant RBNS R values and CLIP S/B values in the same range as the previous set (CLIP+/RBNS+); and (iii) a negative control group of sequences that lacked enrichment by CLIP or RBNS (CLIP–/RBNS–) (Figure 6A). Comparing the splicing regulation of cassette exons whose downstream introns contain 6mers from each set revealed a clear pattern: exons associated with the CLIP+/RBNS+ set had significantly higher MZ scores than those associated with either control 6mers, or with CLIP+/RBNS– 6mers. Furthermore, the CLIP+/RBNS– set was no more likely to be associated with high MZ values than the control set (Figure 6C). Thus, no evidence was found that the CLIP+/RBNS– set of motifs has regulatory activity in vivo. Instead, the simplest explanation is that these motifs result from transient nonspecific interactions of protein with RNA, with U-rich sequences simply being captured much more often than other nonspecifically bound RNAs. An alternative would be that they result from preferential crosslinking of U-rich sequences in the vicinity of specific binding directed by protein complexes involving RBFOX that do not predominantly activate splicing, or that the high CLIP S/B is coming from binding to a small subset of functional U-rich motifs. This analysis shows that RBNS can provide essential information for interpretation of CLIP-Seq data. On the other hand, the observation that essentially all significant RBNS 6mers also had high CLIP S/B values argues against the existence of a class of CLIP-invisible (e.g., uncrosslinkable) RNA motifs, at least for RBFOX2.
RBNS motifs are conserved across mammals

We expect that motifs that contribute to regulation of conserved alternative splicing events will often be evolutionarily conserved. Accordingly, the canonical binding motifs of RBFOX2, MBNL1 and CELF1 are highly conserved in introns flanking alternative exons and in 3′ UTRs (Sugnet et al. 2006; Wang et al. 2008; Daughters et al. 2009; Merkin et al. 2012; Wang et al. 2012). Adapting a method previously developed to assess conservation of microRNA target sites in mRNAs (Friedman et al. 2009), we assessed the conservation of significant RBFOX2 RBNS motifs in orthologous UTRs of 23 mammalian species. UTRs were chosen over introns because they can generally be more reliably aligned. For this analysis, we calculated for each 6mer the fraction of its occurrences in conserved UTRs that were evolutionarily conserved over at least a minimum evolutionary branch length (the “signal”), and measured a similar fraction for a cohort of control 6mers matched for genomic abundance, C+G% and CpG dinucleotide content, defining the mean conserved fraction over these control 6mers as the “background”. For RBFOX motifs, almost all 6mers containing the canonical GCAUG 5mer had conservation signal:background (S:B) ratios significantly above 1, indicating preferential conservation (Figure 6D). Furthermore, 6mers containing the alternative motif GCACG had S:B values nearly as high, further supporting the in vivo regulatory function of this motif. Some but not all of the remaining RBNS-detected motifs also showed significant S:B values, supporting a range of functional activity. No significant conservation was detected for the set of CLIP+/RBNS- 6mers (Figure 6E), again suggesting that this set is not enriched for functional regulatory motifs. By contrast, the set of CLIP+/RBNS+ motifs matched for CLIP S:B showed significant conservation (Figure 6E), defining this as a high-confidence set of functional regulatory motifs.
3.4 DISCUSSION

Proteins that bind RNA sequence-specifically play central roles in many aspects of gene expression. The data presented here demonstrate that RBNS – including the described analytical approaches – yields information that is both comprehensive and quantitative about the spectrum of RNA motifs bound by an RBP. As affinities for all kmers are assessed simultaneously, this approach may prove attractive as an alternative to traditional low-throughput quantitative methods.

Complexity of RNA binding affinity spectra

The depth of data generated in this approach yields information across a broad range of binding affinities, particularly when several RBP concentrations are used, enabling detection of weaker but significant motifs, such as GCACG for RBFOX2. For this particular example, the structure of the RBFOX1 RRM domain (which is identical to that of RBFOX2) has been solved by NMR, in complex with RNA representing canonical motif, UGCAUG (Auweter et al. 2006). The substitution of U for C in the fifth position of the 6mer would not introduce a steric clash, and one of the two hydrogen bonds that RBFOX1 makes with U5 would be preserved with a C in this position (Auweter et al. 2006). Together, these observations suggest that RBFOX proteins can bind GCACG in a manner similar to their binding of GCAUG, albeit with somewhat lower affinity. These observations, and similar results for a variety of variants of classical CELF1 and MBNL1 motifs, lead us to conclude that RBPs often have rather complex RNA binding affinity spectra, often centered on a few core essential bases, such as GU and GC in the cases of CELF1 and MBNL1. We also found that GCACG motifs are bound in vivo, and are associated with sequence conservation and splicing regulatory activity to an extent similar to canonical motifs. These and similar observations for a variety of variant CELF1 and MBNL1 motifs argue that
secondary motifs with affinities within an order of magnitude or so of the optimal motif often play functional roles in splicing regulation.

**Effects of structure on RNA binding**

We have also used RBNS data to make inferences about the impact of RNA structure on protein-RNA interactions. For RBFOX2 and CELF1, both of which bind RNA through RRM domains, our RNA folding analyses suggested strong preferences for binding of single-stranded RNA. Analysis of MBNL1, which binds RNA through zinc fingers, revealed a strong preference for unpaired Us but no significant bias for or against unpaired G and C bases in UGC-containing motifs, suggesting either that MBNL can melt paired GC dinucleotides or that it can recognize them even when base-paired. CUG repeat RNA, which is recognized by MBNL proteins (Teplova and Patel 2008a), crystallizes as a hairpin with base pairing of GC repeats separated by unpaired U-U bulges (Mooers et al. 2005a), consistent with the pattern of MBNL binding preferences observed here. Intron 4 of cardiac troponin T (cTNT), a well-characterized MBNL binding and regulatory target, also contains multiple paired GCs flanked by unpaired pyrimidine bulges (Warf and Berglund 2007). Consistently, biochemical evidence has shown that MBNL binds with high affinity to pairs of GC dinucleotides with a wide range (~1-15 bases) of intervening pyrimidines (Goers et al. 2010a). This structural signature is consistent with RNA looping around MBNL proteins such that different zinc fingers interact with different GCs, and we might speculate that looping involving pairing of GCs facilitates MBNL binding by bringing GCs into proximity. RNA looping as a mechanism of RNA recognition has been proposed for PTB and is also consistent with the crystal structure of MBNL1 zinc fingers 3 and 4 (Teplova and Patel 2008b).
RBNS enhances interpretation of CLIP data

The *in vivo* context of CLIP experiments has yielded many insights into the direct regulation of splicing and other post-transcriptional processes by particular RBPs. However, the complexity of these co-regulated processes often results in binding signatures that are difficult to explain solely by direct binding to consensus motifs. In the case of RBFOX2, as roughly half of binding events lack the consensus pentamer and more than 80% of expressed intronic consensus motifs are unbound (Yeo et al. 2009; Jangi et al. in review), additional parameters must be influencing which motifs are detected as bound. We posit that the analysis provided by RBNS serves to filter out noise as well as deconvolute complex biological signals generated in CLIP data.

First, the structural preferences determined by RBNS may help define which motif occurrences are accessible to be bound *in vivo*. As RBNS is performed in the absence of other RBPs or other cellular components, we demonstrated that the effect of structure on binding can be effectively isolated and measured. Applying these parameters to expressed motif occurrences that are not bound *in vivo* will allow loci with sub-optimal structure to be distinguished from those motifs that are made inaccessible by other means, for example due to binding by a competing RBP.

Second, RBNS may yield a less biased portrait sequence binding affinity of an RBP’s than CLIP (including iCLIP, PAR-CLIP) data. For example, when crosslinking to a CLIP+/RBNS– sequence is observed in close proximity to a CLIP+/RBNS+ sequence, it is likely that the protein’s sequence-specific RNA binding domain is actually bound at the location of the CLIP+/RBNS+ motif. To improve the effective resolution of CLIP data, this correction could be
Chapter 3. Comprehensive, quantitative protein-RNA binding

made automatically, using RBNS affinities to narrow the actual binding site. From a more biological standpoint, since RBNS is carried out in the presence of only RNA and one RBP, it can be used to distinguish cases when a protein binds directly to RNA versus indirectly through interaction with another RBP. Many binding events detected in vivo may arise from these secondary interactions, and knowledge of the spectrum of low-affinity binding sites in the absence of other RBPs will distinguish non-specific binding from cooperative binding through another RBP. This approach will become more effective when RBNS data are available from a larger set of RBPs.

Acknowledgements

We thank Andy Berglund for advice on expression of RNA binding proteins, and Tom Cooper, Wendy Gilbert and A. B. for helpful suggestions. This work was funded by an NIH NRSA Postdoctoral Fellowship (N.L.) and by grants from the NIH (C.B.B.).
3.5 MATERIALS AND METHODS

Cloning, expression and purification of proteins
To create a tandem affinity tag, a streptavidin binding peptide tag was added to the pGex-6P1 vector (GE), downstream of the GST tag and PreScission protease site. Full length CELF1, MBNL1 (1-260), and RBFOX2 (100-194) were cloned downstream of the SBP tag by infusion (Clonetech) using BamHI and NotI cloning sites. Both truncated MBNL1 and RBFOX2 constructs contain all RNA binding domains, including all four MBNL1 Zinc finger domains and RBFOX2’s single RNA recognition motif (RRM). The proteins were expressed at 18 degrees for 4 hours in the Rosetta(DE3)pLysS E. coli strain. CELF1, MBNL1 and RBFOX2 were then purified via the GST tag and eluted from GST GraviTrap columns (GE) by cleaving off the GST tag with 120 Units of Prescission protease (GE) in 4 mL of protease cleavage buffer (50mM Tris pH 7.0, 150 mM NaCl, 1mM EDTA, 1 mM DTT) at 4 degrees overnight (~12-16hours) and stored in storage buffer (20 mM Tris pH 7.5, 300mM NaCl, 0.5 mM EDTA, 0.5 mM DTT, 10% glycerol). Protein purity was assessed by running all purified proteins on a SDS-PAGE gel and all protein products were visualized with SimplyBlue SafeStain (Invitrogen).

Preparation of random RNA
RBNS input random RNA was prepared by in vitro transcription using the RBNS T7 template, a DNA oligo containing a random 40mer sequence flanked by priming sites for the addition of Illumina adapters and the T7 promoter sequence. To artificially create a double-stranded T7 promoter, the T7 oligo was annealed to the region of the RBNS T7 template corresponding to the T7 promoter sequence by heating the template, T7 oligo and water at 65 degrees for 5 minutes and then allowing the solution to cool at room temperature for 2 minutes. The RBNS input RNA pool was then in vitro transcribed with T7 polymerase using Ampliscribe (epibio) or HiScribe T7 In vitro transcription kits (NEB). The in vitro transcribed RNA was then gel-purified on a 6% TBE-Urea polyacrylamide gel. The resulting RBNS input RNA pool:

GAGTTCTACAGTCCGACGATC(N)40TGGAATTCTCGGGTGTCAAGG.

RBNS
CCTTGACACCCGAGAATTCACA(N)40GATCGTCCGAGGTAGAAGACCCCTATAGTGAAGT

T7 oligo: TAATACGACTCACTATAGGG

RBNS
RBNS was performed after purifying a given RBP and in vitro transcribing RBNS input RNA. 7-10 concentrations of RBP, including a no RBP condition was equilibrated in 250ul of binding buffer (25mM tris pH 7.5, 150 mM KCl, 3mM MgCl2, 0.01% tween, 1 mg/mL BSA, 1 mM DTT, 30 ug/mL poly I/C (sigma)) for 30 minutes at room temperature. RBNS input random RNA was then added to a final concentration of 1uM with 40 U of Supersasin (Ambion). RBP and RNA were incubated for 1 hour at room temperature. During this incubation streptavidin magnetic beads (Invitrogen) were washed 3 times with 1 mL of wash buffer (25mM tris pH 7.5, 150 mM KCl, 60 ug/mL BSA, 0.5 mM EDTA, 0.01% tween) and then equilibrated in binding buffer until needed. To pull down tagged RBP and interacting RNA each RNA/protein solution was then added to 1 mg of washed streptavidin magnetic beads and incubated for one hour. Unbound RNA was removed from the beads and the beads were washed once with 1 mL of wash.
buffer. The beads were incubated at 70 degrees for 10 minutes in 100 uL of elution buffer (10mM tris pH 7.0, 1mM EDTA, 1%SDS) and the eluted material collected. Bound RNA was extracted from the eluate by phenol/chloroform extraction and ethanol precipitation. Half of the extracted RNA from each condition was reverse transcribed into cDNA using Superscript III (Invitrogen) according to manufacturer’s instructions using the RBNS RT primer. To control for any nucleotide biases in the input random library, 0.5 pmol of the RBNS input RNA pool was also reverse transcribed and Illumina sequencing library prep followed for all experimental conditions as outlined below. To make Illumina sequencing libraries, primers with Illumina adapters and sequencing barcodes were used to amplify the cDNA by PCR using high fidelity Phusion (NEB) with 8-10 amplification cycles. PCR primers always included RNA PCR 1 (RP1) and one the indexed primers listed below. PCR products were then gel-purified from 8% TBE polyacrylamide gels. Sequencing libraries corresponding to all concentrations of a given RBP were pooled in a single lane and the random 40mer was sequenced on the HighSeq2000.

RBNS RT primer: GCCTTGGCACCACCGAGAATTCCA
RNA PCR (RP1)
AATGATACGGCGACCA
CAAGCAGAAGACCGCATACGAGATCTACACGTTCCAGAGTCTACAGTCCGACGATC
Index 1 (RP11)
CAAGCAGAAGACCGCATACGAGATCGTGATGTGACTGGAGTTCTTGGCACCACCGAGAATTCCA
Index 2 (RP12)
CAAGCAGAAGACCGCATACGAGATACGGTGACTGGAGTTCTTGGCACCACCGAGAATTCCA
Index 3 (RP13)
CAAGCAGAAGACCGCATACGAGATGCTTGGCACCACCGAGAATTCCA
Index 4 (RP14)
CAAGCAGAAGACCGCATACGAGATTGCTACGACTGGAGTTCTTGGCACCACCGAGAATTCCA
Index 5 (RP15)
CAAGCAGAAGACCGCATACGAGATCTGTGTGACTGGAGTTCTTGGCACCACCGAGAATTCCA
Index 6 (RP16)
CAAGCAGAAGACCGCATACGAGATATTGGCGTGACTGGAGTTCTTGGCACCACCGAGAATTCCA
Index 7 (RP17)
CAAGCAGAAGACCGCATACGAGATCTCTGGTGACTGGAGTTCTTGGCACCACCGAGAATTCCA
Index 8 (RP18)
CAAGCAGAAGACCGCATACGAGATTTGCTGAGTCTGGAGTTCTTGGCACCACCGAGAATTCCA
Index 9 (RP19)
CAAGCAGAAGACCGCATACGAGATCTGATCTGACTGGAGTTCTTGGCACCACCGAGAATTCCA
Index 10 (RP110)
CAAGCAGAAGACCGCATACGAGATAATGGTGACTGGAGTTCTTGGCACCACCGAGAATTCCA

Modeling of RBNS experiment
To generate a model of the expected conservation signal to background from first principles, we simulated the binding of a protein \( R \) with a ligand pool \( L \) using custom Matlab scripts, using a range of protein concentrations \([R]_T\) from 1 nM to 10 \( \mu \)M and holding the ligand pool concentration \([L]_T\) at 1 \( \mu \)M. Where specified, we also included background, nonspecific binding sites \( B \) such that nonspecific binding of the ligand pool \([BL]_T\) would either be completely absent, 1 nM (“low”), 10 nM (“medium”), or 100 nM (“high”). The concentration of specific motifs \([L_1]_T\) and \([L_2]_T\) was calculated by multiplying the total ligand concentration \([L]_T\) by the probability of an arbitrary motif of length 6 occurring by chance within a ligand of length 40 nt, with the remaining ligand pool containing neither motif denoted as \([L_o]_T\). Setting the dissociation constant for the “strong” motif \( K_{d,L_1} = 5 \) nM, and that of the moderate motif \( K_{d,L_2} = 30 \) nM, we
constructed the linear system of equations relating the concentrations of the free and bound states of $R$, $B$, $L_1$, $L_2$, and $L_o$ to the $K_d$ values for each protein–ligand pair and the total concentrations of each species. This system was solved numerically for each input value of $[R]_T$. Ligand occupancy was calculated by dividing the total bound concentration (including specific and non-specific binding) of a particular motif by the total concentration of that motif in the library. Motif conservation signal to background was calculated as follows: the total bound concentration (specific and non-specific) of a particular motif was first divided by the total bound concentration of all ligands multiplied by the 35, the number of motifs of length 6 within a ligand of length 40. This value was then divided by $(0.25)^6$, the background expectation of any particular motif of length 6 in a library without enrichment for any specific motif.

**Streaming kmer Analysis (SKA)**

SKA is an algorithm designed to estimate the fraction of RBNS reads which are bound at each possible kmer for a single RBNS library and a given value of $k$. Tests on randomly simulated data indicate it is highly accurate given the assumptions of the simulation. The algorithm starts with a uniform distribution of binding weights over all $k$mers and iterates through the reads, incrementally adjusting the binding weights. The algorithm passes through the reads multiple times in order to converge on the correct distribution of weights. The first pass is used to obtain an initial estimate of the fraction of RBNS reads which are bound at each possible kmer and subsequent passes are used to converge to the correct distribution.

The algorithm works as follows. The binding weights for all $4^k$ kmers are initialized with a pseudocount of 1.0. This is equivalent to assignment of binding to one read to each kmer. During the first pass through the reads, the weights are updated continually by the following process. For each read, the $k$mers present within the read’s sequence are enumerated (one kmer beginning at each position from 1 to $\lambda - k + 1$ in the read). The read’s total weight of 1.0 is fractionally assigned to the kmers present in its sequence by the following process. The current estimates of the binding weights are summed up for this set of kmers to obtain a normalization factor. The binding weights of each of this set of kmers are incremented by the current weight of the kmer normalized by the summed weights of all the kmers present. As this process continues, highly represented kmers will have increasing weights relative to other kmers. Because of this, the algorithm “learns” to assign a larger fraction of a read to the more highly represented kmers than other weaker kmers also present in that read.

After the reads have been processed once, subsequent passes are done slightly differently. A new set of weights for every kmer is initialized to 0. The reads are processed and fractionally assigned to the kmers based on the weights in the previous pass. Once all the reads have been assigned, this new set of weights is taken to be the best current estimate of the true binding distribution. It can either be normalized and returned or plugged into another pass through the reads depending on how close the algorithm is to convergence (which can be assessed by the magnitude of change in the weights between successive passes). We find that the number of passes required depends on the number of reads in the library and that with our libraries two passes are sufficient.

**Meta motif plots**

The meta motif plots were generated using Samtools and custom Python scripts and by iterating over each intron (not including introns which overlap with a cassette exon) and each 3' UTR and selecting all examples of each kmer which were at least 40 nt from the edge of the region (either
the splice sites, the Stop codon or the cleavage site). For each kmer the (i)CLIP density over all instances of the kmer was summed up and plotted. The CLIP signal to background was calculated as the ratio of the average coverage over the kmer in all cases selected as described above to the average coverage over the regions 80 to 40 nt upstream of the kmer and 40 to 80 nt downstream of the kmer.

Monotonicity Scores
The monotonicity scores were calculated as introduced in Wang et al. Each of the eight RNA-seq libraries was mapped to the mouse genome (mm9) with Tophat and the alternative splicing of skipped exon (SE) events was analyzed with MISO as follows. All pairwise comparisons between the libraries were done and the significantly (Bayes factor $\geq 5.0$) changing events were identified. The difference between the number of comparisons where the higher RBFOX concentration showed significantly more inclusion and the number where the lower RBFOX concentration showed more inclusion was calculated for all events. For each skipped exon event the monotonicity score was defined to be the z score of this difference out of a control set of differences generated by shuffling the order of the Fox concentrations.

Monotonicity of the false positives
To assess splicing effects of the false positive kmers, the total sets of true and false positives as defined by the rectangle in 6A. kmers were subsampled from the false positive set to match the size of the true positive set. For each subsample, the set of exons whose downstream introns contained a kmer from one but not the other set of kmers were separated out and CDFs for the corresponding monotonicity scores were generated. The figure shows the median CDF for each set of kmers. The true negatives were compared to the true positives analogously.

Conservation signal to background
The conservation signal to background was calculated in a similar way to as described by Friedman et al. For each kmer probed a control set of 25 kmers was generated such that the CpG content was maintained and the number of occurrences in UTRs was similar to the tested kmer. The number times each kmer (tested kmer and those in the control set) was conserved to each possible branch length was calculated. The S/B was calculated as the ratio of the fraction of kmer occurrences that were conserved to a branch length of 1.5 to the mean conserved fraction of the control set.

3.6 REFERENCES
Chapter 3. Comprehensive, quantitative protein-RNA binding


Chapter 3. Comprehensive, quantitative protein-RNA binding


Robertson A. 2013. Understanding Regulation of mRNA by RNA Binding Proteins. in *Computational and Systems Biology*. Massachusetts Institute of Technology, Cambridge, MA.


Chapter 3. Comprehensive, quantitative protein-RNA binding


3.7 FIGURES

Figure 1. RNA Bind-n-Seq overview and motif enrichment analysis.
A. Overview of the experimental method. In the experiment tagged protein is incubated with random RNA oligos (typically 40mers with sequencing tags) at each of several concentrations of protein with a fixed concentration of RNA oligo. The RBP is pulled down and the associated RNA is sequenced along with the input in multiplex. The counts of sequences in this library are used to estimate concentrations of bound RNA molecules.
B. Stacked histogram of RBFOX2’s Bind-n-seq signal to background (S/B) for every 6mer sequence at RBFOX2 concentration of 365nM. Color scale shows fraction of sequences...
within the bin that contain the known motif, GCAUG and the novel secondary motif GCACG. The Y axis plotted on log scale.

C. Histogram of CELF1’s RBNS R for every 7mer sequence at CELF1 concentration of 1μM. Stacked bars are colored as indicated in legend.

D. Histogram of MBNL1’s RBNS R for every 7mer sequence at MBNL1 concentration of 250nM. Stacked bars are colored as indicated in legend.

E. Aligned enriched CELF1 motifs clustered by separation of GU’s. Red sequences indicate two UGU’s are present; orange indicates one is present. Pictographs of each cluster are shown below.

F. Aligned enriched Mbnl1 motifs clustered by separation of GC’s. Orange indicates GCUU is present; gray indicates no GCUU is present. Pictographs of each cluster are shown below.
Figure 2. Estimation of dissociation constants using RBNS data.
A. Model of RBNS enrichment profiles under basic assumptions. Standard binding curves for two motifs of different binding affinities.
B. Predicted RNA Bind-n-seq for a single motif with background nonspecific binding (NSB) at various strengths (dashed: no NSB, dash/dotted low NSB, solid moderate NSB, dotted high NSB).
C. Predicted S/B in Bind-n-seq for a strong motif (red) and 10 weaker motifs (orange) with background nonspecific binding.
D. RBNS R values for several top enriched 6mers and several random 6mers are shown as a function of RBP concentration for each RBP studied, RBFOX2, MBNL1 and CELF1. For RBFOX2 UGCAUG and UGCACG are shown. For CELF1 we show the four sequences UGUUUGU. For MBNL1 we illustrate the spacing of GCs within Us.
E. Correlation of relative K_d s for several kmers for RBFOX2 as estimated by Bind-n-seq (at RBFOX concentration 121nM) and SPR. Correlation is significant by Pearson test (R=0.935, P=7e-4). Motifs are colored as in Figure 1B.
Chapter 3. Comprehensive, quantitative protein-RNA binding
Chapter 3. Comprehensive, quantitative protein-RNA binding

Figure 3

A. For every occurrence of the RBFOX2 primary motif (UGCAUG) in each RBFOX2 selected library, the (Vienna RNAfold) predicted probability that each base in the motif is paired is averaged over the 6 bases of the motif. The reads are binned based on this average probability and the RBNS R value is calculated. The R values of these bins are plotted for several concentrations for the three proteins.

B. For a top motif for each of the proteins, the probability that each base is paired is calculated for each oligo in the selected library and in the input control library. The ratio of probabilities that the base is paired in the selected library to the probability that it is paired in the input control library is shown on a log scale.

C. The ratio of base pairing probabilities for GCUU motifs within alternative exons of different evolutionary ages and constitutive exons is calculated and plotted on a log scale.
Chapter 3. Comprehensive, quantitative protein-RNA binding

Figure 4
A. The iCLIP crosslinking sites are shown for an example intron in the kinase Dyrk1a centered on an example secondary motif.
B. Meta-motif plots are shown for RBFOX2 iCLIP in mESCs over all occurrences of the primary motif in introns (top left panel) and in 3' UTRs (top right panel). Meta-motif plot of RBFOX2 iCLIP crosslinking sites in mESCs over secondary Fox 5mer motif (GCACG) in
in introns (left) and 3' UTRs (right). Bottom panels show the negative control meta-motif plots for Upf1 CLIP over the same regions. Numbers indicate the scale of the y axis.

C. Top Panels: meta-motif plot of MBNL1 CLIP coverage in C2C12 cells over the top MBNL1 6mer motif (GCUUGC) in introns (left) and 3' UTRs (right). Bottom panels: meta-motif plot of CELF1 CLIP coverage in C2C12 cells over the top CELF1 6mer (UUUUGU) motif in introns (left) and 3' UTRs (right).

D. For each 6mer the (i)CLIP S/B (see online methods) is plotted against the RBNS R for the most enriched concentration of RBP for each RBP. Left panel shows RBFOX2 mESCs iCLIP signal in introns. Yellow points indicate the secondary motif 5mer is present in the 6mer. Middle panel shows CELF1 CLIP coverage in introns for C2C12 cells. Right panel shows MBNL1 CLIP coverage in C2C12 in 3' UTRs. Red points indicate 6mers which are significantly enriched in the Bind-n-seq experiment. Histograms at right of each scatter plot indicate the normalized distributions of CLIP S/B for 6mers grouped by color as in previous figures.
Chapter 3. Comprehensive, quantitative protein-RNA binding

Figure 5
A. RBFOX western is shown for Dox inducible Fox2 mES cell lines. Cells are treated with either a control hairpin targeting GFP (left lanes) or Fox2 (right lanes). Cells were treated with 0, 0.05, 0.1 or 1 µg/mL of Dox. Western shows endogenous and tagged Dox induced Fox2 as well as Vinculin control.
B. For each of the 8 possible levels of Fox2 (two hairpins x four levels of Dox) a plot of the RNA-seq reads mapping to junctions of a highly Fox-sensitive skipped exon in the pyrophosphorylase Uap1. Distributions of estimated PSI values are shown at right of each RNA-seq profile.

C. Monotonicity scores (see online methods) are calculated for 1442 skipped exons in mESC expressed genes. For each 6mer, the average monotonicity score is plotted against RBNS R for all skipped exons where the kmer is present in the downstream intron (within 200 nt). Points are colored by sequence as in previous plots. The histogram at right shows the distributions of monotonicity scores for the enriched and unenriched sequences (significantly different by KS test, p=2e-7). Histogram at bottom shows the distribution of RBNS R values.

D. Same as C for CELF1 motifs in upstream introns within 200 nt of cassette exons for a time course in mouse muscle as CELF increases. The MZ scores differ significantly between enriched and unenriched RBNS sequences (significantly different by KS test, p=2e-18).
Chapter 3. Comprehensive, quantitative protein-RNA binding

A

B

C

D

E
**Chapter 3. Comprehensive, quantitative protein-RNA binding**

**Figure 6**
A. Bind-n-seq indicates sequence biases in iCLIP. As in (Figure 4D) RBFOX2 iCLIP S/B in UTRs is plotted against RBNS R. Points are colored by the number of U bases present in the sequence and the histogram at the side indicate the distributions of iCLIP S/B’s for each number of U’s present. Pink rectangle indicates the set of 6mers enriched in CLIP, but not in Bind-n-seq (CLIP+/RBNS-). Green rectangle indicates the set of kmers enriched in both CLIP and Bind-n-seq (CLIP+/RBNS+). The dark gray box indicates the 6mers enriched in neither (CLIP-/RBNS-).

B. Cumulative distribution of MZ scores for skipped exons that contain sequences from the three sets enumerated in A.

C. RBFOX primary motifs are selectively present near to crosslinked CLIP+/RBNS- sequences. For each CLIP+/RBNS- motif in either introns (left) or UTRs (right) fraction of these motifs that had a primary GCAUG within 40nt was calculated for all motif occurrences that was crosslinked in iCLIP or uncrosslinked.

D. A plot of the S/B conservation of the top RBFOX2 Bind-n-seq motifs in mammalian 3’ UTRs (see online methods). Motifs are listed in descending order by R and colored as previously.

E. Box plots of the distributions of conservation signal for the CLIP/RBNS kmer sets as defined in Figure 6A.
**Chapter 3. Comprehensive, quantitative protein-RNA binding**

Supplemental Figure S1

A. Comparison of R values for oligos of length 10 and 40 nt. Plot shown for ~120 nM RBFOX2.

B. As in S1B but for B values calculated using equation 1. Plot shown for 121 nM RBFOX2.

C. SKA library fractions shown for in the 10mer oligo and the 40mer oligo data. The 10mer data is at 120 nM RBFOX and the 40mer data is at 265nM RBFOX.
Chapter 3. Comprehensive, quantitative protein-RNA binding
Supplemental Figure S2
A. A Correlation of relative $K_d$s for several kmers for RBFOX2 as estimated by Bind-n-seq (at RBFOX concentration 121nM) and SPR data from Auweter et al. Correlation is significant by Pearson test ($R=0.958$, $P=0.003$)
Supplemental Figure S3
Meta motif plots of (i) CLIP density versus R for 6mers for RBFOX2 (A), MBNL1 (B) and CELF1 (C). Plot shows 3' UTR positions.
Chapter 3. Comprehensive, quantitative protein-RNA binding
Supplemental Figure S4
A. Average CELF1 MZ scores for SEs which contain each of the 6mers within the exon.
Chapter 3. Comprehensive, quantitative protein-RNA binding
Chapter 4

Co-transcriptional splicing regulation by Rbfox2

The work described in this chapter is an initial characterization of mechanisms of co-transcriptional splicing mediated by Rbfox2. The correlations made through computational analyses will need further experimental validation before specific mechanistic conclusions can be reached. We propose alternate interpretations of the data and highlight the particular questions that will require future investigation. This chapter is adapted from the following manuscript in preparation:


Experimental contributions:

Mohini Jangi performed cell fractionations, chromatin immunoprecipitations, co-immunoprecipitations, and all computational analyses. Prakriti Paul performed splicing assays.
Chapter 4. Co-transcriptional splicing regulation by Rbfox2

4.1 ABSTRACT

The majority of mammalian splicing occurs co-transcriptionally. However, the impact of the chromatin environment on alternative splicing remains unclear. Here, we investigate mechanisms of co-transcriptional splicing by the RNA binding protein Rbfox2. Using chromatin immunoprecipitation (ChIP) of epitope-tagged Rbfox2 coupled to high-throughput sequencing, we identify thousands of Rbfox2-bound regions in mouse embryonic stem cells. Rbfox2 binding sites are enriched within gene promoters and transcriptional enhancers and are not biased toward genes containing Rbfox2-dependent splicing events. Rbfox2 ChIP signal is strongly coincident with serine 5-phosphorylated RNA polymerase II (Pol II), and Rbfox2 co-immunoprecipitates with Pol II in an RNA-independent manner. Modulation of chromatin structure by a small-molecule inhibitor of histone deacetylases leads to increased Rbfox2-dependent splicing activity, consistent with decreased Pol II elongation rates leading to a kinetic effect on Rbfox2-mediated co-transcriptional splicing. We propose that Rbfox2 is recruited to sites of transcription initiation and is influenced by transcription elongation rates as it tracks with Pol II to be deposited at sites of splicing regulation.
Chapter 4. Co-transcriptional splicing regulation by Rbfox2

4.2 INTRODUCTION

Alternative splicing is a major contributor to context-dependent diversification of the transcriptome and proteome. Over the past three decades, much insight has been gained in the function of interactions between cis-regulatory elements within nascent transcripts and trans-acting splicing factors (Barash et al. 2010). Increasing evidence that splicing occurs co-transcriptionally (Osheim et al. 1985; Listerman et al. 2006; Pandya-Jones and Black 2009) and is functionally coupled to transcriptional elongation (McCracken et al. 1997; Lin et al. 2008; Singh and Padgett 2009) has sparked more recent interest in the regulatory potential of chromatin and transcription in splicing.

While early studies demonstrated a link between transcription and constitutive splicing (Tennyson et al. 1995; Furger et al. 2002; Pandya-Jones and Black 2009), observations that introns were not necessarily removed sequentially suggested that co-transcriptionality could also impact alternative splicing (LeMaire and Thummel 1990; Kessler et al. 1993; Bauren and Wieslander 1994; Attanasio et al. 2003). Chromatin structure and histone modifications may play an important role in this process. A series of genome-wide studies demonstrated that nucleosomes are positioned on exons (Schwartz et al. 2009; Spies et al. 2009; Tilgner et al. 2009). In addition, the H3K36me3 modification is significantly enriched on exons, with a stronger enrichment on constitutive exons than alternative exons, while the first intron is marked by H3K79me2 (Kolasinska-Zwierz et al. 2009; Huff et al. 2010). Splicing-dependent feedback between RNA polymerase II phosphorylated on serine 2 of the C-terminal domain heptad repeats (Pol II ser2p) and the H3K36me3 mark through recruitment of the H3K36 methyltransferase SetD2 by Pol II ser2p has been proposed to explain the differential distribution of the H3K36me3 modification (de Almeida et al. 2011; Kim et al. 2011).
Chapter 4. Co-transcriptional splicing regulation by Rbfox2

Two general models of co-transcriptional alternative splicing regulation have emerged to explain these and other pieces of evidence (Luco et al. 2011). The kinetic model posits that the speed of the elongating polymerase influences competition between splice sites, driving alternative exon choice. This builds on the observation that promoter strength and identity are able to influence alternative exon inclusion due to promoter-specific recruitment and co-transcriptional tracking of splicing machinery (Cramer et al. 1997). More directly, elongation rates of RNA polymerase II (Pol II) can determine the outcome of competing co-transcriptional splicing reactions (Kadener et al. 2001; de la Mata et al. 2003; Batsche et al. 2006). Gaining more recent support is the recruitment model, whereby splicing regulators may bind specific histone modifications to increase the local concentration of splicing factor in a cell type-specific manner. Perhaps the strongest argument for the recruitment model is the causal regulation of PTB-dependent splicing by H3K36me3, mediated through the adaptor protein MRG15 that recognizes the H3K36me3 mark and interacts with PTB (Luco et al. 2010). Similarly, the H3K4me3-binding adaptor CHD1 recruits U2 components to the nascent transcript to alter splicing efficiency (Sims et al. 2007).

In this study, we have sought to address the contribution of these two models of co-transcriptional splicing to regulation by the alternative splicing factor Rbfox2. While extensive evidence from our own and other studies has supported a model in which Rbfox2 binds UGCAUG motifs to repress downstream exons or activate upstream exons, several observations suggest more complex regulation (Jin et al. 2003; Nakahata and Kawamoto 2005; Underwood et al. 2005; Auweter et al. 2006; Ponthier et al. 2006; Yeo et al. 2009; Jangi et al. submitted). A subset of cassette exons predicted to be Rbfox2 targets based on proximity to conserved motifs is regulated in opposite directions in brain and muscle, two tissues in which Rbfox protein
expression is abundant (Zhang et al. 2008). Moreover, Rbfox2 crosslinking immunoprecipitation sequencing (CLIP-seq) in human and mouse embryonic stem cells shows many conserved motifs in expressed genes where binding is not detected (Yeo et al. 2009; Jangi et al. submitted). This implies a mechanism to select specific subsets of UGCAUG motifs to be active in a particular cellular condition.

To deconvolute these observations, we have performed ChIP-seq for Rbfox2 in mouse embryonic stem cells and recovered thousands of high-confidence Rbfox2 binding sites across the genome. This revealed a strong enrichment for widespread binding to transcription start sites and enhancers of actively transcribing genes. We further demonstrate that this is likely mediated through interaction with serine 5-phosphorylated Pol II. While promoter recruitment of Rbfox2 is not specific to Rbfox2-regulated genes, splicing of Rbfox2 target exons can be altered by changing chromatin structure using a histone deacetylase inhibitor. We propose that Rbfox2 is primed for splicing by recruitment to active promoters and may be further regulated through Pol II pausing at intragenic chromatin marks, in particular those marking transcriptional enhancers. These findings have broad implications both in the mechanistic models of splicing machinery recruitment and in cell type-dependent alternative splicing regulation.

4.3 RESULTS

A subpopulation of Rbfox2 fractionates with chromatin independent of RNA

We performed cellular fractionation into cytoplasmic, nucleoplasmic, and chromatin fractions to determine the partitioning of various components of the splicing machinery. To also determine the effect of active transcription on the subcellular localization of splicing machinery, cells were treated with the reversible PTEF-b inhibitor flavopiridol for 1 hour prior to harvesting,
and control cells were allowed to recover in fresh media for 1 hour after the flavopiridol pulse (Rahl et al. 2010). Upon recovery from transcriptional inhibition, all splicing components examined showed a fraction associated with the chromatin, although this was highest for Rbfox2 and PTB, two alternative splicing regulators, and lowest for the U1 components U170K and U1C (Fig.1A). The small and large subunit of U2-associated factor, U2AF35 and U2AF65, showed equivalent partitioning between cytoplasm, nucleoplasm, and chromatin, indicative of nucleocytoplasmic shuttling. Sf3b3 and the U5 component Prp8, both present in the active spliceosome, fractionated equivalently in the nucleoplasm and chromatin. This suggested that core components of the splicing machinery that persist in later steps of the splicing reaction, as well as the alternative splicing regulators PTB and Rbfox2, were more likely to partition with chromatin.

Inhibition of transcriptional elongation by flavopiridol generally resulted in decreased chromatin fractionation of all spliceosomal components assayed; however, appreciable amounts of Rbfox2, PTB, Sf3b3, and Prp8 remained in the chromatin fraction. Treatment with RNase did not greatly alter the nucleoplasmic-chromatin distribution after transcriptional inhibition, suggesting these associations were unlikely to be mediated through Pol II-tethered transcripts not released from the chromatin. Interestingly, RNase treatment increased chromatin association of PTB, U170K, and U1C, perhaps indicating a default association with chromatin that is then competed away by transcribed RNA or the U1 snRNA.

**Rbfox2 ChIP peaks are enriched at the 5’ ends of genes**

To gain insight into regulatory mechanisms afforded by the interaction of Rbfox2 with chromatin, we performed ChIP-seq for epitope-tagged and endogenous Rbfox2 in mouse
Chapter 4. Co-transcriptional splicing regulation by Rbfox2

embryonic stem cells (Fig.1B). For the tagged ChIP-seq experiments, we stably transduced V6.5 mESC with a lentiviral vector carrying FLAG-HA tagged or untagged human RBFOX2 (FHFOX2 or UntagFOX2) expressed from a doxycycline-inducible promoter. Titration of doxycycline allowed for expression of the transgene at levels similar to endogenous Rbxo2, as previously described (Jangi et al. in review). As a comparison, we also performed ChIP-seq using a monoclonal antibody against U2AF65, which showed transcription-dependent chromatin association (Fig.1A), and verified all observed trends with a second α-U2AF65 antibody (Supplemental Fig.S1A). Mapped reads for all datasets were clustered using the MACS algorithm, using UntagFOX2 as the control for the FHFOX2 ChIP or chromatin from whole-cell extract (WCE) as the control for endogenous Rbxo2 or U2AF65 ChIP. Compared to WCE, UntagFOX2 yielded 935 significantly enriched peaks ($P < 1e^{-7}$). These are most likely attributed to “hyper-ChIPpable” regions of highly expressed genes determined in a recent study in S. cerevisiae (Teytelman et al. 2013) and do not factor into our analysis because FHFOX2 ChIP peaks are called using UntagFOX2 as the background. Consequently, FHFOX2 vs UntagFOX2 yielded 12,500 statistically significantly enriched peaks ($P < 1e^{-7}$) while endogenous Rbxo2 vs. WCE yielded 25,690 peaks. As comparisons between the two sets of called peaks showed that they were significantly overlapping and consistent in all of the trends described below, we will describe the results for the FHFOX2 vs. UntagFOX2 peaks due to the increased stringency against non-specific chromatin interactions.

We first determined the distribution of FHFOX2 ChIP peaks within various genomic regions. When compared to a set of randomly-distributed peaks, FHFOX2 ChIP peaks were significantly depleted in intergenic regions as well as introns but were strongly enriched in all other genic categories (Fig.1C,D). Surprisingly, the most strongly enriched categories were
those occurring near the transcription start site (TSS), including promoter regions (defined as 2kb upstream of TSS), 5’UTR, and first coding exon. Together with the intronic depletion, this suggested that the pattern of Rbfox2 recruitment to the chromatin was distinct from its eventual recruitment to sites of regulation within the pre-mRNA, which we and others have previously demonstrated to be proximal to alternative exons.

**Rbfox2 accumulates around active transcription start sites and transcriptional enhancers**

As a parallel method of subcategorizing FHFOX2 peaks, we next intersected FHFOX2 peaks with a diverse set of peaks from ChIPseq datasets representing distinct histone marks and chromatin-associated complexes (Mikkelsen et al. 2007; Chen et al. 2008; Ku et al. 2008; Marson et al. 2008; Seila et al. 2008; Creyghton et al. 2010; Kagey et al. 2010; Rahl et al. 2010; Whyte et al. 2012). A correlation heatmap was generated to examine the frequency of co-occurrence of any set of marks, considering only those peaks that intersected with FHFOX2 peaks (Fig.2A). Each pairwise comparison between all datasets was summarized by a correlation coefficient, with 1 signifying, for example, that all peaks of mark A that overlap with FHFOX2 also overlap with mark B, and -1 signifying that no peaks of mark A that overlap with FHFOX2 overlap with mark B. Two clusters of marks emerged that clustered tightly and represented a large fraction of FHFOX2 peaks. The first contained H3K4me1, H3K27ac, and Lsd1, which have all been reported to be associated with active transcriptional enhancers as well as active promoters to a lesser extent (Whyte et al. 2012). The second set contained H3K4me3 and H3K79me2 in addition to total and serine 5-phosphorylated RNA Polymerase II (Pol II ser5p). While H3K4me3 is a marker of active promoters and tends to accumulate most sharply over the first 5’ splice site, H3K79me2 has been shown to mark the first intron of active genes.
and may be a general indicator of the initiation of productive elongation (Huff et al. 2010). Pol II ser5p, on the other hand, is thought to accumulate at a pause site soon after transcription initiation and prior to productive elongation (Rahl et al. 2010). Together, this cluster most likely represented active promoters committed to entering productive elongation. FHFOX2 peaks also strongly anticorrelated with repressive marks such as H3K27me3 and components of the repressive PRC2 complex, Ezh2, Suz12, and Ring1b. When the same binning and correlation analysis was carried out for U2AF65, the enhancer/promoter signal was less prominent than for FHFOX2. A cluster characterized by H3K36me3 and Pol II ser2p was also evident, suggesting a distinct localization pattern more closely coupled with transcriptional elongation (Supplemental Fig.S1B,C).

The above analysis argued that the majority of Rbfox2 did not appear to be localizing to regions associated with splicing regulation, namely introns and exons internal to the gene body, but was instead being recruited to active genes and enhancers early in the transcription cycle, prior to transcriptional elongation. To determine whether FHFOX2 ChIP signal was accumulating near TSS and transcriptional enhancers, we generated metagene plots aligning summits of FHFOX2 ChIP peaks either to UCSC-annotated TSS or to intergenic or intragenic transcriptional enhancers, defined by the presence of H3K4me1 and H3K27ac and the absence of H3K4me3 (Fig.2B). FHFOX2 ChIP peaks were strongly enriched around TSS and showed roughly 1/5 as much signal at intergenic and genic enhancers, lending support to the model that recruitment was occurring at sites of transcription initiation. In contrast to this distribution and in support of the overlap with Pol II ser2p, U2AF65 peaks accumulated most strongly downstream of transcription termination sites as opposed to around transcription start sites and enhancers (Supplemental Fig.S1D).
Chapter 4. Co-transcriptional splicing regulation by Rbfox2

Rbfox2 interacts with Pol II

We next took a complementary approach to specifically determine which activating marks were most likely to be responsible for defining FHFOX2-bound genomic regions. Subsetting FHFOX2 peaks into TSS-proximal, intronic, or intergenic categories, coverage of various histone marks and transcriptional machinery was plotted around the summits of FHFOX2 peaks in each category (Fig. 3A). In all three cases, the most highly enriched factor analyzed was Pol II ser5p, followed closely by total Pol II. Consistent with the enhancer signature seen earlier, both intronic and intergenic FHFOX2 peaks were flanked by H3K4me1 and H3K27ac and showed little H3K4me3 accumulation.

Taken cumulatively, our results most parsimoniously argue that Rbfox2 is likely not directly recruited to specific sites in the chromatin through interaction with particular histone marks. Instead, Rbfox2 may translocate with the active polymerase from the point of transcription initiation until being unloaded onto the pre-mRNA upon emergence of a functional consensus motif. This would require an interaction with Pol II, either directly or through a protein or RNA intermediate. To test this hypothesis, we performed co-immunoprecipitations using an antibody against total Pol II and assayed for pull-down of FHFOX2. Indeed, both in the absence and presence of the general nuclease Benzonase to eliminate DNA- or RNA-dependent interactions, a small yet reproducible fraction of FHFOX2 co-immunoprecipitated with Pol II (Fig. 3B). To further validate the RNA independence of this interaction, we generated a FHFOX2 construct with a point mutation in the RRM that decreased RNA affinity by >1000 fold (FHFOX2 F126A) (Auweter et al. 2006). The interaction with Pol II persisted under these conditions, demonstrating that Rbfox2 was associating with Pol II not through the pre-mRNA intermediate but through direct contact or a protein partner (Fig. 3C).
Chapter 4. Co-transcriptional splicing regulation by Rbfox2

**Pol II pausing deposits Rbfox2 at sites of splicing regulation**

Previous evidence in yeast and mammals has demonstrated that polymerase pausing, marked by ser5p, can occur within the gene body and may be coupled to splicing (Batsche et al. 2006; Alexander et al. 2010). In the case of the human CD44 gene, Muchardt and colleagues argued that chromatin remodeling by the SWI/SNF component Brm mediates Pol II pausing. Association of U1 and U5 components with Brm, as well as kinetic favorability of slowed transcription for recognition of weak splice sites, were hypothesized to be responsible for increased alternative exon inclusion in the presence of Brm (Batsche et al. 2006). This group also demonstrated that presence of the H3K9me3 mark within the region encoding the CD44 variable region results in decreased Pol II elongation rates and increased exon inclusion (Saint-Andre et al. 2011). In an analogous manner, we predicted that Pol II ser5p pause sites within gene bodies might allow for a prolonged accumulation of Pol II-associated splicing components, thus leading to more efficient splice site recognition in line with the kinetic model of co-transcriptional splicing. Intersection of FHFOX2 ChIP peaks, Pol II ser5p ChIP peaks, and annotated cassette exon loci from MISO revealed a strong overlap between the three categories, providing scope for this mode of regulation (Fig.4A). Plotting ChIP signal around the set of concordant Rbfox2-dependent splicing changes defined previously (Jangi et al. in review), we detected no obvious trend suggesting a positional relationship between FHFOX2 ChIP peaks and the direction of Rbfox2-dependent splicing regulation. However, manual inspection of individual cassette loci with overlapping FHFOX2 and Pol II ser5p ChIP peaks indicated that this mechanism may be active and detectable at lower resolution at a subset of Rbfox2-responsive splicing events. We examined four example loci, Hnrnpa1, Src, Srsf3, and Fam113b, showing intermediate to high levels of FHFOX2 and increased local density of Pol II ser5p (Fig.
4B-E). Importantly, FHFOX2 iCLIP signal that we previously generated (Jangi et al. in review) also exhibited higher depth within these loci; iCLIP signal near Srsf3 and Hnrnpa1 in particular was strongly coincident with FHFOX2 and Pol II ser5p ChIP peaks (Fig. 4B-E), suggesting increased deposition at Rbfox2 binding sites on the nascent transcript due to higher local concentration of Rbfox2 recruited by the pausing polymerase.

**Modulation of histone acetylation alters Rbfox2-dependent splicing**

As our results suggested that Rbfox2 was co-transcriptionally tracking with Pol II, we reasoned that modulating Pol II elongation could change local concentrations of Rbfox2 and alter the splicing of Rbfox2 target exons. The histone deacetylase (Hdac) inhibitor sodium butyrate has been demonstrated to have varying effects on transcription elongation. A subset of genes show attenuated gene expression upon Hdac inhibition due to decreased Pol II elongation rates (Heruth et al. 1993; Rada-Iglesias et al. 2007; Daroqui and Augenlicht 2010). Another study showed by exon array that for the subset of genes that increases in expression upon Hdac inhibition, exon inclusion decreases, while genes that are unchanged or decrease in expression exhibit higher levels of exon inclusion, arguing for gene-specific effects on Pol II processivity (Hnilicova et al. 2011).

To query the effect of altered chromatin states on Rbfox2-dependent splicing, we treated mESC with sodium butyrate for 8h and measured by RT-PCR the splicing of seven Rbfox2-dependent cassette exons identified in our RNAseq analysis (Jangi et al. in review), four of which show Rbfox2-dependent exon inclusion (Fig.5A) and three showing Rbfox2-dependent exon skipping (Fig.5B). These exons were also chosen for the presence of transcriptional enhancer marks within the flanking introns in at least one of nine mouse tissues examined from
Chapter 4. Co-transcriptional splicing regulation by Rbfox2

the ENCODE project (Encode Project Consortium 2011), suggesting that these may be dynamically regulated by chromatin-dependent mechanisms. For three of the seven exons, treatment with sodium butyrate led to further splicing changes in the direction of increased Rbfox2 activity, while the remaining four exons were unchanged. Specifically, Rbfox2 enhances the inclusion of an exon within the gene Myo9b, and treatment with sodium butyrate led to increased exon inclusion (Fig.5A). As treatment with a hairpin against Rbfox2 did not result in complete Rbfox2 knockdown (~10-20% protein levels remaining), exon inclusion was also increased upon sodium butyrate treatment in the Rbfox2 knockdown compared to control knockdown. Conversely, Rbfox2 represses inclusion of cassette exons in Numb and Smtn, and sodium butyrate treatment resulted in a pronounced decrease in exon inclusion in control and to a lesser extent in Rbfox2 knockdown (Fig.5B). While determining the mechanism of altered splicing will require further investigation, these observations are consistent with increased Rbfox2-dependent splicing activity due to Hdac inhibition. This may be caused by decreased Pol II elongation rates, allowing Rbfox2 to exert its effects on splicing more effectively.

4.4 DISCUSSION

Here, we sought to investigate mechanisms of co-transcriptional splicing mediated by Rbfox2. We performed chromatin immunoprecipitation for endogenous and epitope-tagged Rbfox2 and found it localizes to thousands of transcription start sites and enhancers across the genome. Metagene and correlation analysis with ChIP datasets of various histone marks and transcriptional machinery revealed that the sites to which Rbfox2 is recruited are transcriptionally active and dominated by a Pol II ser5p signature. In agreement with this observation, Rbfox2 co-immunoprecipitates with Pol II in an RNA-independent manner. Most
notably, the pattern of chromatin binding is distinct from the pattern of RNA binding we previously reported (Jangi et al. in review). Rbfox2 is not specifically enriched at the promoters of genes whose splicing is Rbfox2-dependent but instead is generally recruited upon transcription initiation. Specific RNA binding to regulated loci may occur during transcriptional elongation when accessible Rbfox2 consensus motifs are transcribed. Modulation of histone acetylation results in increased Rbfox2-dependent splicing activity, suggestive of decreased Pol II elongation rates.

Our results support a model in which Rbfox2, perhaps with other constitutive and alternative splicing regulators, is indiscriminately recruited to active transcription start sites and enhancers prior to recognition of its consensus UGCAUG motif in the nascent transcript. Evidence for this type of early recruitment has been observed for constitutive components of the U1 snRNP, which were also shown to feed back into transcription to enhance productive elongation (Das et al. 2007). The SR proteins SC35 and SF2/ASF were also recently shown to accumulate at sites of transcription initiation, indicating that this may be a common trait of much of the spliceosomal machinery (Ji et al. 2013). Such a model is distinct from the recruitment model suggested to be at play for PTB and the U2 snRNP, in which particular patterns of histone marks within gene bodies may directly bind splicing components to activate a locus for splicing (Luco et al. 2010). In the case of Rbfox2, we speculate that co-transcriptional tracking with Pol II would serve to increase the local concentration of Rbfox2 to enhance the recognition of nascently-transcribed UGCAUG motifs and thus activate splicing. This model is diagrammed in Figure 6.

An important component of this model is the kinetic coupling between transcription and splicing. We found that Rbfox2 accumulations within gene bodies as well as at promoters are
accompanied by accumulations of Pol II ser5p, suggestive of intragenic transcriptional pause sites (Alexander et al. 2010). In this case, pausing of Pol II by any number of stimuli would result in an increased concentration of Rbfox2 in the vicinity of the pause site and an increased likelihood of Rbfox2-dependent splice site recognition. Histone deacetylase inhibition has been suggested to induce pauses after transcription initiation and decrease Pol II elongation rates (Heruth et al. 1993; Daroqui and Augenlicht 2010), which may explain the increased Rbfox2-dependent splicing activity seen in our study upon sodium butyrate treatment. Many of the intragenic pause sites we detected also bear H3K4me1 and H3K27ac marks indicative of transcriptional enhancers, which were recently shown to be enriched for nearby alternatively-spliced exons (Mercer et al. 2013). While our observations do not exclude the possibility that Rbfox2 is directly recruited to the chromatin at transcriptional enhancers, the biochemical interaction and overlapping ChIP signal with Pol II is more directly explained by recruitment to the transcriptional machinery, followed by kinetic coupling. However, further investigations will be necessary to conclusively distinguish between Pol II-dependent and –independent effects of Hdac inhibition on Rbfox2 activity.

Increasing evidence in support of the kinetic model of splicing raises the possibility that altering promoter activity or strength and transcription elongation rates could lead to the regulation of constitutive and alternative splicing. Early observations that changes in promoter strength of the Fn1 gene lead to changes in splicing of exon 33 indicate long-range impacts of the promoter on co-transcriptional splicing (Cramer et al. 1997; Cramer et al. 1999). One model that has been proposed to explain this observation is the differential recruitment of splicing machinery to different promoters. Our results and those of others (Ji et al. 2013) would argue against this, as Rbfox2 promoter recruitment is not limited to a subset of active promoters.
However, this does not preclude other splicing factors from being recruited only to specific promoters. Focusing on shorter range effects of transcription rates, Batsche et al argued that removing polymerase pause sites by knocking down the SWI/SNF chromatin remodeling complex component Brm alters variable domain splicing of the cell surface co-receptor CD44 (Batsche et al. 2006). More directly, the introduction of MAZ elements to induce transcriptional pausing around the Rbfox2-regulated cassette exon in the Src gene leads to increases in exon inclusion (Roberts et al. 1998).

The role of intragenic enhancers in gene regulation remains unclear. One recent study suggested that intragenic promoter-like modifications are enriched around alternative exons and may arise from three-dimensional interactions between promoters and intragenic regulatory sites (Mercer et al. 2013). The possibility that these interactions may regulate transcription kinetics could implicate cell type-specific enhancer marks in splicing regulation (Mercer et al. 2013). We have observed several instances of alternative exon-proximal transcriptional enhancers hinting at a complex regulatory function. For example, the Pkm gene contains a mutually exclusive splicing event that results in a switch in glucose metabolism depending on whether the Pkm1 or Pkm2 isoform is produced (Christofk et al. 2008; Anastasiou et al. 2012). The distribution of iCLIP signal as well as splicing changes upon Rbfox2 knockdown support Rbfox2 repression of the upstream mutually exclusive exon, resulting in Rbfox2-dependent production of the m2 isoform. However, in brain and muscle where Rbfox proteins are highly expressed, the predominant isoform is m1. Intriguingly, the Rbfox2-bound region is also specifically marked with a transcriptional enhancer in brain and muscle, suggesting an additional layer of regulation that either alters Rbfox2 activity or changes splicing regulation of Pkm through other means.
Future studies will be necessary to determine whether changes in chromatin structure mediated, for example, by cell type-specific intragenic enhancers affect Rbfox2 occupancy or transcriptional kinetics. Our results provide additional support for the concept that the regulation of alternative splicing by polymerase pausing, either through transcriptional enhancers, promoter effects, or other histone marks, could have widespread effects on the establishment and maintenance of splicing networks.

### 4.5 MATERIALS AND METHODS

**Cell fractionation**

Two hours prior to harvesting, control V6.5 mESC were treated with 1uM flavopiridol for 1h, then allowed to recover in fresh media without drug for 1h. Treatment cells were treated with 1uM flavopiridol for 1h prior to harvesting. Cytoplasmic, nucleoplasmic, and chromatin fractions were purified from cells as previously described (Wysocka et al. 2001). 100ug/ml of RNase A was added to nuclear fraction during the 30min incubation at room temperature to separate the soluble nucleoplasmic compartment. Cell equivalents of each fraction were loaded on a 10% Tris-glycine SDS-PAGE gel, transferred to PVDF, and blotted for splicing factors.

**ChIP library preparation**

Mouse V6.5 ES cells were cultured as previously described. Cells were infected with lentivirus expressing doxycycline-inducible untagged or FLAG-HA-tagged human RBFOX2 (UntagFOX2, FHFOX2) and selected in 150ug/ml hygromycin for 6 days. Uninfected V6.5 cells or FHFOX2 and UntagFOX2 cells induced for 24 hours with 1ug/ml doxycycline were crosslinked with 1% formaldehyde solution, collected, and flash frozen. Cells were lysed in modified RIPA (50mM

179
Tris pH 8.0, 140mM NaCl, 1mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% Na-deoxycholate), sonicated, and immunoprecipitated overnight with HA, Rbfox2, or U2AF65 antibody pre-conjugated to magnetic Protein G beads. Washed and eluted immunoprecipitates were subjected to RNase A and proteinase K treatment and purified by phenol-chloroform extraction. End-repaired, A-tailed chromatin fragments were ligated to Illumina-compatible barcoded adapters and PCR-amplified for 18 cycles. Fragments 300-500nt in length were purified from a 2% agarose gel. Samples were pooled and sequenced on the Illumina HiSeq platform.

**Co-immunoprecipitation**

Cells expressing FHFOX2, FHF12A, or UntagFOX2 were lysed in Nuclear Lysis Buffer (1% NP-40, 125mM KCl, 1mM EDTA, 20mM HEPES pH 7.9) in the presence or absence of Benzonase. Lysates were immunoprecipitated overnight with Pol II N20 antibody or IgG conjugated to magnetic Protein A beads, washed, and visualized by western blot as described above.

**Histone deacetylase inhibition and splicing analysis**

Mouse V6.5 ES cells were infected with lentivirus expressing shRNA targeting control (shLuciferase) or Rbfox2 (shFox2). After 3 days of selection in 1ug/ml puromycin, cells were treated for 8h with vehicle (HBS) or 5mM sodium butyrate to inhibit HDAC activity. RNA was collected using Trizol, reverse transcribed, and amplified for 35 cycles using primers flanking alternative exons. Spliced products were resolved on a 2% ethidium bromide gel, scanned on a Typhoon imager and quantified using ImageQuant software.
Chapter 4. Co-transcriptional splicing regulation by Rbfox2

Metagene analysis

BedTools and custom Python scripts were used to quantify ChIP coverage with respect to various genomic features.

4.6 REFERENCES

Chapter 4. Co-transcriptional splicing regulation by Rbfox2


Chapter 4. Co-transcriptional splicing regulation by Rbfox2


Chapter 4. Co-transcriptional splicing regulation by Rbfox2


Figure 1. Rbfox2 associates with chromatin.
A) mESC were fractionated into cytoplasm, nucleoplasm, and chromatin after transcriptional inhibition with flavopiridol, or inhibition followed by recovery. RNA dependency was determined by RNase A treatment of the nuclear pellet. Shown is Western blot of splicing components in each condition for cell equivalents of each subcellular fraction. B) mESC expressing FLAG-HA-tagged human RBFOX2 were formaldehyde crosslinked and prepared for ChIP. After immunoprecipitation with anti-HA antibody, input, flowthrough, and IP were resolved by SDS-PAGE and probed with the indicated antibodies. C) Distribution of FHFOX2 ChIP summits across various genomic categories, not normalized for the nucleotide length of each category. D) Enrichment of FHFOX2 ChIP summits in genomic categories over a shuffled control distribution. P-values determined using Chi-square test.
Chapter 4. Co-transcriptional splicing regulation by Rbfox2
Figure 2. Rbfox2 chromatin association is enriched around promoters and transcriptional enhancers.
A) Cross-correlation of histone marks or chromatin factors at Rbfox2-bound regions. Boxes represent closely clustered sets of histone marks or chromatin factors. Top row and left column represent fraction of FHFOX2 peaks overlapping other histone marks or chromatin factors. B) Metagene plots showing distances from summits of Rbfox2 peaks to transcription start sites (TSS), intragenic enhancers, or genic enhancers, averaged using a sliding window of 100nt. Enhancers defined by overlapping H3K4me1 and H3K27ac peaks, with no H3K4me3 peak. Negative values indicate 5’ with respect to the gene, positive values indicate 3’.
Figure 3. Rbfox2 interacts with RNA polymerase II.
A) Metagene plots showing normalized enrichment of proteins and histone marks around FHFOX2 ChIP summits located within 1kb of the TSS, within introns, or in intergenic regions, averaged using a sliding window of 100nt. B) Immuno precipitations of total Pol II or control IgG in cells expressing FHFOX2, with or without nuclease treatment (Benzonase), shows specific co-immunoprecipitation of FHFOX2 with Pol II. C) Co-immunoprecipitation with Pol II is detectable using a mutant FHFOX2 construct with strongly diminished RNA binding affinity.
Chapter 4. Co-transcriptional splicing regulation by Rhfox2
Figure 4. A subset of Rbfox2 chromatin interactions occur around cassette exons and are associated with Pol II ser5p.

A) Venn diagram showing the intersection of FHFOX2 ChIP peaks, Pol II ser5p ChIP peaks, and cassette exons and their flanking introns. B-D) Integrative Genome Viewer screenshots of Rbfox2-regulated exons with coincident FHFOX2 and Pol II ser5p ChIP signal.
Chapter 4. Co-transcriptional splicing regulation by Rhfox2
Figure 5. Histone deacetylase inhibition increases Rbfox2-dependent splicing activity.

A,B) cDNA generated from mESC treated with 5mM sodium butyrate (NaB) or vehicle (HBS) for 8h was amplified with primers in constitutive exons flanking Rbfox2-enhanced (A) or Rbfox2-repressed (B) cassette exons. Quantitation of average percent inclusion ($\Psi$) from two biological replicates is shown below a representative image of PCR products resolved on an EtBr-stained agarose gel.
Chapter 4. Co-transcriptional splicing regulation by Rbfox2
Figure 6. Model for co-transcriptional splicing regulation by Rbfox2.
Schematic showing Pol II transcription of a gene containing an Rbfox2-regulated exon (shown in red). As Pol II (shown in green) initiates transcription, it is loaded with Rbfox2 (orange) and other transcriptional and splicing components not shown here. Fast elongation rates and additional signals may prevent UGCAUG motifs from being recognized during transcription, preventing the inclusion of nearby alternative exons (dark green exon). Upon encountering signals such as histone modifications (histones shown in gray, modifications as small colored circles), Pol II elongation rates may decrease, resulting in preferential recognition of UGCAUG motifs around the regulated exon (shown in red) and increased exon inclusion in the final transcript.
Chapter 4. Co-transcriptional splicing regulation by Rbfox2
Supplemental Figure S1. U2AF65 ChIP signal coincides with elongating Pol II and is enriched at transcription termination sites.

A) mESC were formaldehyde crosslinked, prepared for ChIP, and immunoprecipitated with two separate antibodies recognizing endogenous U2AF65. Western blot shows enrichment of U2AF65 in the immunoprecipitate. All following analysis shown is from the ChIPseq library generated using the MC3 monoclonal antibody. B) Distribution of U2AF65 ChIP summits across various genomic categories, not normalized (left panel) or normalized (right panel) for the nucleotide length of each category. C) Cross-correlation of histone marks or chromatin factors at U2AF65-bound regions. B) Metagene plots showing distances from summits of U2AF65 peaks to transcription start sites (TSS), intragenic enhancers, genic enhancers, or transcription termination sites (TES), averaged using a sliding window of 100nt. Enhancers defined by overlapping H3K4me1 and H3K27ac peaks, with no H3K4me3 peak. Negative values indicate 5' with respect to the gene, positive values indicate 3'.
Chapter 4. Co-transcriptional splicing regulation by Rhfox2
Supplemental Figure S2. Enhancer-marked alternative exons show Rbfox2-dependent splicing changes.

A-C) Sashimi plots of cassette exons in shLuc and shFox2-1 show Rbfox2-dependent splicing regulation (top two tracks). Integrative Genome Viewer screenshots showing FHFOX2 iCLIP, Pol II ser5p, H3K27ac, H3K4me1, FHFOX2 ChIP.
signal and FHFOX2, Pol II ser5p, H3K27ac, and H3K4me1 ChIP signal in the Rbfox2-regulated loci.
Chapter 5

Conclusions and Future Perspectives

The era of genomics has ushered in a new perspective on fundamental questions in biology and medicine. Integrative analyses between studies of transcription, RNA processing, and translation using biochemical and computational approaches will be key in understanding complex regulatory networks. This thesis has described the characterization of multiple regulatory layers that together establish one such network, the Rbfox2 splicing network in mouse embryonic stem cells (mESC), providing important groundwork for investigating systems-level regulation of splicing by other tissue-specific RNA binding proteins.

5.1 Summary

In Chapter 2, we identified thousands of in vivo Rbfox2 binding sites and hundreds of Rbfox2-regulated splicing events across the mESC transcriptome using iCLIP and RNAseq. Because of the parallel nature of these approaches, we were able to distinguish direct from indirect splicing changes arising from Rbfox2 loss. Importantly, this allowed us to discover a large class of Rbfox2-regulated splicing events that are rendered silent by nonsense-mediated mRNA decay (NMD). Such targets would not have been identifiable relying only on splicing
changes inferred from RNAseq data, as only the stable isoform is detectable, while the other is largely degraded due to the splicing-dependent introduction of a premature termination codon. A significant fraction of these NMD splicing events fall in the category of conserved autoregulatory splicing events within RNA binding proteins (RBPs) that have been previously described (Lareau et al. 2007; Ni et al. 2007). We found that cross-regulation of this autoregulatory splicing by Rbfox2 serves to set the efficacy of autoregulation, and thus the steady-state expression of the gene. This is not limited only to RBPs; we found evidence of Rbfox2 regulating splicing-coupled NMD and gene expression of hundreds of other genes including several kinases and ubiquitin ligases. We proposed that Rbfox2, and most likely many other splicing factors, tunes the autoregulation and gene expression of a network of RBPs, and that changes in Rbfox2 expression such as during development activate direct Rbfox2-dependent splicing changes as well as a cascade of secondary splicing changes. These events, together, define the complete Rbfox2 splicing regulatory network. This analysis thus highlights how the integration of high-throughput techniques reveals general principles of gene regulation that can be extended to other splicing regulators in different contexts.

Another critical aspect in defining splicing factor networks is the identification of RNA motifs recognized by particular RBPs. Chapter 3 described the method and applications of RNA Bind-n-Seq (RBNS), a novel technique for discovering sequence and structural specificity of RBPs in vitro. A more quantitative and comprehensive approach to motif discovery than many existing techniques, RBNS can be combined with in vivo techniques to infer principles of binding specificity. RBNS was used here to quantitatively identify the spectrum of motifs bound by Rbfox2, Mbnl, and Celf/Cugbp1, three splicing factors that may function together in many physiological contexts, particularly cardiac development (Kalsotra et al. 2008). In addition to the
UGCAUG motif we identified by iCLIP for Rbfox2 in Chapter 2, we found that a secondary motif, GCACG, is functional in Rbfox2-dependent splicing regulation. While motifs derived from RBNS and iCLIP/CLIPseq correlate well for all three factors analyzed, we demonstrated a role for RBNS in distinguishing direct protein-RNA interactions from noisy or complex multi-protein interactions with RNA that may be detected by in vivo techniques such as iCLIP/CLIPseq. Analysis of regions bound exclusively in vivo, once corrected for biases in the CLIP technique, may reveal novel insights into the determinants of motif-independent binding, such as cooperativity and multimerization of splicing factors or, as discussed in Chapter 4, chromatin interactions.

Much recent work in the splicing field has focused on the integration of chromatin dynamics and splicing regulation. In Chapter 4, we described our investigation into the mechanisms of co-transcriptional splicing regulation by Rbfox2. We found that Rbfox2 is recruited early in transcription to the majority of active genes, as well as a large number of transcriptional enhancers. Due to an interaction with RNA Pol II, we proposed that Rbfox2 is likely to co-transcriptionally track with the transcriptional machinery to be deposited at regulatory sites. Indeed, we found that modulation of histone acetylation, which has been reported to affect transcriptional elongation (Hnilicova et al. 2011), results in altered splicing of Rbfox2 target exons. While demonstrating the potential of the chromatin environment to impact regulated splicing, this study raises many further questions regarding the specific mechanisms of co-transcriptional splicing by Rbfox2 and other splicing factors. In particular, the regulation of Pol II pausing through chromatin modifications will be a crucial component in understanding the kinetic coupling of transcription and splicing.
5.2 Mechanistic insights from global splicing analyses

In the past five years, the development of techniques to assay protein-RNA interactions on a transcriptome-wide scale in vivo has provided a novel angle from which to study splicing regulation. In particular, comparison of iCLIP/CLIPseq datasets for several splicing factors suggests the existence of general principles by which RNA binding proteins mediate exon activation or repression.

Our results, along with those of others (Ule et al. 2006; Xue et al. 2009; Huelga et al. 2012) have demonstrated that motif-dependent binding upstream of the alternative exon tends to repress its inclusion, while binding downstream tends to enhance exon inclusion. It remains an open question in the field why exon repression is generally mediated by inhibiting binding of U2 components and exon activation by enhancing binding of the U1 snRNP. As the interaction of U1 with the 5’ splice site is a much more transient interaction than that of the U2 snRNP (Hoskins et al. 2011), it is possible that strengthening this interaction increases the likelihood of commitment to splicing, while obstructing the stronger U2 interaction is necessary to prevent commitment. This implies that splicing factors that are able to both enhance and repress exon inclusion must be able to stabilize the U1 snRNP, either through direct binding or through interactions with other RNA binding proteins known to bind U1, such as TIA1 (Forch et al. 2002). In the case of Rbfox2, a documented interaction with the U1 component U1C likely explains exon activation when bound downstream of the alternative exon (Huang et al. 2012).

We also observed, however, that binding of Rbfox2 to distal UGCAUG motifs near the downstream 3’ splice site correlates with enhanced exon inclusion. For a subset of exons, this was recently reported to occur through the formation of hairpin structures that bring the distal motifs proximal to the regulated exon (Lovci et al. 2013). Alternatively, considering the kinetics
of co-transcriptional splicing, it is possible that binding to the downstream constitutive 3’ splice site delays its recognition and allows time for the alternative exon to be defined, in effect co-opting the mechanism used by Rbfox2 to repress alternative exon inclusion for exon activation. Analogous splicing regulation through binding flanking constitutive splice sites has also been reported for Nova and may thus represent a general principle of splicing regulator activity (Ule et al. 2006).

Analysis of our iCLIP data also revealed a large number of apparently constitutive splicing events that are bound by Rbfox2 in mESC. These events are defined as constitutive by the condition that junction reads indicative of alternative splicing were never detected in mESC at steady-state, upon Rbfox2 or Upf1 knockdown, or after cycloheximide treatment. It remains formally possible that these events are alternatively spliced in other conditions, such as in other cell types, and are bound but not alternatively spliced due to the absence of additional factors. This is consistent with the observation that constitutive exons bound by Rbfox2 have weaker splice sites than unbound constitutive exons. However, it may also be the case that these exons are in fact regulated by Rbfox2 in mESC, but the spliced products are not detectable. In light of our results detailed in Chapter 2, one possibility to explain the lack of detectable alternative splicing is that missplicing in the absence of Rbfox2 results in degradation by NMD. It will be of interest to determine computationally whether these “constitutive” exons do indeed generate NMD substrates when excluded, and experimentally, whether Rbfox2 regulates their proper splicing.

Rbfox consensus motifs are frequently clustered at closer distances than would be expected by chance. Our iCLIP results confirmed binding to many examples of these clustered motifs, and this gives ground to speculate that as with other RNA binding proteins such as
hnRNPs (Xue et al. 2009), Rbfox2 binding may be cooperative. Using the RBNS technique introduced in Chapter 3, it will be possible to directly assess the cooperativity of Rbfox2 binding \textit{in vitro}. In addition, we found motifs for Celf/Cugbp1 and hnRNP L present in close proximity to Rbfox2 motifs. These factors may be functioning cooperatively, or perhaps because of the short distances between the motifs, may be competing for binding at a particular site. It has been previously observed that the consensus motifs for these factors are enriched around exons alternatively spliced during cardiac development (Kalsotra et al. 2008). As the expression patterns of Rbfox2 and Celf/Cugbp1 change across this transition, it is of importance to determine at what stages these factors are co-expressed to be able to conclude whether they are acting competitively or cooperatively. In either case, it is likely that the regulation of individual splicing events by multiple RNA binding proteins allows for sharper and more robust developmental transitions.

5.3 \textbf{Physiological impact of splicing-coupled NMD}

The results described in Chapter 2 revealed that the scope of splicing-coupled NMD (AS-NMD) may be more broad than previously appreciated. While numerous studies in mouse and human have demonstrated the critical role of NMD in organismal development (Hwang and Maquat 2011 and references therein), the contribution of AS-NMD to these phenotypes remains difficult to parse. A technical hurdle to addressing this question is that transcriptome-wide approaches such as RNAseq underestimate the quantitation of NMD isoforms due to their relatively short half-lives. Furthermore, inhibition of NMD may be insufficient to capture the complete pool of AS-NMD substrates due to active repression of particular NMD isoforms by RNA binding proteins such as Rbfox2. To determine the extent to which any particular splicing
factor relies on AS-NMD to control its splicing and gene expression networks, it will be necessary to perform global transcriptome analyses upon simultaneous inhibition of NMD and depletion of the RBP. In the case of Rbfox2, we predict that NMD inhibition and Rbfox2 knockdown together will unmask the silent splicing events we identified by iCLIP, allowing a quantitation of the effect on a genome-wide scale. Presumably, the widespread regulation of AS-NMD is not specific to Rbfox2 and may be a common feature of many RBPs. Future work will be needed to determine the AS-NMD networks of tissue-specific splicing regulators and the contexts in which they function.

Beyond its roles in maintaining steady-state protein expression, AS-NMD may be particularly relevant in transient physiological states, for example during responses to stress signals. As the decision to splice a stable isoform or an NMD isoform can be regarded as a post-transcriptional regulatory choice, AS-NMD may be ideal for achieving rapid differential responses to external cues. In support of this, a significant fraction of genes that are differentially spliced upon UV-induced DNA damage encode the NMD isoforms of RNA binding proteins (Ip et al. 2011). Similarly, a pool of partially-spliced transcripts encoding the Clk1/4 SR protein kinases undergoes a stress-dependent maturation of the stable isoform to produce active Clk1/4 protein, and homeostatic levels of the kinases are eventually reestablished through the production of the NMD isoform due to SR protein phosphorylation (Ninomiya et al. 2011). These observations suggest that by controlling gene expression at the level of splicing as opposed to transcription, the cell may be able to more rapidly respond to stress signals and initiate a self-limiting cascade of transcriptome changes to return to homeostasis. It will be of interest to determine whether such phenomena are more widespread, for example within other stress responses. More importantly, examining whether there is a coherence to the ontology of
non-RBP genes containing NMD splicing events will be essential in identifying gene expression programs that are activated through context-specific AS-NMD. A systems approach to these questions, as has been taken to demonstrate autoregulation in transcriptional responses (Becskei and Serrano 2000), will further increase our understanding of the functional relevance of AS-NMD in various settings.

5.4 Chromatin as a marker and effector of regulated splicing

A major frontier in the study of splicing regulation is the coordination of splicing with other mechanisms of gene regulation. The relationship between splicing and other pre-mRNA processing pathways such as capping and polyadenylation has been well characterized through biochemical and molecular studies. Within the past five years, the role of chromatin in co-transcriptional splicing has become a new focus.

As described in Chapter 4, kinetic coupling of splicing and transcription or the direct recruitment of splicing machinery to the chromatin affords an additional regulatory layer in splicing (reviewed in Luco et al. 2011). Importantly, as chromatin marks are highly cell type-specific and are maintained across cell divisions, such a mechanism could play a role in maintaining stable splicing patterns within a particular cell type. The extreme example of Drosophila Dscam1 splicing, in which individual neurons express distinct subsets of 8-50 isoforms from a possible ~19,000 isoforms, is a case in which chromatin-dependent effects on splicing may be at play (Neves et al. 2004; Hattori et al. 2009). As each neuron expresses a relatively small number of isoforms to ensure dendritic self-avoidance mediated by homotypic interactions of the Dscam1 adhesion molecule, it is possible that particular subsets of exons exist in a chromatin state permissive for exon inclusion within each cell. The diversity of isoform
expression between cells is unlikely to be explained solely by differences in splicing factor expression, as individual cells of the same cell type can express different patterns of Dscam1 isoforms (Neves et al. 2004). Similar mechanisms of chromatin-dependent commitment to splicing patterns could be envisioned in other contexts. For example, splicing events in which the direction of splicing is inconsistent with the RNA binding protein repertoire of that cell type may be candidates for this type of regulation.

The recent publication of ChIPseq and RNAseq datasets from various human and mouse tissues and cell types as part of the ENCODE project will provide a valuable resource for computational investigations into the links between chromatin and splicing regulation (Encode Project Consortium 2011). However, it will be equally important to functionally validate the global trends that have thus far been reported to elucidate the mechanistic details of co-transcriptional splicing. In particular, distinguishing between effects on the kinetics of transcription and the recruitment of splicing machinery will be a primary question to address.

5.5 RNA binding proteins as master regulators of cell state

In less than a decade, the application of genome-wide RNA sequencing approaches to the study of splicing has revealed the ubiquity of alternative isoform production. Although technical limitations and the high cost of sequencing are still somewhat prohibitive, the wealth of information provided by each study has provided fertile ground for deciphering what has come to be known as the splicing code. Several groups have undertaken computational analyses that integrate RBP consensus motifs and other RNA features with transcriptomics and RNA-protein interaction maps to identify tissue-specific splicing parameters (Barash et al. 2010; Zhang et al. 2010; Lim et al. 2011). These splicing codes have made apparent that different subsets of
genomically-encoded cis elements are active in different tissues, suggesting that the balance of splicing factors expressed in a given cell type is a primary determinant of its splicing patterns. The further integration of experimental approaches such as CLIPseq, RNAseq, and motif discovery tools are becoming critical in applying the splicing code to different physiological contexts.

Studies in the definition of cell identity have revealed that the expression of a relatively small number of individual transcription factors, termed pioneer transcription factors, can initiate commitment along a diverse array of developmental lineages (Young 2011). Characterization of splicing changes across developmental transitions and in differentiated tissues and cell lines (reviewed in Blencowe et al. 2009), as well as RNA-protein interaction maps (reviewed in König et al. 2011) has raised the intriguing possibility that an analogous set of master splicing regulators exists. In this case, the up- or downregulation of an RNA binding protein, most likely one that is expressed in a tissue-restricted manner, could initiate a splicing network and downstream gene expression changes that lead to a stable change in cell identity. Such a cascade has been demonstrated in human embryonic stem cells, where expression of MBNL1 initiates differentiation along the mesodermal lineage, primarily due to altered splicing of the transcription factor FOXP1 (Han et al. 2013). Similarly, depletion of RBFOX2 or exogenous expression of the Epithelial Splicing Regulatory Protein 1 (ESRP1) in human mesenchymal breast cancer cells causes a partial reversion to an epithelial phenotype, suggesting that these factors may play critical opposing roles in defining epithelial and mesenchymal cell states (Shapiro et al. 2011). An examination of the expression patterns of lineage-restricted splicing factors will provide insights into the contexts in which they may be sufficient to induce a rewiring of splicing and gene expression networks.
Potential pioneer splicing factors such as MBNL are likely to be highly regulated to allow their splicing networks to be properly activated in the correct settings. Although at least one Rbfox protein is expressed in a large number of tissues, differential activity of each factor may allow the Rbfox network to be regulated by unique signals in each tissue. In the brain, Rbfox1 and Rbfox2 show non-redundant splicing activity despite being expressed in many of the same neuronal cell types (Gehman et al. 2012). This is consistent with different modes of regulation for each Rbfox protein. For example, we have found that Erk1/2 activation by growth factor stimulation results in rapid phosphorylation of Rbfox2 in many cell types (M. Jangi and P. Sharp, unpublished observations). If such post-translational modification results in differential localization or association with the splicing machinery, it is possible that the Rbfox2 splicing network could be toggled by growth factor signaling. In contrast, Rbfox1 localization and splicing activity is altered in response to neuronal depolarization (Lee et al. 2009), while no analogous change in localization has been demonstrated for Rbfox2. Thus, since each Rbfox protein may respond differently to extracellular signals and may have unique binding partners, they could potentially regulate non-overlapping splicing networks in each cell type. This would allow the Rbfox proteins to function as tissue-restricted splicing factors that act at key nodes in the establishment of gene expression networks.

While the above examples demonstrate that modulation of a single RNA binding protein can initiate cell differentiation, our results suggest that a large supporting cast of splicing factors is likely to also play crucial roles in establishing cell identity. Variations in the expression of Rbfox2, perhaps functioning here as a pioneer splicing factor, cause concomitant changes in the expression of other splicing factors, many of which belong to the ubiquitous SR and hnRNP protein families. This leads to indirect splicing changes comprising a significant fraction of the
Rbfox2 splicing network. Understanding the co-regulation of these secondary splicing changes, as well as other gene expression changes arising from AS-NMD, will be critical in determining the roles of both pioneer and supporting RNA binding proteins in establishing tissue transcriptomes.

The studies presented in this thesis support growing evidence in the field of splicing regulation that the RNA binding protein repertoire of a given cell type can play a significant role in maintaining cell state and coordinating developmental transitions. Here, we have demonstrated the contribution of co-transcriptional splicing and NMD in establishing the Rbfox2 splicing network in embryonic stem cells. It is likely that such mechanisms represent general principles in defining the interconnected regulatory networks of many other RBPs across mammalian development.

5.6 REFERENCES


Hwang J, Maquat LE. 2011. Nonsense-mediated mRNA decay (NMD) in animal embryogenesis: to die or not to die, that is the question. *Current opinion in genetics & development* **21**: 422-430.


Chapter 6. Biographical Note

Mohini R. Jangi
Massachusetts Institute of Technology, Department of Biology
Koch Institute for Integrative Cancer Research
77 Massachusetts Ave., 76-411, Cambridge, MA 02139
Office: 617-253-0265
Mobile: 508-333-0478
Email: mohini.jangi@gmail.com

Education
Massachusetts Institute of Technology
Ph.D., Department of Biology
2007-2013

Brandeis University
B.S. with High Honors, Department of Biology
Summa cum laude
2003-2007

Research Experience
Ph.D. Candidate, Department of Biology, MIT
Advisor: Phillip A. Sharp, Ph.D.
2008-present

Thesis: Multiple regulatory layers in the establishment of Rbfox2 splicing networks
• Conceptualized, designed, and executed experiments and computational analyses to elucidate mechanisms of Rbfox2 activity in mouse embryonic stem cells (mESC). Identified a network of RNA binding proteins regulated by Rbfox2 in normal and disease contexts and a position-specific functional association with alternatively spliced transcripts
• Investigated quantitative RNA-protein interactions in Rbfox2-mediated splicing regulation in collaboration with the laboratory of Christopher Burge (MIT)
• Established reproducible, improved protocols for single-nucleotide resolution crosslinking immunoprecipitation sequencing (CLIP-seq) and chromatin immunoprecipitation sequencing (ChIP-seq) of RNA binding proteins in collaboration with a postdoctoral fellow and graduate student
• Carried out integrative analysis using ENCODE mRNA and chromatin tissue and cell line data to quantitate the global impact of histone marks on alternative splicing regulation
• Examined the role of Rbfox2 in a cell culture model of epithelial-mesenchymal transition and identified Rbfox2-dependent alternative splicing events with putative functions in tumor cell migration and invasion

Research Assistant, Department of Biology, Brandeis University
Advisor: Bruce L. Goode, Ph.D.
2005–2007

Thesis: Identification of actin binding surfaces on the Crn1p β-propellor
• Performed a screen for functional residues in the yeast Coronin protein by alanine scanning mutagenesis
• Validated and quantified functional Coronin-actin interactions using biochemical actin binding assays and immunofluorescence microscopy

Skills and Techniques
Experimental: Cloning, RT-PCR, qPCR, RNA/DNA isolation, northern blots, SDS-PAGE, western blots, immunofluorescence microscopy, mammalian tissue culture, lentiviral infection, cell migration assays, embryonic stem cell differentiation, chromatin immunoprecipitation sequencing (ChIP-seq), RNA sequencing (RNA-seq), crosslinking immunoprecipitation sequencing (CLIP-seq)

Computational: Experienced with tools for RNA-seq/ChIP-seq/CLIP-seq analysis including Bowtie/Tophat/Cufflinks suite, samtools, BEDTools, MACS, MISO, DESeq, pathway analysis, Gene Set Enrichment Analysis; competent in Unix, Python, and R programming
Chapter 6. Biographical Note

**Teaching Experience**
Mentor for an undergraduate researcher, Sharp Lab, MIT  June 2012-present
Teaching assistant, Cell Biology (7.06), MIT  Fall 2010
Co-mentor for two undergraduate researchers, Sharp Lab, MIT  Summer 2009
Teaching assistant, Introductory Biology (7.013), MIT  Spring 2009
Tutor for high school geometry, trigonometry, and pre-calculus  2004-2005

**Honors and Awards**
David Koch Graduate Fellowship, MIT  2011-2012
Phi Beta Kappa, Brandeis University  2007
Dean’s List, Brandeis University  2003-2007
Presidential Scholarship, Brandeis University  2003-2007
CRC Press Freshman Chemistry Achievement Award, Brandeis University  2004
Salutatorian, Ashland High School, Ashland, MA  2003

**Publications**


*These authors contributed equally.


**Select Poster Presentations**


**References**
Phillip A. Sharp, Ph.D.
Institute Professor
Massachusetts Institute of Technology
77 Massachusetts Ave, 76-461A
Cambridge, MA 02139
T: 617-253-6421  F: 617-253-3867

Jacqueline A. Lees, Ph.D.
Associate Director
Koch Institute for Integrative Cancer Research
77 Massachusetts Ave, 76-461A
Cambridge, MA 02139
T: 617-252-1972  F: 617-253-9863

Christopher B. Burge, Ph.D.
Associate Professor
Massachusetts Institute of Technology
77 Massachusetts Ave., 68-271A
Cambridge, MA 02139
T: 617-258-5997  F:617-452-2936