# Understanding Regulation of mRNA by RNA Binding Proteins



by

Alexander De Jong Robertson

B.S., Stanford University (2008)

Submitted to the Graduate Program in Computational and Systems Biology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computational and Systems Biology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2014



Author .....
Graduate Program in Computational and Systems Biology
December 19th, 2013

Certified by .....
Christopher B. Burge
Professor
Thesis Supervisor

Accepted by .....
Christopher B. Burge
Computational and Systems Biology Ph.D. Program Director

# Understanding Regulation of mRNA by RNA Binding Proteins

by

Alexander De Jong Robertson

Submitted to the Graduate Program in Computational and Systems Biology
on December 19th, 2013, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computational and Systems Biology

## Abstract

Posttranscriptional regulation of mRNA by RNA-binding proteins plays key roles in regulating the transcriptome over the course of development, between tissues and in disease states. The specific interactions between mRNA and protein are controlled by the proteins' inherent affinities for different RNA sequences as well as other features such as translation and RNA structure which affect the accessibility of mRNA. The stabilities of mRNA transcripts are regulated by nonsense-mediated mRNA decay (NMD), a quality control degradation pathway. In this thesis, I present a novel method for high throughput characterization of the binding affinities of proteins for mRNA sequences and an integrative analysis of NMD using deep sequencing data. This thesis describes RNA Bind-n-Seq (RBNS), which comprehensively characterizes the sequence and structural specificity of RNA binding proteins (RBPs), and application to the developmentally-regulated splicing factors RBFOX2, MBNL1 and CELF1/CUGBP1. For each factor, the canonical motifs are recovered as well as additional near-optimal binding motifs. RNA secondary structure inhibits binding of RBFOX2 and CELF1, while MBNL1 favors unpaired Us but tolerates C/G pairing in UGC-containing motifs. In a project investigating how NMD shapes the embryonic transcriptome, this thesis presents integrated genome-wide analyses of UPF1 binding locations, NMD-regulated gene expression, and translation in murine embryonic stem cells (mESCs). Over 200 direct UPF1 binding targets are identified using crosslinking/immunoprecipitation-sequencing (CLIP-seq). Results from ribosome foot printing show that actively translated upstream open reading frames (uORFs) are enriched in transcription factor mRNAs and predict mRNA repression by NMD, while poorly translated mRNAs escape repression.

Thesis Supervisor: Christopher B. Burge
Title: Professor

# Acknowledgments

There are a great number of people who helped and guided me on my journey and to whom I am indebted. First, my advisor, Christopher Burge gave me the time to explore and learn and who demonstrated how to do clear scientific research. The Burge lab was an excellent environment to working and it was a privilege to be there. I owe thanks to my thesis committee members, Phillip Sharp and Wendy Gilbert who gave me insight and inspiration throughout my time at MIT and who were extremely supportive. Their advice has proved invaluable. My collaborators on the projects in this thesis helped me enormously. My thanks go out especially to Jessica Hurt and Nicole Lambert, who I worked with closely and who each taught me much about wet lab science and RNA biology. Members of the Burge lab, past and present, have created a great community to learn in. Interactions with lab members on a day to day basis were one of the best parts of my time here. There were a few teachers and mentors that I was fortunate enough to work with prior to coming to MIT that I would like to mention. Vijay Pande and Edgar Luttmann introduced me to the process of doing computational biology research and provided the confidence to pursue it further. Thomas Artiss and Hans De Grys were superb science teachers and inspired me to pursue it as a career. I was very fortunate to learn from them. Finally, I thank my family: Hady, Iain and Johanna, who supported me throughout these endeavors.

Alex Robertson,

December 20th, 2013

# Contents

# Chapter 1

# Introduction

## 1.1 Overview

The central dogma states that DNA gives rise to mRNA via transcription which in turn gives rise to protein via translation, indicating a straight forward pathway from the genome to the organism's proteins. However, decades of subsequent research have demonstrated that while the central dogma holds true there are many layers of regulation going both forwards and backwards along the canonical pathway as well as other players, such as non-coding RNA (ncRNA). Both transcription and translation are regulated at multiple steps adding diversity of function, feedback and many other features. A major goal in RNA biology is characterizing the role proteins play in directly regulating mRNA. I present novel experimental and computational analysis methods to elucidate this role, as well as insights gleaned from these methods in mammalian and in vitro systems.

### 1.1.1 Protein-RNA interactions

The mammalian genome contains tens to hundreds of thousands of genes, the majority of which can produce multiple different RNA and protein isoforms (Wang, Sandberg,

Luo, Khrebtukova, Zhang, Mayr, Kingsmore, Schroth & Burge 2008, Pan et al. 2008). Over a thousand proteins are thought to be RNA binding (ENS; GO:0003723), and therefore have the possibility of directly regulating mRNA. Further adding to the complexity, many proteins recognize RNA structures, while others bind motifs only in a single strand context. Considering that there are so many RNA binding proteins (RBPs) and potentially tens of thousands of distinct mRNAs the space of $10^7$-$10^8$ of potential protein-RNA interactions seems beyond the scope of individual testing. In order to make this problem tractable, methods have been developed which identify protein-RNA interactions and the downstream effects of these interactions. From such methods one can derive principles for predicting effects of changes in mRNA sequence on protein-RNA interactions.

There are many processes by which a protein may affect the expression of an mRNA at each step in the RNA life cycle. Transcription factors affect the level to which an mRNA is transcribed, splicing factors affect which exons are included in the mRNA (Schwarzbauer et al. 1987, Breitbart et al. 1987, Black 2003) and continue to coat the mRNA until it is translated, and cleavage and polyadenylation factors regulate the 3' UTR of the mRNA (Takagaki et al. 1996). Once the mRNA has been fully spliced many other proteins regulate its export, localization and degradation. By characterizing the features of mRNA that lead it to being a target, one can extrapolate to other mRNAs. Furthermore, homologous proteins with greater than 70% sequence identity in their RNA binding domains appear to recognize the same RNA sequence features (Ray et al. 2013). To understand and predict the aggregate effects of RBPs on mRNAs, we need to integrate models of binding and the downstream effects of binding for individual RBPs.

Splicing, the process by which introns are excised from the primary transcript, generates a diversity of sequences from a single gene locus and is therefore especially amenable to study by deep sequencing methods. Since deep sequencing reads can often uniquely identify a gene's isoform, they allow the quantification of each potentially expressed isoform. This technique has been used to demonstrate that

alternative splicing is widespread in mammalian genomes and is regulated between tissue types (Wang, Sandberg, Luo, Khrebtukova, Zhang, Mayr, Kingsmore, Schroth & Burge 2008, Pan et al. 2008). A single modern deep sequencing experiment can quantify the splicing patterns of expressed genes for several tissues/cell types simultaneously (Mortazavi et al. 2008, Fox-Walsh et al. 2011). Analyzing the transcriptomes of cells depleted for individual splicing factors has been used to characterize its effects across all expressed mRNAs (Blanchette et al. 2009). Sequence motifs have been identified for many splicing factors, which can allow the splicing effects to be predicted for genes which are not expressed in the cell types tested directly (or for exogenous genes) (Ray et al. 2013). It is the sequence motif that defines the set of mRNAs which an RBP will regulate and thus the sequence motifs which must be understood in order to make have a deep understanding of RBPs effects on the transcriptome. There has been much progress in unraveling these sequence preferences which has greatly accelerated in recent years.

### 1.1.2 Goal of the Thesis

The goal of this thesis is to quantitate the binding affinities of splicing factors for their target mRNA sequences and uncover determinants of how nonsense-mediated mRNA decay (NMD) targets mRNAs for degradation. A major way that splicing affects the transcriptome is by targeting transcripts for degradation (Wen & Brogna 2008, Kervestin & Jacobson 2012). Exons which contain in-frame stop codons far upstream of the true stop codon produce unstable transcripts, which are targeted by the quality control pathway, NMD, a highly conserved, essential (in mammals) pathway for identifying errors in transcripts which could produce truncated proteins (Leeds et al. 1991, Conti & Izaurralde 2005, Lareau, Brooks, Soergel, Meng & Brenner 2007, Isken & Maquat 2007, Saltzman et al. 2008). NMD serves as a coupling process between splicing and degradation, linking splicing factors to mRNA levels. In order to fully characterize this phenomenon, both splicing factor affinities and NMD must be well understood.

## 1.2 Splicing

Eukaryotes use splicing to excise intronic sequence from the primary transcript (usually cotranscriptionally) of mRNA and reattach flanking segments together to form a mature mRNA transcript, in a regulated manner (Matlin et al. 2005, Chen & Manley 2009, Kornblihtt et al. 2013). For most mRNA introns (with the exceptions of XBP1 (Calfon et al. 2002) and introns spliced by the minor spliceosome) this process is catalyzed by the major spliceosome, a large molecular machine composed of a core set of required splicing machinery and non-core splicing factors. The spliceosome includes the small nuclear RNAs (snRNAs) U1, U2, U4, U5 and U6 which bind several proteins to form the five major small nuclear ribonucleic particles (snRNPs). The spliceosome assembles on the nascent transcript, usually as it is being transcribed (Tilgner et al. 2012). The intron contains a 5' splice site, a branch point, a polypyrimidine tract and a 3' splice site. Since the splice site sequences are fairly degenerate, the splicing pattern is not uniquely defined by a gene's primary sequence. Splicing is regulated in cis by the strength of these sequences and other sequence elements within or close to the alternatively spliced exons (Wang et al. 2004) or the flanking exons (Han et al. 2011).

### 1.2.1 Splicing Mechanism

Assembly of the spliceosome occurs through a series of steps defined by the binding of different factors to the pre-mRNA. First, U1 recognizes the 5' splice site and a splicing factor, SF1 binds the branch point. An auxiliary factor to U2, U2AF is recruited to the 3' end of the intron. U2AF is a heterodimer composed of U2AF35, which recognizes the 3' splice site and U2AF65 which recognizes the polypyrimidine tract. In a phenomenon known as exon definition, interactions between components of the splicing machinery at the 3' splice site interact with components across the exon at the next intron's 5' splice site. Exon definition enhances recognition of the splice sites in mammalian pre-mRNA transcripts. U2AF's role is to recruit U2 to the

branch point where it replaces SF1. In a transition from exon definition to intron definition, interactions between U2 and U1 lead to the splice sites being brought together, with the intronic sequence looped out. U4 and U6 form strong interactions with each other and U5 to make up the tri-snRNP, which is recruited to U1 and U2 on the intron. At this point the intron is committed to being spliced at the two splice sites that have been recognized.

Through a series of rearrangements the spliceosome becomes catalytically activated and two transesterification reactions occur. First the 2' hydroxyl group of the branch point residue nucleophilically attacks the RNA backbone at the 5' splice site. Second, the 3' hydroxyl which was a leaving group of the first reaction nucleophilically attacks the 3' splice site, which removes the intron and splicing the exons together at once. Each of the steps of spliceosome assembly is a target for regulation by RNA binding proteins which recognize sequence elements in the pre-mRNA to enhance or silence splicing (splicing mechanism reviewed by (Black 2003, Shin et al. 2004, Matlin et al. 2005, Chen & Manley 2009)) and in some rare cases the transesterification steps may be regulated (Lallena et al. 2002).

### 1.2.2 Alternative Splicing

Whereas most splicing occurs constitutively, a class of over ten thousand exons are included alternatively (Wang, Sandberg, Luo, Khrebtukova, Zhang, Mayr, Kingsmore, Schroth & Burge 2008, Castle et al. 2008) depending on interactions with splicing regulatory proteins and other influences, such as coupling to transcription (Das et al. 2007) or secondary structure in the RNA (Buratti & Baralle 2004, Kreahling & Graveley 2005, Tu et al. 2000). A major goal in splicing biology is to uncover the code that determines splicing regulation across cell types (Barash et al. 2010) and individuals (Pickrell et al. 2010). One approach to understanding the splicing code is to group together all features of a gene which potentially affect splicing and use machine learning to build models of how this feature set regulates splicing in different

tissues. A second approach is to look at sequence variation across individuals and identify which sequences are correlated to splicing differences. The methods used in this thesis take third route. Since the goal here is to understand the interactions of the individual proteins with mRNA, one must isolate, rather than aggregate the effects of individual RBPs.

### 1.2.3 Splicing Factors

Splicing factors contain one or more RNA binding domains (RBDs), such as RNA recognition motifs (RRM), hnRNP K homology domains or zinc fingers, and usually another functional domain which modifies the splicing of its targets (Wu & Maniatis 1993, Dreyfuss et al. 2002). The function of the protein is separated by domain, such that the RBDs determine the binding affinity and whether it activates or represses splicing is determined by its other functional domains. In vitro studies have found that truncated protein fragments which include the RBDs and some flanking amino acid residues have essentially the same binding preferences as the full-length protein (Ray et al. 2009). Replacing the RBD of an RBP with one from another splicing factor will change the binding preferences of the protein without affecting the functions of other domains (Shi et al. 1997, Wang, Gostissa, Yan, Goff, Hickernell, Hansen, Difilippantonio, Wesemann, Zarrin, Rajewsky, Nussenzweig & Alt 2009). However, while the RBDs function independently of other protein domains, the individual RBDs of splicing factors interact with each other by unpredictable non-additive processes. For example the four zinc fingers of Muscleblind proteins differ in the extent to which they contribute to the overall binding preferences (Purcell et al. 2012). Also, the free energies of binding of individual protein domains do not sum to the free energy of binding to the full-length protein (Shamoo et al. 1995). Thus, whereas one does not lose information by studying the binding preferences of a protein truncated to contain only the RBDs, one most likely cannot get that same information from integrating the binding preferences of the individual RBDs and integrating the results (Ray et al. 2013).

SR proteins are a class of conserved splicing factors which promote splicing (Fu 1995). SR proteins have an arginine/serine (RS) rich domain as well as at least one N terminal RRM and perform various roles in RNA processing (Zahler et al. 1992, Zahler et al. 1993, Sanford et al. 2005). In a series of classic experiments the RS domains of SR proteins were fused to an MS2 coat protein's RNA binding domain and tethered to several substrates in vitro (Graveley et al. 1998). These hybrid proteins are able to activate splicing of their targets because the RS domain is sufficient to confer activating functionality when appropriately localized. In another experiment, the functionality of SXL was reversed from a repressor to an activator by fusion of an RS domain (Valcarcel et al. 1993).

Another class of splicing regulators is the extended family of heterogeneous nuclear ribonucleoproteins (hnRNPs). In contrast to the SR proteins, hnRNPs are splicing repressors (Krecic & Swanson 1999, Huelga et al. 2012). A particularly interesting hnRNP, hnRNP A1 is discussed in depth below.

### 1.2.4 Splicing Regulatory Elements

The pre-mRNA sequences which regulate the splicing patterns of their genes are divided into four categories: exonic splicing enhancers (ESE), exonic splicing silencers (ESS), intronic splicing enhancers (ISE) and intronic splicing silencers, collectively referred to as splicing regulatory elements (SRE) (reviewed by (Cartegni et al. 2002)). These location distinctions are important because the splicing effects of a given sequence are position specific; a sequence which silences splicing when located in an intron may enhance the splicing of an exon when located within the exon (McCullough & Berget 1997, Chen et al. 1999). These SREs have been shown to have predictive capability for splicing of exons that they are associated with (Fairbrother et al. 2002). SREs regulate the splicing of both constitutive and alternative exons, but they are overall more highly conserved near alternative exons (Sorek & Ast 2003). There are a number of well characterized mechanisms by which these cis-elements operate to pro-

mote or prevent splicing, but most require recognition by an RBP with then interacts later on with the core splicing machinery.

SREs can be identified by a number of techniques, which differ in the degree to which they provide mechanistic insight as well as predictive capability. SREs can be identified by in vitro splicing activity screens (Liu et al. 1998), splicing reporters systems in cell lines (Wang et al. 2004, Culler et al. 2010, Wang, Ma, Xiao & Wang 2012), identifying in vivo binding sites of known splicing factors (Huelga et al. 2012) or computational genome-wide analyses (Fairbrother et al. 2002, Zhang & Krainer 2004, Yeo et al. 2007, Suyama et al. 2010). Among these possible methods, this thesis focuses on methods where the interacting trans-factor is known. After doing a cell-based screen which identified ISS motifs, Wang et al. used RNA affinity purification followed by mass-spec to determine which if any proteins are bound to the sequence (Wang et al. 2013). This analysis verified that the ISSs bound to known splicing factors, identified a network of interactions between them and suggested possible mechanisms for their effects. To test that the ISSs were active in full organismal context, a genome-wide analysis was done on human deep sequencing data, which demonstrated that alternate exons proximal to these ISSs were less highly included than expected based on a control set of exons. Together these constitute systematic methods for identifying silencers and gaining mechanistic insight into their functions.

### 1.2.5 Mechanisms of Regulation

Much work has been done in uncovering the mechanisms by which splicing factors affect the splicing of their targets once they have bound through the complementary approaches of traditional gene-by-gene experiments, genome-wide studies and reporter-based screens (reviewed by (Chen & Manley 2009)). hnRNP A1 provides an illustrative example of the diversity of different mechanisms by which splicing factors interact with SREs to regulate splicing (reviewed by (Jean-Philippe et al. 2013)). hnRNP A1 can bind to an ESS which overlaps with an ESE in the HIV-1 tat exon

2, potentially blocking the SR-protein, SC35 (also called Srsf2) from binding to that position (Zahler et al. 2004). When SC35 is blocked the exon is excluded, while when SC35 binds the exon is included. In this case the splicing of the alternate exon depends on the nuclear concentrations of each protein and the competition for binding to the mRNA, which is thought to be determined by their relative binding affinities among other factors. A1 can also hinder the recruitment of SR proteins by cooperatively binding the transcript in multiple copies blocking the 3' splice site and polypyrimidine tract of the HIV tat pre-mRNA (Okunola & Krainer 2009). Lastly, A1 can actually loop out an alternate exon bringing the splice sites of its flanking exons into close proximity and leading to its exclusion. This loop formation occurs when separate A1 proteins bind to each of the introns flanking the alternate exon and interact with each other (Blanchette & Chabot 1999, Nasim et al. 2002). Likewise, splicing activation can function via a variety of mechanisms. Together these examples show the importance of understanding the binding affinity of splicing factors (both repressors and activators) to individual sequences and the potential for cooperativity in binding to RNA.

## 1.3 Nonsense-mediated mRNA decay

NMD is a broadly conserved pathway for specifically degrading transcripts which have premature termination codons (PTCs). It is present in all eukaryotic organisms studied to date and in all cases requires the function of the highly conserved helicase UPF1 (also called RENT1) (Leeds et al. 1991). UPF1 and the other key factors in NMD are essential for mammalian development (Medghalchi et al. 2001, Weischenfeldt et al. 2008, McIlwain et al. 2010). Thought to be a quality control mechanism, NMD will selectively degrade transcripts where the open reading frame (ORF) is truncated by a PTC upstream of the true stop codon. In mammals the canonical rule is that NMD will be triggered if the ribosome translates an ORF which terminates 50 nucleotides (nt) or more upstream of an exon-exon junction (Cheng & Maquat 1993).

Full-length ORFs can be truncated by one of four possible mechanisms. One is that there is an error in the genomic sequence and all proteins expressed from that allele will be truncated. The second possibility is that there was an error in transcription which introduced a nonsense mutation into the mRNA. By preventing truncated ORFs from being translated, NMD protects from the effects of expressing truncated proteins. A truncated protein containing only an N terminal domain could bind a target molecule but lack the ability to affect the necessary downstream interactions or could misfold or aggregate either of which has the potential to cause dominant negative effects. The third, and potentially most interesting, way that an ORF can be truncated is through regulated alternative splicing (Lewis et al. 2003). The fourth possibility is that RNA editing introduces a termination codon (Chester et al. 2003).

### 1.3.1 Splicing mediated NMD

The simplest splicing event which truncates the ORF is the introduction of a poison exon into a transcript. Such events often have extremely high sequence conservation (Lareau, Brooks, Soergel, Meng & Brenner 2007). Inclusion of such an exon is often used to create an autoregulatory negative feedback loop. There are 25 splicing factors and core spliceosomal proteins that have conserved, potentially NMD-inducing, splicing events nine of which were shown to be regulated by NMD (out of 11 tested) (Lareau, Brooks, Soergel, Meng & Brenner 2007, Ni et al. 2007, Saltzman et al. 2008). This feedback loop is a way of coupling the functional level of the protein to the expression of the RNA. There are at least 11 splicing activators with ultraconserved (perfect sequence identity in at least two species) exons whose inclusion is predicted to lead to degradation by NMD (Lareau, Brooks, Soergel, Meng & Brenner 2007). Splicing repressors by contrast often regulate the levels of their own transcripts by excluding exons whose inclusion is required to preserve frame (Ni et al. 2007). If the length of a coding cassette exon is not a multiple of three then its exclusion will cause the downstream sequence to be translated out of frame and the ORF will typically be truncated. Such events are ultraconserved in at least five splicing repressors (Ni

et al. 2007).

Depending on context, either retention or inclusion of an intron can lead to NMD. Introns located in the coding sequence trigger NMD when retained as they will likely contain a termination codon in-frame and therefore truncate the ORF. Because of the way that NMD is thought to detect target transcripts (discussed below), splicing of an intron at least 50 nt into the 3' UTR may trigger NMD, while retention of the intron will not. This mechanism has been shown to be the autoregulatory mechanism of SC35 which has an alternate intron in its 3' UTR the inclusion of which is regulated by SC35 itself through binding to an ESE (Liu et al. 2000, Sureau et al. 2001, Lareau, Brooks, Soergel, Meng & Brenner 2007).

### 1.3.2 Mechanisms of NMD

According to current understanding of the mechanism of NMD in mammals, NMD relies on the typical lack of exon-exon junctions far downstream of the termination codon in mammalian genes. The number of exons in protein coding genes in the human genome ranges from one to 364 (in the case of Titin) with an average nine exons. While the farthest 3' exon is usually the longest, earlier exons are shorter; sized on the order of a hundred nt (ranging up to 17,000; with an median of 122 nt). In a very large majority of coding genes the protein coding sequence begins in one of the first few 5' exons and almost always continues to either the last exon or within 50 nt of the 3' end of the penultimate exon. In this majority of cases the final exon contains all or almost all of the 3' UTR. During splicing the spliceosome deposits a set of proteins, including the NMD-related proteins UPF2 and UPF3, in a complex called the exon junction complex (EJC). The EJC is located approximately 24 nt upstream of the exon-exon junction and is present for the majority (>80%) of exon-exon junctions (Le Hir, Izaurralde, Maquat & Moore 2000, Singh et al. 2012).

After splicing is complete the transcript and associated bound proteins are exported from the nucleus for translation. At this point the transcript has EJCs at

exon-exon junctions. It is bound by a protein complex at its 5' cap, known as cap binding complex (CBP) and has poly(A) binding protein bound to its 3' poly(A) sequence. The first ribosome to translate the transcript displaces EJCs in a 5' to 3' direction as it translocates along the ORF. If the ORF is full-length, then by the time the ribosome reaches a stop codon all the EJCs will have been removed from the message. The message will be clear of EJCs if the stop codon is up to 50 nt upstream of the final exon-exon junction, since the EJC is deposited 24 nt upstream of the junction and the ribosome has a wide footprint on the mRNA.

The presence of an EJC downstream of a stopped ribosome is the mark of a truncated ORF, which is recognized by the NMD machinery. If the ORF is truncated more than 50 nt upstream of the last exon-exon junction, then when the translating ribosome reaches the termination codon, then there will still be at least one EJC still present on the message. We refer to such downstream exon-exon junctions as dEJs. This dEJ mechanism appears to be the key for triggering NMD in vertebrates. The release factors eRF1 and eRF3 are recruited to the stopped ribosome and the stop codon. Together the release factors and the EJC components UPF2 and UPF3 work to recruit UPF1 (Le Hir, Gatfield, Braun, Forler & Izaurralde 2001, Chamieh et al. 2008, Chakrabarti et al. 2011). CBP is present during the pioneer round of translation when NMD is thought to occur (Lejeune et al. 2002), but is later replaced by eIF4E (Ishigaki et al. 2001). There is some evidence that CBP promotes this recruitment via interactions with Upf1 (Ishigaki et al. 2001, Hosoda et al. 2005, Hwang et al. 2010), but it is not strictly required as eIF4E bound transcripts have been shown to be NMD targets (Rufener & Muhlemann 2013). Once UPF1 is in place Smg1 is recruited, phosphorylates UPF1 and together with the other factors forms the SURF complex (Smg1-UPF1-release factors) (Yamashita et al. 2001, Kashima et al. 2006, Ivanov et al. 2008). At this point the mRNA is committed to degradation, which occurs rapidly by both endo- and exo-nucleolytic degradation (Gatfield & Izaurralde 2004, He & Jacobson 2001, He et al. 2003).

Crucially, one corollary of the dEJ NMD mechanism is that there is a set of PTCs

which cannot be detected: those that occur very late in the ORF and are still less than 50 nt upstream of the last exon-exon junction. This exception has important implications for the exon structure of genes and protein domains and human disease.

### 1.3.3 uORFs

A direct corollary of the dEJ NMD mechanism is that genes with upstream ORFs (uORFs) are likely to be NMD targets if the uORF is translated before the first round of translation of the primary ORF unless the ribosome reinitiates and translates the primary ORF as well. uORFs are short open reading frames starting in the 5' UTR of an mRNA. The canonical theory of NMD predicts that if a uORF is translated and the ribosome terminates at the end of the uORF there will still be EJCs present on the mRNA. In some cases ribosomes can reinitiate translation after translating uORFs, in which case the primary ORF is also translated, removing the EJCs. Reinitiation depends on the exact spacing of the ORFs and the availability of initiation factors (Vattem & Wek 2004). Luciferase and qRT-PCR assays on a panel of 26 uORF-containing human 5' UTRs showed that uORFs downregulated protein expression but generally did not have significant effects on mRNA level (Calvo et al. 2009). Several uORFs have also been shown to specifically escape NMD (Stockklausner et al. 2006). In genome-wide studies of NMD some uORFs have been shown to be NMD targets (Ramani et al. 2009, Mendell et al. 2004, Yepiskoposyan et al. 2011), however, the work presented in this thesis gives the first genome-wide evidence that translation of the uORF triggers NMD.

### 1.3.4 Alternate Mechanisms of NMD

While the canonical mechanism of NMD is thoroughly demonstrated, it does not completely explain the NMD pathway. In saccharomyces cerevisiae, where most genes are not spliced, the mechanism is less clear cut. The model, generally referred to as

the faux 3' UTR model says that PTC containing messages are distinguished from messages with full-length ORFs by the distance from the first in frame stop codon to the poly(A) tail, i.e. the length of the apparent 3' UTR. If there is an error which places a stop codon early in an ORF, then to the translational and NMD machinery the apparent 3' UTR appears elongated (Behm-Ansmant et al. 2007). 3' UTR length is thought to be detected by the proximity of PABP which is localized to the poly(A) tail. Evidence for this comes from tethering PABP to the 3' UTR just downstream of an apparent PTC via an MS2 coat protein tag (Amrani et al. 2004, Behm-Ansmant et al. 2007). The proximal PABP protects the mRNA from degradation, equivalent to having a shorter UTR. Furthermore, several studies show that messages with long wildtype 3' UTRs are upregulated when NMD is inhibited by knocking down or deleting one of the essential factors or by blocking translation by a drug (Buhler et al. 2006, Hansen et al. 2009, He et al. 2003, Muhlrad & Parker 1999).

There are several problematic implications of the faux-3' UTR model of NMD. Firstly, mRNA transcripts are circularized and so the actual proximity of PABP and the length of the apparent 3' UTR are not directly related. The second objection is that the lengths of 3' UTRs vary dramatically between mammalian genes, and targeting messages as truncated due to changes in 3' UTR length will lead to poor specificity and/or selectivity.

It has been shown that UPF1 binding in mammalian 3' UTRs can lead to degradation (Hogg & Goff 2010). Hogg and Goff propose a length-sensing mechanism role for Upf1 which potentiates mRNA decay (Hogg & Goff 2010). Based on reporter genes, they show that UPF1 can bind to mRNA sequences which are not being actively translated by the ribosome. When a region of mRNA is left untranslated, such as if a PTC occurs upstream of it, then UPF1 binding is undisturbed. They propose that a sufficient amount of undisturbed Upf1 binding potentiates the transcript for degradation via an unknown mechanism. In a separate pathway UPF1 is also known to degrade certain mRNAs based on binding to 3' UTRs directed by Staufen1, Staufen-mediated decay (SMD) (Kim et al. 2005). Stauffen1 (STAU1) binds to STAU1-binding sites

(SBSs) in the 3' UTR and recruits Upf1 to the message. SBSs are double stranded motifs which can be formed by either intramolecular hairpin formation within the 3' UTR or intermolecular base pairing between two half-sites, one in the 3' UTR and the second in ncRNA (SMD reviewed by (Park & Maquat 2013)). Since SMD and NMD both require Upf1 they are competitive pathways and increased SMD has been shown to lead to decreased NMD efficiency (Park et al. 2013).

Overall, the evidence indicates that in mammals NMD is largely directed by the EJC rule, but its efficiency is influenced by the length of the UTR, but Upf1's role as a 3' UTR directed regulator of transcript stability is also observed.

## 1.4 Relevant Technology

Recent years have seen a surge in the availability of technology used in systems biology. This section summarizes some of these advancements, the technologies they replace and the necessary novel computational tools needed to deal with the increased depth of information. One primary driver of this surge is the advent of deep sequencing technologies, which allow genomes to be sequenced cheaply and can be used to quantitate transcriptomes. Deep sequencing has also been adapted into Ribo-seq which is used to characterize which part of the transcriptome is being translated. Furthermore, deep sequencing has been adapted to study protein-DNA and protein-RNA interactions, offering advantages over previous methods.

### 1.4.1 Transcriptomics

Deep sequencing can be used to perform a deeper examination of the transcriptome than previous technologies. Prior to the wide adoption of deep sequencing technology, exon junction and tiling microarrays were often the best available methods for profiling splicing transcriptome-wide (Clark et al. 2002, Johnson et al. 2003, Ule

et al. 2005, Pan et al. 2006, Sugnet et al. 2006). Deep sequencing is the sequencing of hundreds of millions of short (40-500 nt) reads. RNA-seq is a version of deep-sequencing where RNA or cDNA is sequenced in order to characterize the transcriptome. In RNA-seq RNA from the sample of interest is isolated, processed and sequenced. Reads are then mapped to the relevant genome to identify their genes of origin. The distribution of reads across genes allows one to quantify their expression. Expression of a gene is frequently expressed as the fragments per kilobase per million mapped reads (FpKM), so the expression is normalized to the total amount of reads in the sample. However, counting cells and using spike-in controls allows one to calculate average copies per cell for a given gene or transcript (Jiang et al. 2011). Even without such controls RNA-seq is an advance over microarrays as it allows direct comparisons of genes relative to each other. In addition to quantifying gene expression, RNA-seq reads may give isoform specific information, if they are located on a cassette exon or if the read spans an exon-exon junction which uniquely identifies a cassette exon as being skipped.

Ribo-Seq is a recently developed technique for profiling translation across the transcriptome. Ribosomes are immobilized on their messages (by either flash freezing or drug treatment) and the footprint of mRNA being covered by the ribosome is isolated and sequenced. These short sequences (approximately 30 nt) are mapped to the relevant transcriptome and give detailed information about exactly which parts of the transcriptome are being translated under the examined conditions. Ribo-seq experiments have given important insight into the reaction of yeast to glucose starvation (Ingolia et al. 2009), translational changes during embryoid body formation (Ingolia et al. 2011) and the translational response to heat stress (Shalgi et al. 2013). The precise locations of ribosome footprints has also been used to identify translated uORFs (Ingolia et al. 2009).

Novel, more efficient, algorithms needed to be developed due to the increased depth of data generated by deep-sequencing experiments. Early methods of mapping short reads to a genome or genomes, such as BLAST or BLAT, would take impracti-

cally long to sequence the results of a modern deep-sequencing experiment (Altschul et al. 1990, Kent 2002). New software was required to map these reads in a reasonable amount of time. Many such software packages have been published, often based on the Burrows-Wheeler transform, an algorithm for mapping short sequences to very long sequences rapidly (Burrows & Wheeler 1994, Li & Durbin 2009, Langmead et al. 2009). The algorithm has been adapted to map spliced mRNA reads to the transcriptome. Also, the algorithm has identified novel splicing isoforms which had not previously been annotated by ESTs (Trapnell et al. 2009).

### 1.4.2 Techniques for measuring protein-nucleic acid interactions

Surface Plasmon Resonance (SPR) is a biophysical technique which exploits certain properties of total internal reflection (TIR) to make highly accurate measurements of the binding affinities of protein-nucleotide interactions (reviewed by (Zeng et al. 2011)). One species - either the protein or an oligonucleotide - is immobilized on a gold surface in a glass flowcell and the other species is flowed over. A laser is shined at the bottom of the glass and spread by the prismatic properties of the glass slide as it enters the slide. The angularly dispersed light reflects off the bottom of the slide via TIR, with the exception of a band which reaches the interface at the exact angle required to excite an electromagnetic wave parallel to the surface, called a surface plasmon. This absorption of light into the surface wave leaves behind a dark band in the light reflected off the surface, the exact location of which is determined by the mass of material deposited on the gold surface. SPR exploits this by flowing the antagonist to the immobilized species through the flowcell at increasing or decreasing concentration and measuring the resulting shifts in resonant angle to measure both the association rates, dissociation rates and thus the resulting dissociation constant ($K_d$). The advantages of SPR are that it is highly quantitative and that it gives both kinetic and equilibrium measurements. The disadvantages are that it requires

expensive equipment, is low-throughput and is laborious.

Electrophoretic mobility shift assays (EMSA), more commonly known as gel shifts, are a more commonly used technique which exploits the decreased mobility of protein-RNA (or protein-DNA) complexes relative to each species alone in a gel (Garner & Revzin 1981, Fried & Crothers 1981). When the protein and RNA are incubated together and then run in a polyacrylamide or agarose gel then there will be a distinct band for the protein-RNA complex, shifted higher than the RNA alone. By quantitating the band intensities at different concentrations of protein one can calculate the apparent $K_d$. This method is simpler as it requires less equipment than SPR but is still low-throughput and laborious.

Systematic evolution of ligands by exponential enrichment (SELEX) is a method for measuring protein-RNA or protein-DNA interactions when the binding motif is unknown (Ellington & Szostak 1990). The method is begun by generating a randomized single stranded library of possible sequences flanked by the necessary primers for amplification. In successive rounds of enrichment the library is selected for sequences which bind the protein of interest, usually with protein bound to beads (Ogawa & Biggin 2012). These selected oligos are eluted and amplified and then re-selected in several rounds of increasing stringency. After several rounds of selection the remaining oligos are sequenced, yielding the binding motif or motifs. The advantage of SELEX is that it will reliably select for strong-binding sequences from a large set of potential sequences. The disadvantages are that it is not quantitative and will not resolve the most strongly binding sequences over those merely strong enough to make it to the final pool (Carothers et al. 2006).

High-throughput sequencing - fluorescent ligand interaction profiling or HiTS-FLIP is a high throughput and quantitative in vitro method for profiling the binding affinities of protein for DNA sequences (Nutiu et al. 2011). It takes advantage of the Illumina Genome Analyzer 2's capability of measuring the fluorescence of millions of immobilized DNA clusters -each made up of many hundred molecules of the same

sequence (Bentley et al. 2008). By flowing fluorescently tagged protein over the DNA clusters, the extent of protein binding can be measured by the standard imaging software normally used to sequence the library. Normalizing to the cluster size allows calculation of the extent of binding to each DNA sequence present in the library. While this is an extremely powerful method for profiling protein-DNA interactions, an adaptation to protein-RNA binding has yet to be reported.

Crosslinking immunoprecipitation followed by sequencing (CLIP-seq) is a powerful high resolution method for in vivo characterization of protein binding (Sanford et al. 2009, Chi et al. 2009). CLIP-seq works by crosslinking RBPs to their target mRNAs via UV light, isolating the complexes via immunoprecipitation and then sequencing the resulting fragments. This allows transcriptome-wide profiling of RBP binding sites with a resolution of about 40 nt; in the iCLIP variant single nucleotide resolution is achieved by sequencing up to the crosslinking site (Konig et al. 2010). CLIP has been used to show that Mbnl can either repress or activate splicing depending on its binding location relative to the exon. CLIP is highly specific, allowing one to probe the binding pattern of a protein in specific cell line under chosen conditions, though one must be careful as the patterns identified may not extend to other systems. Another issue is that detection of binding is proportional to the level of gene expression which must be carefully controlled for.

As with new technologies in transcriptome profiling, the advent of novel high-throughput technologies for measuring biomolecular interactions requires new algorithms and software to analyze the increased amounts of data. Traditional motif finding algorithms such as MEME (Bailey & Elkan 1994, Bailey et al. 2009) are able to identify sequence motifs in small sets of sequence space, such as the promoters of a set of several hundred genes, but are less well equipped for distilling the information derived from deep sequencing experiments which can yield hundreds of millions of sequences in a single experiment.

# Chapter 2

# Nonsense-mediated mRNA Decay

## 2.1 Introduction

The multi-step nature of eukaryotic gene expression and RNA processing enables multiple layers of regulation but also introduces more opportunities for error. Nonsense-mediated mRNA decay (NMD) is a highly conserved RNA surveillance pathway that oversees mRNA translation and targets those mRNAs harboring premature termination codons (PTCs) for decay, preventing the cell from producing potentially deleterious truncated proteins . As a translation-dependent process, NMD is triggered when a ribosome stalls at the termination codon (TC) of a target RNA and recruits the RNA helicase UPF1 (reviewed by (Kervestin & Jacobson 2012)). UPF1 is conserved in all studied eukaryotes and strictly required for NMD activity (Leeds et al. 1991, He et al. 1993, Hodgkin et al. 1989, Page et al. 1999, Gatfield et al. 2003, Arciga-Reyes et al. 2006, Sun et al. 1998, Czaplinski et al. 1995). The NMD pathway has important implications in human disease, as ~11% of disease-causing mutations result in the production of nonsense-containing mRNAs (Mort et al. 2008) and frequently result in haploinsufficiency phenotypes (reviewed by (Kuzmiak & Maquat 2006, Holbrook et al. 2004)).

Interestingly, while NMD is traditionally considered to be required to prevent the translation of aberrant mRNAs that harbor mutations or result from errors in transcription or splicing, this pathway is also implicated in regulating the expression of many normal (wildtype) genes and mRNAs (reviewed by (Nicholson et al. 2010, Chang et al. 2007, Rehwinkel et al. 2006)). These include mRNAs harboring upstream open reading frames (uORFs), selenocysteine codons, long 3' UTRs, or alternative splicing events that generate isoforms with PTCs. While this last mode is used to regulate the levels of specific factors, particularly splicing factors (Lareau, Inada, Green, Wengrod & Brenner 2007)(Ni et al. 2007)(Saltzman et al. 2008), in general, the regulation of and importance of this pathway's effects on wildtype gene expression remains poorly understood. A large fraction of the mammalian genome appears to be regulated by NMD; two recent studies have estimated that between one sixth and one quarter of mammalian genes are affected by this pathway (Weischenfeldt et al. 2012, McIlwain et al. 2010). Mice homozygous null for key NMD factors die during embryogenesis (Medghalchi et al. 2001, McIlwain et al. 2010, Weischenfeldt et al. 2008), leading to the hypothesis that aberrant expression of NMD target mRNAs contributes to these phenotypes. While distinguishing primary from secondary effects of inhibition of the NMD pathway remains challenging, how NMD activity regulates mammalian transcriptome early in development is not well understood.

In mammals, targeting of UPF1 to mRNAs that harbor a PTC is primarily thought to occur via its specific interactions with additional, strategically positioned NMD factors. Pre-mRNA splicing results in the deposition of a multi-protein complex, known as the exon junction complex (EJC), ~20-24 nt upstream of the exon-exon junction. The EJC can recruit many different factors that affect mRNA metabolism, including the NMD factors UPF2 and UPF3 (Le Hir, Moore & Maquat 2000, Le Hir, Izaurralde, Maquat & Moore 2000, Le Hir, Gatfield, Izaurralde & Moore 2001). When recruited to the EJC and sufficiently downstream of a TC, UPF2 and UPF3 can stabilize UPF1 interactions at the terminating ribosome and stimulate both its ATPase and its helicase activity (Chamieh et al. 2008, Chakrabarti et al. 2011) as well as its phos-

phorylation by the kinase SMG1 (Yamashita et al. 2001). These activities, in turn, trigger a cascade of events resulting in degradation of the target mRNA. Exon-exon junctions positioned >50 nt 3' of the TC (downstream exon-exon junctions or dEJs) trigger NMD of the host mRNA (Cheng & Maquat 1993), a distance likely reflecting the sizes of the terminating ribosome and EJC. Since EJCs are normally displaced by a transiting ribosome during the first or pioneer round of translation (Lejeune et al. 2002), typical mammalian mRNAs lacking dEJs (Nagy & Maquat 1998, Giorgi et al. 2007), will be cleared of EJCs in this process and will therefore fail to recruit UPF1 and will escape from NMD.

An additional feature of mRNAs that enhances NMD susceptibility is extended 3' UTR length (Buhler et al. 2006, Hansen et al. 2009, Ramani et al. 2009, Yepiskoposyan et al. 2011). Factors that associate with polyA tails (mainly cytoplasmic poly(A) binding protein, PABPC1) can compete with UPF1 for binding to the terminating ribosome (Behm-Ansmant et al. 2007, Singh et al. 2008) and modulation of the intra-PABP-TC distance alters message stability (Eberle et al. 2008, Singh et al. 2008, Amrani et al. 2004). Recent studies demonstrated that UPF1 can associate with 3' UTRs of some mRNAs (Hogg & Goff 2010, Kurosaki & Maquat 2013) including an endogenous, long 3' UTR previously shown to be sufficient for decay (Hogg & Goff 2010). However, specificity of UPF1 for particular UTRs is not understood, and the transcriptome-wide binding profile of UPF1 remains largely unknown. Furthermore, the relative contributions of dEJs and 3' UTR length to NMD of endogenous mRNAs have not been assessed genome-wide.

Despite progress in understanding NMD mechanisms, the canonical determinants of NMD- 3' UTR length and presence of a dEJ- do not fully explain the observed impact of NMD on the transcriptome. For example, many mRNAs that appear as NMD targets in genome-wide studies lack these canonical features and many transcripts that harbor these traits are not repressed, suggesting that they possess features that enable full or partial escape from degradation. Genome-wide, presence of an upstream open reading frame (uORF) in a gene's 5' UTR has been associated with

NMD (Ramani et al. 2009, Mendell et al. 2004, Yepiskoposyan et al. 2011). However, detailed analysis of specific uORF-containg mRNAs has revealed that only a fraction is actually targeted by this pathway (Linz et al. 1997, Zhao et al. 2010, Stockklausner et al. 2006). Genes with longer than average 3' UTRs have been associated globally with decay (Mendell et al. 2004, Buhler et al. 2006, Hansen et al. 2009, Ramani et al. 2009). However, only a few specific UTRs have been shown to confer this activity (Singh et al. 2008, Yepiskoposyan et al. 2011). Similarly, direct binding of mRNAs by UPF1 has been associated with NMD for only a handful of metazoan messages (Hwang et al. 2010, Hogg & Goff 2010). Thus, large-scale identification of direct NMD targets remains challenging and the transcriptome-wide binding and activity of UPF1 poorly characterized.

Here we sought to define the role of UPF1 in gene expression of an early developmental system, murine embryonic stem cells (mESCs), by identifying UPF1 binding locations within the transcriptome and globally measuring the changes in mRNA abundance and translation following perturbations to the NMD pathway. We associate uORF translation with NMD susceptibility and identify a class of UPF1-bound mRNAs that undergo repression by NMD in the absence of canonical NMD features. Interestingly, the set of messages bound by UPF1 in mESCs is regulated by NMD in other mouse cells/tissues, and NMD regulation of this group of mRNAs is conserved to human. Our results enabled us to describe additional features associated with NMD, to quantify the contributions of these and canonical NMD-triggering features to the decay of endogenous mRNAs, and to better understand the role of NMD in embryonic cells.

## 2.2 Results

### 2.2.1 Hundreds of mRNAs with dEJs and long 3' UTRs are de-repressed by UPF1 depletion and translational inhibition in mESCs

To identify NMD-regulated genes and isoforms in and early developmental system, we performed RNA-Seq analysis of wildtype mESCs (v6.5) depleted of UPF1 or treated with cycloheximide (CHX). CHX is a potent translation elongation inhibitor and relatively short treatment of cells with this drug results in the stabilization of NMD-targeted mRNAs (Carter et al. 1995, Noensie & Dietz 2001). We reasoned that use of multiple methods to inhibit NMD would increase our ability to identify authentic NMD target mRNAs. Stable mESC lines were derived using two independent shRNA sequences targeting Upf1 (denoted Upf1-1 and Upf1-2) or a control shRNA targeting GFP. Cells infected with Upf1-specific shRNAs had UPF1 protein levels between 21% and 37% of the levels in control cells (Fig. S1A) and Upf1 mRNA levels between 14% and 25% of those in controls (Fig. S1B). OCT4 levels and alkaline phosphatase staining remained similar between UPF1-depleted and control-depleted cells, supporting that ESC state is maintained in the knockdowns (Figs. S1C,D). Translational inhibition using CHX was performed on wildtype mESCs for 2 hours, a duration that caused significant increase in abundance of known NMD target mRNAs without causing visible cytotoxicity.

RNA-Seq reads were mapped to the mouse genome and exon-exon junctions and both gene and isoform-specific abundances were calculated (Trapnell et al. 2010). Pair-wise comparisons of gene and mRNA expression values for each experiment were made relative to control, e.g., v6.5 CHX to v6.5, Upf1-1 to GFP, and Upf1-2 to GFP, following normalization (see Methods). As expected, the number of mRNA isoforms whose expression increased or decreased in both of the UPF1 knockdown experiments was greater than that observed between either of the these and translational inhibi-

tion, indicating that translational inhibition and UPF1 knockdown likely have some targets unique to each treatment (Fig. 1A). However, mRNA isoforms whose expression increased or decreased by more than 1.1-fold in each NMD inhibition treatments separately (Upf1-1 and Upf1-2 knockdowns, and CHX) were twice as likely to overlap between the three experiments as expected by chance. The extent of this overlap rose with increasing fold change, indicating consistency in the gene expression response to the three treatments (Fig. S1E). The extent of overlap above background was greater for mRNAs that increased in abundance after treatment than for those that decreased, consistent with NMD functioning as a negative regulator of mRNA levels as has been described in other systems (Fig. S1E) (Mendell et al. 2004).

We next sought to address the extent to which the NMD pathway regulates canonical NMD targets in mESCs. We evaluated the changes in mRNA abundance that resulted after NMD inhibition for mRNAs having one or more downstream exon-exon junctions (dEJs) and for mRNAs with varying 3' UTR lengths. We defined a dEJ as an exon-exon junction located 50 nt 3' of an annotated TC (Nagy & Maquat 1998). dEJ-containing isoforms include both mRNAs that harbor a PTC (e.g., as a result of alternative splicing) and mRNAs with introns in their 3' UTRs. While not a universal rule (Sauliere et al. 2012, Singh et al. 2012), these mRNAs are likely to have a EJC between the TC and dEJ, thus enabling stimulation of UPF1 activity. The abundance of messages harboring a dEJ increased significantly following UPF1 knockdown relative to mRNAs without dEJs (Fig. 1B). Since mRNAs whose expression change similarly in the three NMD inhibitory treatments are likely enriched for authentic NMD targets, we developed a consistency criterion and identified mRNAs that consistently increased, consistently decreased, or remained unchanged within the three experiments (see Methods). These three groups are designated as consistent mRNAs (~3900 mRNAs), with a similar classification applied at the gene level (~4500 genes) (Table S2; available online). Indeed, the consistent subset of dEJ-containing mRNAs showed stronger de-repression upon UPF1 knockdown than the full set of dEJ-containing messages (Fig. 1B), supporting the enrichment for authentic NMD

targets by applying this filter. The median log2 fold change (LFC) in expression of these mRNAs following the three treatments (relative to the that of non-dEJ mRNAs) was on average 1.19-fold (P <1E-7, difference in median of logged values shown in Fig. 1C). Similar comparisons between two control clones expressing a GFP-targeting hairpin yielded much smaller fold changes of 4% (NS, not shown).

We next assessed the relationship between UTR length and expression regulation resulting from UPF1 depletion or translational inhibition. In general, mRNAs with longer 3' UTRs (>1500 nt) were de-repressed compared to those with shorter 3' UTRs (50-350 nt) following NMD inhibition. Similar to dEJs, the degree of de-repression between these groups of mRNAs was largest when considering only consistent mRNAs (Fig. 1D, E) but was also significant when including all mRNAs (data not shown). This trend was observed in all three NMD inhibition experiments with the median difference in de-repression being somewhat greater (on average 1.26-fold) for the UPF1 knockdowns than for the CHX treatments (1.15-fold, all P <2E-23, Fig. 1E). In all cases, we observed the magnitude of the 3' UTR length effect seen in the experimental treatments exceeded that observed between controls (see Methods). Interestingly, the de-repression associated with 3' UTR length was not dependent on the presence of a dEJ (Fig. 1F), as constitutive 3' UTR mRNAs lacking dEJs exhibited similar de-repression as all mRNAs with long 3' UTRs upon NMD inhibition (Fig. S1F). Notably, the observed de-repression of mRNAs harboring a dEJ or long 3' UTR following NMD inhibition persisted at higher minimum expression thresholds, indicating the robust nature of these changes in our system (Fig. S1G).

In summary, dEJs and long 3' UTRs are both associated with NMD activity in mESCs. While the median expression changes associated with presence of a dEJ or long 3' UTR were moderate, some mRNAs changed much more than this. For example, 24 consistent dEJ messages and 33 consistent messages with 3' UTRs >1500 nt were de-repressed over 2-fold following NMD inhibition (based on the geometric mean of the fold changes of the three treatments) (Table S2; available online). The observed fold changes almost certainly underestimate the magnitude of NMD's effects,

since the ~3-4-fold UPF1 knockdown achieved likely does not completely abolish NMD activity and many of this pathway's most efficient targets are likely to still be repressed under these conditions. Importantly, messages that possessed either a dEJ or a long 3' UTR were not universally de-repressed by NMD inhibition. While some of these messages may have escaped our detection as NMD targets because of residual UPF1 activity or incomplete translational inhibition, it is also possible that some of these mRNAs may have additional features that render them insensitive to NMD.

### 2.2.2 Repression afforded by a dEJ is strongest when within a short 3' UTR

While addition of a dEJ to an mRNA with a long 3' UTR can enhance NMD-associated repression (Singh et al. 2008), the relationship between 3' UTR length and presence of a dEJ as NMD determinants has not been addressed genome-wide in mammalian cells. To assess the degree of de-repression afforded by a dEJ in different contexts, we compared the changes in abundance of dEJ and non-dEJ subsets of mRNAs in different 3' UTR length classes. We found that, in general, presence of a dEJ was associated with increased de-repression irrespective of 3' UTR length class (Fig. 1F). However, after correction for the de-repression associated with UTR length, the relative increase in expression associated with presence of a dEJ was much higher (1.63-fold) for mRNAs with short 3' UTRs (50-350 nt) than for those with longer UTRs (1.14-fold for UTRs longer than 800 nt, Fig. 1F). This finding suggests that NMD triggered by a downstream EJC is most active for transcripts with short 3' UTRs, and that transcripts with longer UTRs are less sensitive to the presence/absence of a dEJ.

### 2.2.3 Genes derepressed following NMD inhibition are enriched for transcription factors

Analysis of the biological functions of de-repressed genes revealed expected results as well as some surprises. Several known NMD-targeted isoforms increased in abundance upon NMD inhibition in mESCs, including isoforms of genes involved in pre-mRNA splicing and NMD (Table S2; available online). In addition, one of the largest and most strongly enriched categories among derepressed genes was transcriptional regulators, including many DNA binding transcription factors (GO:0045449~regulation of transcription P = 1.5E-11, Benjamini-corrected P = 4.4E-9) (Table S3; available online). Previously, regulation of some transcription factors (TFs) by NMD has been observed in mouse embryonic fibroblasts (MEFs) and HeLa cells (McIlwain et al. 2010, Wang, Wengrod & Gardner 2011). As transcriptional regulation plays an essential role in mESC pluripotency and differentiation programs (Young 2011), modulation of TF abundance by NMD could be of special importance in these instances.

### 2.2.4 Translated but not untranslated uORFs are associated with NMD

mRNAs harboring upstream open reading frames (uORFs) may be susceptible to NMD. If a uORF is translated prior to initial translation of the main ORF in a gene with typical intron distribution, downstream EJCs will be present when the ribosome terminates. Additionally, the typically large distance from the uORF to the poly-A tail could trigger NMD. Under this model, however, decay is triggered only if the uORF is translated.

A previous integrative study reported that genes with uORFs tend to produce ~10-40% less protein than those without uORFs, with less significant effects on mRNA levels (Calvo et al. 2009). Furthermore, several cases of uORFs that seemingly

escape NMD have been described (Stockklausner et al. 2006), leaving the question open as to the degree that uORFs globally effect mRNA stability. Triggering of NMD by uORFs has been characterized at the single gene level (Linz et al. 1997, Zhao et al. 2010) or inferred based on global expression analysis (Mendell et al. 2004, Ramani et al. 2009). Only recently has the translational status of uORFs been assessed genome-wide (Ingolia et al. 2009, Ingolia et al. 2011). Here we sought to identify uORFs that are actively translated and to assess their contribution to NMD in the mESC transcriptome. We used ribosome footprint profiling (Ingolia et al. 2009) to map ribosome locations within mRNAs and to assess the translational status of uORFs.

Ribosome footprinting was performed using wildtype, UPF1-depleted and control-depleted mESCs, and the density of footprint reads was used to distinguish actively translated uORFs (tuORFs) from non-translated uORFs (ntuORFs) in each cell line (Table S4; available online). In our classification scheme we only considered uORFs located completely upstream of the annotated translation start site in order to cleanly distinguish footprint reads belonging to the uORF and to the main ORF. Overall, the density of footprint reads in uORFs was well correlated between experiments (Spearman $= 0.86$ between Upf1-1 and control cells) (Fig. S2A). We defined a tuORF as a uORF that had ribosome footprint coverage at least 5-fold greater than that of surrounding regions. An ntuORF was defined as a uORF that had footprint coverage no greater than the coverage of surrounding regions (see Methods). These definitions are conservative, enabling determination of translation status when the evidence is fairly strong, but leaving some uORFs unclassified. Genes were then classified by the presence and translation status of their upstream uORFs, such that any gene harboring one or more tuORFs was classified as a tuORF gene and any gene that contained one or more ntuORFs and lacked tuORFs was classified as an ntuORF gene. Ribosome footprint data for a typical tuORF-containing gene, the transcription factor Dmtf1, and a ntuORF-containing gene, Armc1, are shown in Figure 2A. Using these definitions, we identified 392 and 464 tuORF genes in control

and UPF1 knockdown cells, respectively with most (347) in common between the two sets. Conversely, we identified 237 and 204 ntuORF genes in control and UPF1 knockdown cells, respectively. Most of the ntuORFs identified in each cell line also had low uORF to background footprint density ratios in the other cell line (Fig. S2B), and <1% of all uORFs classified as either translated or untranslated in one cell line switched classification in the other. For downstream analyses we used uORF classifications derived from UPF1 knockdown cells, as we reasoned that this condition would give us a better opportunity to observe isoforms that are actively targeted by NMD.

Notably, tuORF genes were modestly but significantly de-repressed relative to ntuORF genes following UPF1 depletion (P <0.001 for both hairpins, Fig. 2B,C). While the degree of tuORF-associated de-repression was strongest for the subset of consistent genes, it was also significant for all tuORF-containing genes (Fig. 2B). Furthermore, tuORF genes were significantly de-repressed upon NMD inhibition compared to uORF-containing genes overall (Fig. 2B and S2C). Similar trends with smaller magnitudes were observed following translational inhibition (Fig. 2C and S2C). Together, these results suggest that regulated uORF translation can often modulate mRNA stability via NMD.

Interestingly, tuORF-containing genes were enriched for transcriptional regulators compared to all expressed genes (GO:0045449~regulation of transcription, P = 3.7E-5, Benjamini-corrected P <0.05). Furthermore, we observed that genes encoding transcriptional regulators were enriched 1.5-fold for tuORFs compared to all expressed genes (P = 4.1E-6) and this enrichment increased to 2-fold for consistently de-repressed messages (Fig. 2D). Together these findings suggest that NMD triggered by uORF translation is an important mechanism of gene expression regulation in mESCs, and particularly for modulators of transcription.

### 2.2.5 Identification of hundreds of mRNAs bound by UPF1, mostly in 3' UTRs

One challenge facing study of mammalian NMD and of UPF1, in particular, is the identification of direct regulatory targets. While UPF1-bound mRNAs have been associated with NMD in yeast genome-wide (Johansson et al. 2007), metazoan studies have inferred UPF1 targets using indirect evidence such as changes in gene expression following UPF1 depletion. Here, we identified binding targets of UPF1 in mESCs using CLIP-Seq. Wildtype mESCs were UV-crosslinked, and the resulting RNA-UPF1 complexes were immunoprecipitated using antibodies against endogenous UPF1 after limited RNase digestion. Since the RNase used can impact CLIP-Seq results (Kishore et al. 2011), we prepared libraries using both RNaseA (2 libraries: Upf1.A1 and Upf1.A2) and RNaseI (1 library: Upf1.I) to enhance the robustness of the analysis. Small RNA fragments that co-precipitated with UPF1 were isolated, amplified, and sequenced. Anti-rabbit IgG precipitates harvested in parallel contained little or no cross-linked RNA, indicating low levels of intact background RNA-protein complexes remaining after stringent washing during the CLIP procedure (Fig. S3A). After mapping the resulting CLIP-Seq reads to the mouse genome and transcriptome and subtracting background read density, we determined the fraction of reads mapping to different genic regions (Fig. S3B). The density of CLIP reads per nucleotide was ~10-30-fold higher in exons than introns in all samples, consistent with the expectation that UPF1 interacts predominantly with mature mRNAs in the cytoplasm (Fig. S3C).

Based on the standard model of NMD activity, we had initially hypothesized that the majority of binding events would reside in close proximity to PTCs and/or dEJs. Instead, we observed a pronounced bias for UPF1 binding to occur in mRNA 3' UTRs, which was consistent in all three CLIP libraries (Fig. 3A). When combining the data for all genes in a metagene analysis, the density of UPF1 binding increased rapidly to ~10 times that seen in coding regions just downstream of the TC and

remained high throughout the 3' UTR (Fig. 3B). Preferential binding to the 3' UTRs of specific mRNAs was observed (controlling for gene expression; Fig. S3D), and these preferences were strongly correlated across replicate UPF1 CLIP samples, indicating the gene-specific nature of the UPF1 binding signal (Fig. 3C). UPF1 also exhibited preferential binding to specific locations within 3' UTRs (Fig. 3D). The positions of binding along 3' UTRs were correlated between replicates, and clustered separately from CLIP-Seq locations obtained for other RNA binding proteins AGO2 and MBNL1 in two recent studies of mouse cells (Leung et al. 2011, Wang, Cody, Jog, Biancolella, Wang, Treacy, Luo, Schroth, Housman, Reddy, Lecuyer & Burge 2012), indicating the specificity of the interactions identified. Correlations involving sample Upf1.A2 were less strong than those between samples Upf1.I and Upf1.A1, likely reflecting the lower complexity of the Upf1.A2 library. Comparison of UPF1 and AGO2 binding sites revealed some significant overlaps (Fig. S3E). While overlap by itself does not imply a functional relationship, a previous study showed that AGO2 can inhibit NMD (Choe et al. 2010).

While analysis of UPF1 binding sites in 3' UTRs did not reveal a clear sequence motif, we did find that UPF1-bound regions are enriched for guanosine residues ($P < 0.0001$, Chi-square test, Fig. 3E). Given UPF1's function as an RNA helicase, we also analyzed RNA structural features. We observed that UPF1 binding sites had higher propensity to form secondary structures (more negative Gfolding) than surrounding areas (Fig. 3SF), an effect that was significant overall but could be attributed to increased GC content (not shown). Thus, our data suggest that UPF1's residence within a 3' UTR is biased toward primary sequences rich in G nucleotides or towards structures produced by G-rich RNA. Furthermore, analysis of the two CLIP libraries that had deeper coverage (Upf1.A1 and Upf1.I, Table S1; available online) revealed that the extent of UPF1 binding in the upstream half of 3' UTRs was correlated with the extent of binding to the downstream half of the same UTR (Spearman $= 0.3$ to $0.4$, $P = 0.018$ and $0.0013$, respectively, in the two libraries). This observation might result from sliding (translocation) or UPF1 along the 3' UTRs

of some mRNAs (Melero et al. 2012). Together, the binding data paint a picture of a factor with a moderate degree of specificity for particular mRNAs and locations within their 3' UTRs.

### 2.2.6 Translation displaces UPF1 from ORFs

To ask whether the process of translation influences UPF1 binding locations, we performed CLIP-Seq analysis of UPF1 after 2 hours of CHX treatment. Under these conditions, UPF1 CLIP tags were enriched in mature mRNAs, as in control conditions (Fig. S3C). However, CHX treatment also caused a dramatic redistribution of UPF1 binding within mRNAs, resulting in much higher levels of binding to coding regions (Fig. 3A), with similar densities of binding upstream and downstream of TCs overall (Fig. 3B). The redistribution of UPF1 binding locations following a 2-hour inhibition of translation indicates that UPF1 binding to RNAs is fairly dynamic and suggests that translating ribosomes normally displace UPF1 from ORFs, as likely occurs for other RNA binding factors (Grimson et al. 2007).

We next identified significantly UPF1-bound mRNAs in control conditions by comparing the number of UPF1 bound positions within mRNAs relative to what would be expected if binding were random (controlling for gene length and expression level). Given that UPF1 is an RNA helicase, likely interacting transiently with RNA, we adopted a method of identifying high confidence targets within specific gene regions (3' UTRs or coding regions) rather than specific positions (Methods). After filtering for significant binding in replicate CLIP samples, we identified just over 200 high confidence target mRNAs with significant UPF1 binding in their 3' UTRs and 17 genes with significant binding to coding regions (Table S5; available online). As a control, reads sampled randomly from the RNA-Seq data at comparable 3' UTR depths as the CLIP reads yielded very few significantly enriched genes (Fig. S3G). Unbound mRNAs were defined as those displaying no UPF1 binding in any CLIP experiment. Analyzing genes encoding UPF1-bound mRNAs by Gene Ontology

analysis did not yield significant biases after correcting for multiple comparisons, but we noted that some bound mRNAs encoded proteins involved in NMD, protein localization or ESC biology. These included the estrogen-related-receptor beta mRNA, Esrrb, a factor with documented roles in ESC pluripotency and reprogramming (Feng et al. 2009, Zhang et al. 2008) (Fig. 3F), as well as mRNAs encoding a number of protein localization factors, including karyopherin 1 (Kpna1), cell cycle factors such as aurora kinase A (Aurka), cyclin dependent cyclin 25a (Cdc25a), and thioredoxin interacting protein (Txnip), and the NMD-related mRNAs Smg6 and Smg7. Interestingly, several NMD factors, including SMG6 and SMG7, participate in autoregulatory feedback circuits to regulate their own levels (Huang et al. 2011, Yepiskoposyan et al. 2011). In the case of SMG7 at least, the 3' UTR appears to mediate this regulation (Yepiskoposyan et al. 2011); our data raise the possibility that this regulation occurs through direct binding of UPF1 to this mRNA's 3' UTR.

### 2.2.7 UPF1 binding in 3' UTRs is associated with repression

We next assessed whether UPF1 binding is associated with UPF1 activity by measuring the abundance of mRNAs bound by UPF1 before and after UPF1 depletion and translational inhibition. Given that UPF1 binding occurred predominantly to the 3' UTRs of mRNAs, we chose to focus on messages bound in this region for further analysis. However, we did find that those few genes that were bound by UPF1 in their CDS were on average derepressed following NMD inhibition (Fig. S4). mRNAs with significant UPF1 binding in their 3' UTRs were de-repressed compared to unbound mRNAs following all of NMD-inhibitory treatments (median fold change , $P<1E\text{-}7$), implicating UPF1 binding in regulation of their mRNA levels (Fig. 4A,B). No significant change was observed between two control lines (median fold change $<1.01$, not shown).

Notably, UPF1 also plays a role in other cellular process in addition to NMD, including Staufen-mediated mRNA decay (SMD). SMD is a translation-dependent

decay mechanism in which UPF1 is recruited to mRNA 3' UTRs via Staufen binding (Kim et al. 2005, Gong & Maquat 2011). Importantly, however, SMD is UPF2-independent (Kim et al. 2005). In order to further characterize the regulation of the UPF1-bound mRNAs in this study we took advantage of two recently published mouse RNA-Seq data sets investigating the role of UPF2 and the UPF1 kinase SMG1 in gene expression regulation (McIlwain et al. 2010, Weischenfeldt et al. 2012). The extent to which UPF1 binds to similar targets in different cell types has not been examined comprehensively. However, we observed significant de-repression of mRNAs bound by UPF1 in mESCs in data from Upf2 knockout liver and Smg1 knockout MEFs (median fold changes 1.28 and 1.20, respectively, both P<1E-4), further supporting a connection between UPF1 3' UTR binding and NMD (Fig. 4B,C and S4A). The 3' UTR binding that we observed does not appear to reflect the canonical dEJ-based NMD pathway, as genes encoding 3' UTR-bound mRNAs were not enriched for expression of dEJ-containing isoforms (Fig. 4D). However, we did observe that UPF1-bound 3' UTRs were on average 2145 nt in length, 70% longer than the average 3' UTR length of 1262 nt (difference for all versus bound, P = 3.5E-30, Fig. 4E).

Given that extended 3' UTR length is itself an NMD-triggering feature, we asked whether increased 3' UTR length could explain all of the de-repression associated with UPF1-bound mRNAs following NMD inhibition. For this analysis, we compared the degree of de-repression of genes bound by UPF1 to that for mRNAs similar 3' UTR lengths, controlling also for initial expression level. Interestingly, UPF1-bound mRNAs were de-repressed 1.10- to 1.16-fold relative to control gene sets with similar UTR lengths following UPF1 knockdown (P <5E-5, Fig. 4F, S4C and data not shown). While long UTRs were more likely to exhibit binding, a small subset of mRNAs with 3' UTRs less than 800 nt in length were also bound, and on average this set of genes was also de-repressed upon NMD inhibition, indicating that binding acts in the NMD pathway regardless of 3' UTR length (Fig. S4D,E). These observations suggest that while UPF1 binding may contribute to the association between 3' UTR length and NMD, UPF1 binding contributes directly to mRNA decay independent of

3' UTR length.

Given the high G content of UPF1-binding sites, we also asked whether UPF1-bound 3' UTRs were enriched for G-rich regions. We defined G-rich regions as 50 bp segments with G content within the top 5% of all such segments in 3' UTRs. Indeed, we found that UPF1-bound UTRs have on average nearly twice the number of G-rich regions per kb than do unbound UTRs with similar lengths and expression levels (Fig. 4G).

In order to determine whether UPF1-dependent regulation of mouse mRNAs bound via their 3' UTR is conserved, we assessed whether human homologs of genes encoding mRNAs bound by UPF1 in mESCs were similarly regulated. Interestingly we observed that human homologs of mouse UPF1 targets were significantly de-repressed following UPF1 depletion in two human cell lines, HeLa and U2OS cells (both $P < 0.01$, Fig. 4B,H and S4B) (Cho et al. 2012, Wang, Zavadil, Martin, Parisi, Friedman, Levy, Harding, Ron & Gardner 2011). These findings provide evidence that UPF1-dependent regulation of these genes is conserved in human cells.

### 2.2.8 Genes with low translational efficiency escape NMD

As NMD is a translation-dependent process, we asked whether the level of translational activity influenced susceptibility to NMD. For this purpose, we calculated the average ribosome density also referred to as translational efficiency (TE) (Ingolia et al. 2009) of each message by dividing the ribosome footprint density of the ORF by the RNA-Seq read density of this same region. We then analyzed the effects on mRNA stability of different NMD-associated features as a function of TE values. When comparing TE to gene expression changes, we calculated each measure using data from separate experiments to avoid an established source of false positives (Larsson et al. 2010), and used only consistently behaving mRNAs to enrich for NMD-related effects.

Overall, TE was positively correlated with de-repression following UPF1 knockdown, consistent with the known translation-dependence of NMD (Fig. 5A). Similar results were observed for the other NMD inhibition treatments (not shown). Furthermore, transcripts harboring dEJs derived from genes with very low TE failed to exhibit significant de-repression following UPF1 depletion (Fig. 5B). Similarly, UPF1 binding in the 3' UTR was not associated with significant de-repression following UPF1 depletion for genes with very low TE (Fig. 5C). Low TE genes are likely to encode mRNAs that are translated infrequently, reducing their sensitivity to NMD. For example, such mRNAs might undergo signal-induced translation but otherwise be held in a non-translating (but stable) state. Interestingly, little difference in 3' UTR length-dependent de-repression was observed between genes grouped by TE in the UPF1 depletion experiments (Fig. 5D). Thus, 3' UTR length-dependent NMD may be less reliant on robust translation than regulation based on UPF1 binding or presence of a dEJ. While these patters were somewhat more variable following CHX treatment (not shown), genes with higher TE values were more de-repressed overall after all treatments. Together, these data provide systematic evidence for modulation of NMD activity by translational efficiency.

## 2.3 Discussion

### 2.3.1 NMD-sensitive mRNA features

One goal of this study was to identify features that sensitize an mRNA to NMD. Using our data from mESCs, we observed that established NMD-triggering features long 3' UTR, presence of a dEJ, and presence of a uORF were predictive of consistent derepression in the NMD-inhibitory treatments in this system (Fig. 6A). We also observed that presence of a uORF with ribosome footprint coverage indicative of active translation was more predictive of regulation by NMD than mere occurrence of a uORF (Fig. 6A), emphasizing the need to assay uORF translational status.

Most notably, detection of UPF1 binding in a gene's 3' UTR was as predictive of NMD regulation as was presence of a long 3' UTR or dEJ, whether analyzed at the gene level (Fig. 6A; 37%, 22%, and 32% respectively of genes with these features were responsive to NMD-inhibiton) or at the level of individual mRNA isoforms (Fig. S5A; 15%, 15%, and 12% of respective isoforms). Furthermore, messages with longer 3' UTRs were more likely to be bound, but binding was predictive of NMD regulation independent of 3' UTR length (Fig. 4F). In total, ~30% of genes in the mESC transcriptome contains at least one of the four features characterized here (dEJ, long 3' UTR, translated uORF, or 3' UTR binding by UPF1) and these genes are 1.7-fold more likely to be regulated by the NMD pathway than are genes that do not encode any such features (Fig. 6A). It will be of interest to characterize additional mRNA properties that modify the efficacy of these features in triggering NMD.

We observed that presence of a dEJ is associated with increased mRNA repression regardless of 3' UTR length class (Fig. 1F), and addition of a dEJ to a long 3'UTR was reported to increase repression by NMD in some cases (Singh et al. 2008). We additionally found that the extent of repression associated with presence of a dEJ was greatest for mRNAs with short UTRs (Fig. 1F). This observation may reflect a degree of parallelism in the 3' UTR length- and dEJ-dependent branches of the NMD pathway such that destabilization triggered by one branch reduces the scope of repression achievable by the other branch.

### 2.3.2 UPF1 binds extensively in the 3' UTRs of a cohort of mRNAs

This study provides the first genome-wide identification of UPF1 binding sites within mRNAs. Given the canonical model in which UPF1 is recruited to specific mRNA targets via interaction with termination release factors and components associated with the EJC, and recent report of increased UPF1 association with reporter mRNA 3' UTRs that harbor EJCs as compared to those without (Kurosaki & Maquat 2013),

we initially hypothesized that the majority of CLIP sites would reside near PTCs and/or downstream EJCs. While we did observe a modest number of UPF1 CLIP reads near PTCs of dEJ-containing mRNAs (Fig. S3H), most such sites were not detected consistently between CLIP libraries, and thus were not emphasized here. mRNAs with dEJs may be more efficiently degraded than other classes of NMD targets, reducing the time window during which UPF1 is associated and making detection by CLIP more difficult. Detection of specific binding to PTC-containing splice forms produced by exon skipping is also likely to be extremely challenging given that only those reads that map to the exon skipping splice junction will uniquely identify this isoform.

The majority of UPF1 binding locations were distributed along the 3' UTRs of hundreds of mouse mRNAs (Fig. 3). This finding extends recent reports that UPF1 can associate with several, mostly exogenous, 3'UTRs (Hogg & Goff 2010, Kurosaki & Maquat 2013). Importantly, our genome-wide approach enabled us to find that UPF1 binding sites are not randomly distributed but are concentrated in the UTRs of specific mRNAs (Fig. 3C,D and Fig. S3D). One prior study proposed a 3' UTR length-sensing function for UPF1 (Hogg & Goff 2010) and we show that UPF1-bound 3' UTRs tend to be longer than average (Fig. 4E). However, we also show that UPF1 binding is associated with NMD-dependent repression in excess of that predicted by 3' UTR length (Fig. 4F), that UPF1 has specificity for certain UTRs (distinct from simple UTR length dependence), and that binding events that occur within short 3' UTRs are associated with NMD. Thus, our data indicate that message-specific features beyond 3' UTR length and dEJ presence determine UPF1 binding and associated mRNA decay.

RNA helicases often transit through or rearrange mRNP complexes associated with a wide variety of RNA sequences. Other RNA helicases bind at specific mRNA locations without recognizable primary sequence motifs (Sievers et al. 2012, Bohnsack et al. 2009). The absence of a clear UPF1 binding motif in the CLIP data was therefore not surprising. Two recent reports describing binding of the RNA helicase

eIF4AIII (the core component of the human EJC) confirmed an exceptional degree of positional specificity for regions 20-24 nucleotides upstream of exon-exon junctions (Sauliere et al. 2012, Singh et al. 2012) However, beyond a tendency for binding adjacent to SR protein binding sites, limited sequence specificity was observed.

UPF1's interactions with RNA are dynamic and regulated by both ATP binding and interaction with auxiliary factors. The CLIP assay likely captures multiple distinct states including UPF1 that is stably loaded onto the RNA before interaction with UPF2 or ATP (Chakrabarti et al. 2011), helicase-active UPF1 (post-UPF2 interaction) (Melero et al. 2012), and UPF1 that is actively involved in disassembly of mRNPs as degradation progresses (Franks et al. 2010). UPF1-bound locations displayed increased G content (Fig. 3E). Interestingly, earlier studies using purified hUPF1 found that its ATPase activity is more than 4-fold less active when in the presence of poly(rG) than for any other ribohomopolymer tested (Bhattacharya et al. 2000), suggesting that UPF1 may preferentially pause at G-rich regions. Together, our data indicate that UPF1 resides in particular mRNAs, often in G-rich regions, and localizes mostly to 3' UTRs unless translation is inhibited.

It is interesting to consider how messages targeted by UPF1 binding to the 3' UTR are degraded. Our results in mESCs together with work by other labs in human cell lines indicate that this process is not only UPF1-dependent, but also translation-dependent (Figs. 4B and 5C) and (Kurosaki & Maquat 2013). Additionally, the strong association that we observed between UPF1 binding and de-repression in UPF2 knockout tissue (Figs. 4B,C) suggests that our findings are predominantly relevant to NMD rather than SMD. How might additional NMD factors, in particular UPF2, be recruited to typical UPF1-bound 3' UTRs devoid of exon-exon junctions? One possibility is that EJC components may interact with some 3' UTRs independently of presence of a canonical EJC, thus positioning them to help recruit UPF1 to 3' UTRs for decay. Alternatively, UPF1 bound to 3' UTRs might sometimes interact with free cytoplasmic UPF2 and/or UPF3. Notably, a recent study reported that the degree of UPF1 association with mature mRNA was not affected by depletion of

eIF4AIII, the primary RNA-binding component of the EJC, suggesting that UPF1's association with mRNA may not require fully assembled EJCs (Singh et al. 2012). We suggest that multiple modes of NMD regulation cooperate to control mRNA abundance in mESCs, including both 1) a canonical dEJ-dependent mode, in which UPF1 is recruited to terminating ribosomes and its phosphorylation and ATPase activity are triggered by interaction with stably associated downstream EJC-associated UPF2 and UPF3, and 2) a 3' UTR targeting mode, wherein UPF1 associates with 3' UTRs of messages and interacts with soluble UPF2 and/or UPF3 or with these factors bound to 3' UTRs in the absence of an EJC (Fig. 6B). Interestingly, the proposed 3' UTR targeting mode of regulation could occur after the initial rounds of translation of the message, raising the possibility that NMD could be used to deplete bulk pools of post-pioneer messages.

Our CLIP-Seq analysis of UPF1 binding appears to be far from saturating (Fig. S3G), indicating that UPF1 may bind several hundred or even thousands of mRNAs. Undetected binding by UPF1 may account for a substantial portion of the genes derepressed by NMD-inhibition which lack detectible targeting features (Table S3; available online). Notably, we observed that mRNAs bound in ESCs appear to be similarly regulated in MEFs and in mouse liver (Fig. 3B,C), suggesting that binding to the same or similar sets of mRNAs occurs in other cellular contexts. Intriguingly, human homologs of genes bound by UPF1 in mouse were similarly regulated by UPF1 in human cells (Fig. 4C,D), indicating a conserved and potentially functional role for UPF1 regulation of this set of genes. Together these findings posit that association of UPF1 with the 3' UTRs of many mESC mRNAs plays a widespread role in gene expression and that this mode of regulation is likely conserved in many cell types and organisms.

### 2.3.3 NMD regulation via translation of uORFs

Here we have also identified uORFs that undergo translation transcriptome-wide and observed an association of this activity with UPF1-dependent mRNA repression. The extent of repression attributable to tuORFs was somewhat less than that seen for other NMD regulatory features. This difference may reflect that uORFs are not always translated prior to translation of the main ORF, with the extent of uORF translation in the pioneer round of translation serving to regulate the extent of NMD. Translation of uORFs after the pioneer round may also contribute to NMD, by inhibiting translation of the main ORF and thereby increasing the accessibility of the coding region to UPF1 and/or extending the effective length of the 3' UTR. Regulation of translation of the main ORF via uORF translation is known to be important in the context of cellular stresses, including hypoxia and endoplasmic reticulum stress (Vattem & Wek 2004, Gaba et al. 2001). Our data suggest that regulated translation of uORFs may also commonly contribute to regulation of mRNA stability via the NMD pathway, thereby reinforcing repression at the translational level (Calvo et al. 2009). We found that, as a group, transcriptional regulators are de-repressed in response to NMD inhibition, and this group is also enriched for tuORFs (Table S3; available online and Fig. 2D), suggesting that tuORF-dependent NMD is a common regulator of the expression levels of TFs.

### 2.3.4 Role of NMD in the mESC transcriptional program

Efficient depletion or elimination of several NMD components results in multisystemic developmental abnormalities and eventual embryonic lethality in zebrafish (Anastasaki et al. 2011, Wittkopp et al. 2009), Drosophila (Metzstein & Krasnow 2006, Avery et al. 2011) and mice (McIlwain et al. 2010, Medghalchi et al. 2001, Weischenfeldt et al. 2008), although the strict requirements for different components vary between organisms (Frizzell et al. 2012, Hwang & Maquat 2011). Importantly, several of these components have additional cellular functions which may contribute to these pheno-

types (Frizzell et al. 2012, Varsally & Brogna 2012, Azzalin & Lingner 2006, Kim et al. 2005). However, their common function in the NMD pathway has led to the hypothesis that key developmental transcripts are NMD targets whose aberrant expression in the absence of an intact NMD pathway disrupts normal embryogenesis.

Notably, examination of available ChIP data revealed that 116 genes that were de-repressed following NMD-inhibition have been previously identified as targets of the OCT4 TF, a master regulator of pluripotency ($P < 0.005$ by hypergeometric test, Table S3; available online) (Kim et al. 2008). This observation could indicate a general rewiring of the transcriptional network of mESCs following NMD inhibition. Mis-expression of key developmental regulators has been shown to disrupt cell fate and ultimately embryonic development (Niwa et al. 2000, Kopp et al. 2008, Nichols et al. 1998). Several mRNAs encoding developmentally relevant TFs were de-repressed following NMD inhibition, including the Notch signaling partner Rbpj (Imayoshi et al. 2010), the transcriptional co-repressor Sirt1 (Wang, Sengupta, Li, Kim, Cao, Xiao, Kim, Xu, Zheng, Chilton, Jia, Zheng, Appella, Wang, Ried & Deng 2008), and the pluripotency and reprogramming factor Klf4 (Takahashi & Yamanaka 2006). Interestingly, mRNAs of several developmental regulatory TFs were observed to contain tuORFs, including Nanog (observed previously by (Ingolia et al. 2011)), Klf9, Ncor1 (Jepsen et al. 2000), and Tbx3 (Ivanova et al. 2006, Lu et al. 2011, Davenport et al. 2003), the last 3 of which were de-repressed upon NMD inhibition. While transcriptional networks are complex and discriminating direct regulation from indirect effects is difficult, it is worth exploring whether modulation of these factors' levels by NMD contributes to the developmental abnormalities observed in NMD mutant mice.

Our findings, including the recognition of UPF1 binding to 3' UTRs as a widespread NMD targeting determinant, the identification of hundreds of direct NMD targets, and delineation of the relationships between mRNA translation and NMD susceptibility, provide a context for understanding the role of UPF1 and NMD in development and transcriptome control.

## 2.4 Figures

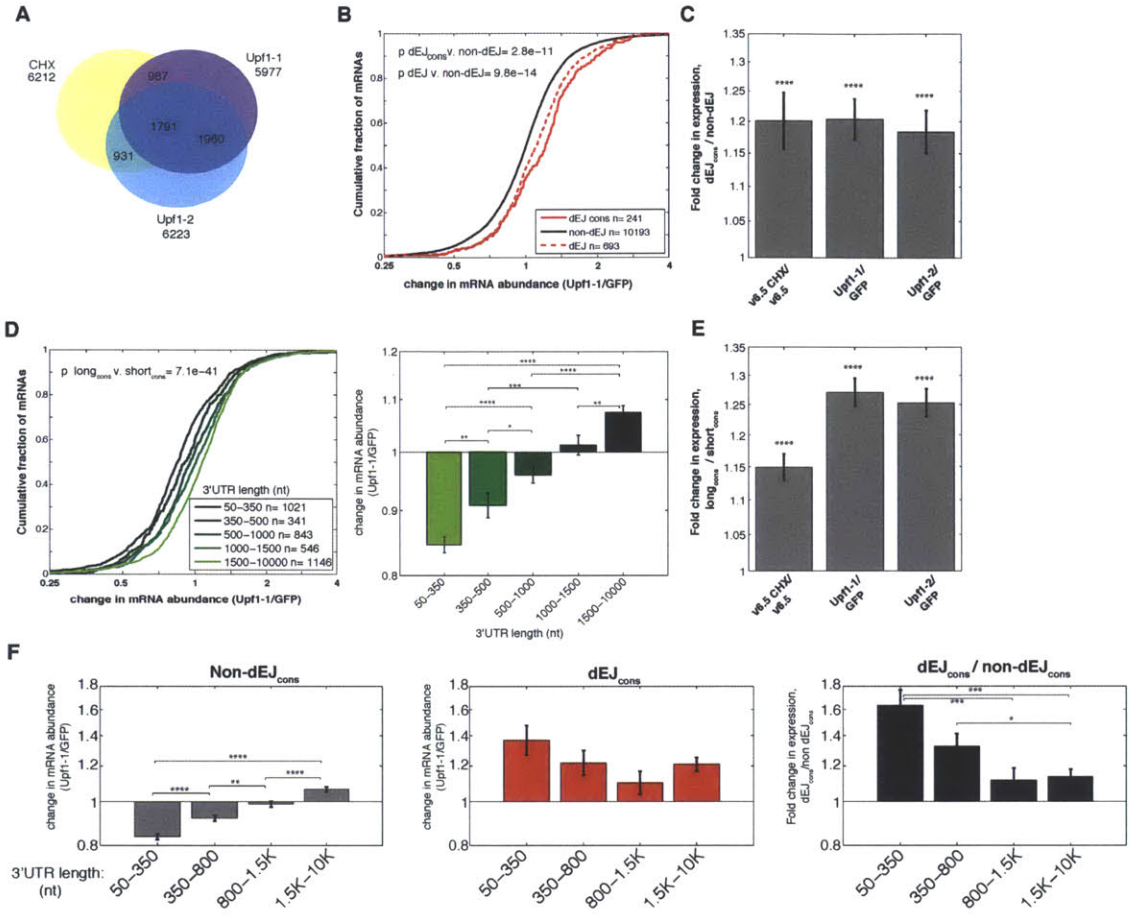



Figure 2-1: Consistent de-repression of hundreds of mRNAs with and without canonical NMD features occurs following UPF1 depletion and translational inhibition.

Figure 2-1 (A) Overlap of mRNAs that changed expression by >1.1-fold in the same direction in each of three NMD inhibition experiments (shRNA Upf1-1, shRNA Upf1-2, and CHX treatment). (B) Depletion of UPF1 or translational inhibition results in derepression of mRNAs harboring dEJs. Cumulative distribution functions of changes in mRNA abundance (log2) following Upf1 depletion (shRNA Upf1-1) for mRNAs either with an annotated dEJ (red) or without an annotated dEJ (black). dEJ mRNAs that behave consistently between the three NMD inhibition experiments (dEJcons) are shown in solid, all dEJ mRNAs are displayed in hashed line. X-axis is plotted on a Log2 scale and P-value determined by Wilcoxon rank sum test. (C) Bar graph illustrates difference in median fold change of abundance of dEJcons and non-dEJ isoforms for the three NMD inhibition experiments. Y-axis is plotted on a Log2 scale. Differences ranged from between (1.18- to 1.20-fold) in the 3 experimental conditions. Error bar represents standard error of the two populations compared. P values determined as in C for each experiment. (D) mRNA derepression following UPF1 depletion or translational inhibition is dependent on 3'UTR length. (D) As in (B) for isoforms behaving consistently (cons) with different annotated 3' UTR lengths (different green lines). Bar graph shows median fold change in expression and standard error of mRNAs with different 3' UTR lengths. See also Fig. S1F. (E) As in (C) for mRNAs with long (1500-10,000 nt) versus short (50-350 nt) 3' UTRs. Differences ranged from between (1.15- to 1.27-fold) in the 3 experimental conditions. (F) Interaction between dEJ and 3' UTR length for consistent isoforms. Bar graph illustrates fold change in expression of mRNAs with different 3' UTR lengths without (left) and with (middle) an annotated dEJ. Difference in expression change between mRNAs with and without an annotated dEJ of a given 3' UTR length (right). Repeated random permutation without replacement of the expression change values of mRNAs between 3' UTR length bins was used to estimate the null distribution of median values (n=2,000). The significance of the observed difference was assessed by comparison to this distribution. This trend was also observed when comparing expression changes between each dEJ isoform and non-dEJ isoforms with equivalent 3' UTR lengths (10%) following NMD inhibition (Fig. S1H). Y-axes is plotted on a Log2 scale.

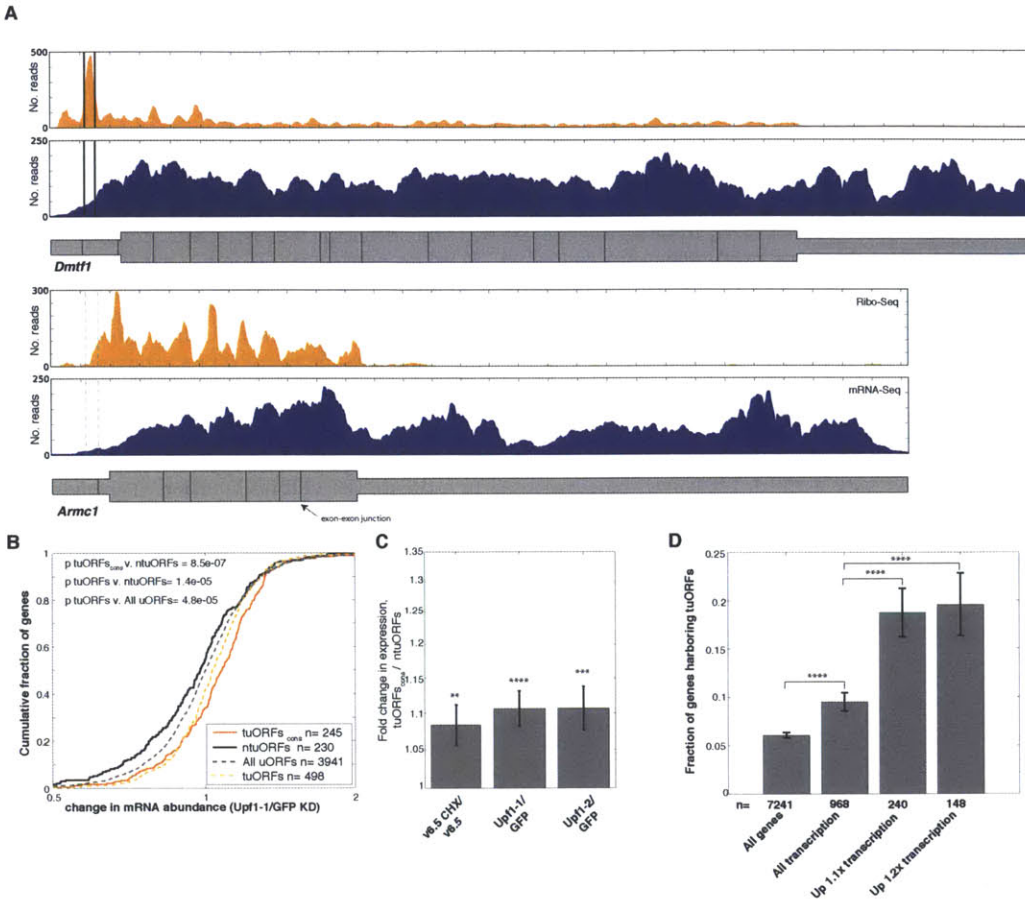



Figure 2-2: Translation of uORFs is associated with UPF1-mediated repression.

Figure 2-2 (A) Classification of translated uORFs (tuORFs) and untranslated uORFs (ntuORFs). mRNA-Seq (blue) and ribosome footprint (orange) reads from UPF1 depleted (shRNA Upf1-1) cells mapping to Dmtf1 and Armc1 mRNAs. Dmtf1 (top) contains a tuORF (outlined with dark grey lines) while Amrc1 contains an ntuORF (hashed lines). Note that Dmtf1 harbors an additional ATG within the tuORF (not marked). (B) Cumulative distribution functions of changes in mRNA abundance (log2) following UPF1 depletion (shRNA Upf1-1) for all genes with a tuORF (yellow dashed), consistent genes with a tuORF (orange), genes with a uORF (gray dashed), and genes with an ntuORF (black). P values determined by Wilcoxon rank sum test. (C) Difference in median LFC in mRNA abundance between consistent genes with a tuORF and genes with an ntuORF for three NMD inhibition experiments. Error bars represent standard error of the two populations compared. P values determined as in (B). See also Fig. S2. (D) Fraction of genes harboring a tuORF for all expressed genes, all expressed transcriptional regulators, and transcriptional regulators de-repressed by more than 1.1- or 1.2-fold in at least 2 out of 3 NMD inhibition experiments. Transcriptional regulators defined as GO:0045449. P-value of enrichment determined by hypergeometric test. Numbers of genes in each category are indicated. Error bars indicate binomial standard deviations. Asterisks as in Figure 1.

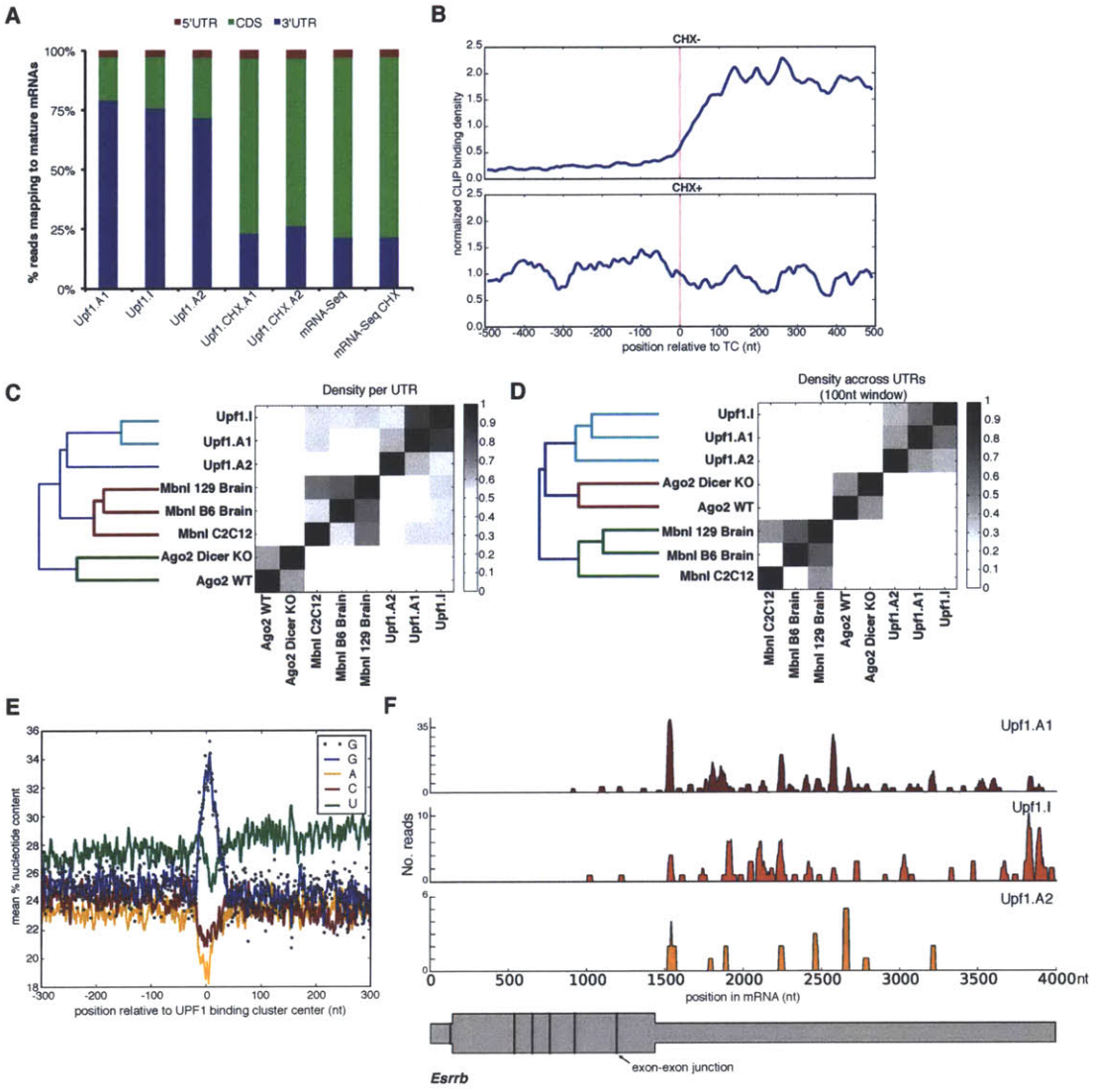



Figure 2-3: UPF1 interacts predominantly with 3' UTRs of mature mRNAs.

Figure 2-3 (A) Distribution of reads mapping to 5' UTRs, coding sequence (CDS), and 3 ' UTRs of mature mRNAs for different UPF1 CLIP experiments using RNAse A (A) or RNAse I (I) and mRNA-Seq experiments, with or without CHX treatment. See also Fig. S3A. (B) Metagene plot of average UPF1 CLIP tag density per gene in 500 nt regions flanking the TC (red line) under normal (top) and CHX (bottom) conditions. Density was smoothed with a 10 nt Gaussian. (C,D) Correlation of UPF1 CLIP samples binding in 3' UTRs of genes with minimum FPKM (fragments per kilobase per million mapped reads) of 50. Correlations of MBNL1 CLIP data in mouse C2C12 cells and two mouse brain samples (Wang, Cody, Jog, Biancolella, Wang, Treacy, Luo, Schroth, Housman, Reddy, Lecuyer & Burge 2012) and of AGO2 CLIP data in wildtype and Dicer null mESCs (Leung et al. 2011) are shown for comparison. In C, correlation coefficients were calculated between densities of CLIP binding over all UTRs. In D, correlation coefficients were calculated between densities over 100 nt windows across all UTRs. Greyscale emphasizes higher values in C. (E) Percent homonucleotide content is shown for 600nt surrounding UPF1 binding sites within 3' UTRs. Values were averaged over all UPF1 binding sites which fall within the 3' UTR. Dot markers indicate average G content for all sites at each position. Lines indicate moving average of values for each nucleotide using a 3 bp sliding window. (F) UPF1 binding to the Esrrb mRNA in 3 CLIP experiments in untreated cells. Schematic of Esrrb mature mRNA displayed below. Vertical black bars indicate exon-exon junctions. Width of grey bar indicates CDS and UTR regions.

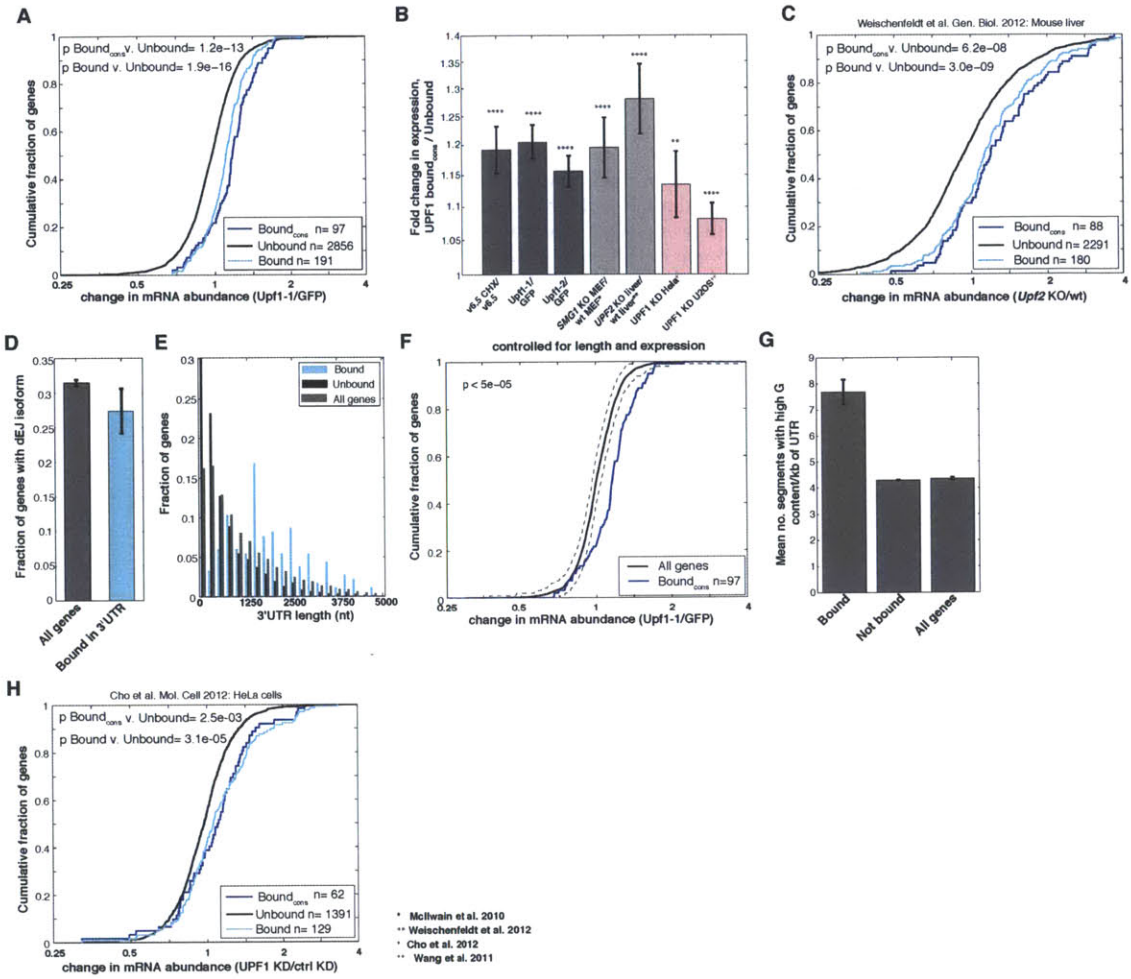



Figure 2-4: 3' UTR binding by UPF1 is associated with mRNA repression.

Figure 2-4 (A) Cumulative distribution functions of gene expression changes following UPF1 depletion (shRNA Upf1-1) for consistently behaving genes bound by UPF1 in the 3' UTR (blue), all genes bound by UPF1 in the 3' UTR (cyan), and unbound genes (black). (B) Difference of median LFC in mRNA abundance (log2) between consistently behaving genes bound by UPF1 in the 3' UTR and unbound genes for three NMD inhibition experiments in mESCs (dark grey), as well as for Smg1 KO in MEFs (McIlwain et al. 2010), and Upf2 KO in mouse liver (Weischenfeldt et al. 2012) (light grey), and for human homologs of these genes following UPF1 depletion in HeLa (Cho et al. 2012) and U2OS cells (Wang, Wengrod & Gardner 2011) (pink). Error bar represents standard error of the two populations compared. (C) As in (A) except mRNA abundance measurements were made in wild type and Upf2 KO mouse liver (Weischenfeldt et al. 2012). (D) Fraction of all genes and genes bound by UPF1 in the 3 ' UTR that have an annotated isoform harboring a dEJ. Error bars indicate binomial standard deviation. (E) Distribution of 3' UTR lengths of genes bound by UPF1 in their 3' UTRs. Lengths were assigned based on the best-annotated isoform for each gene. (F) Cumulative distribution function of changes in mRNA abundance (log2) following UPF1 depletion (shRNA Upf1-1) for consistently behaving genes bound in the 3' UTR by UPF1 and genes sampled with replacement to match the distribution of expression levels and 3' UTR lengths. Sampling was repeated and the mean cumulative distribution function is shown in black; 95% confidence intervals are shown in grey. P value was calculated as the number of times the mean expression change of the matched population exceeded that of the consistent UPF1 bound genes, corrected for number of iterations performed (n = 20,000). Results were similar using shRNA Upf1-2 and CHX treatment (data not shown). (G) Mean number of regions with high G content per kb of 3' UTR for different groups of genes. Fraction G content was calculated for 50 base pair sliding windows (10 bp step) across all 3' UTRs of all of all annotated 3' UTRs without exon-exon junctions. Mean number of windows per kb of UTR that passed the 95th percentile (of genome) for UPF1 bound genes, 1000 randomly selected sets of genes not bound by UPF1 but selected for similar expression levels and lengths, and all genes in total is plotted. Error bars represent standard error of population of UPF1-bound UTRs and all UTRs of the genome. Error bars for UTRs not bound by UPF1 represent standard error of the means of the controlled gene sets. (H) As in (A) except gene names were identified by homology to mouse genes either bound or unbound by UPF1 in their 3' UTR and mRNA abundance measurements were made in control- and UPF1-depleted HeLa cells (Cho et al. 2012). Asterisks as in Figure 1. See also Fig. S4.

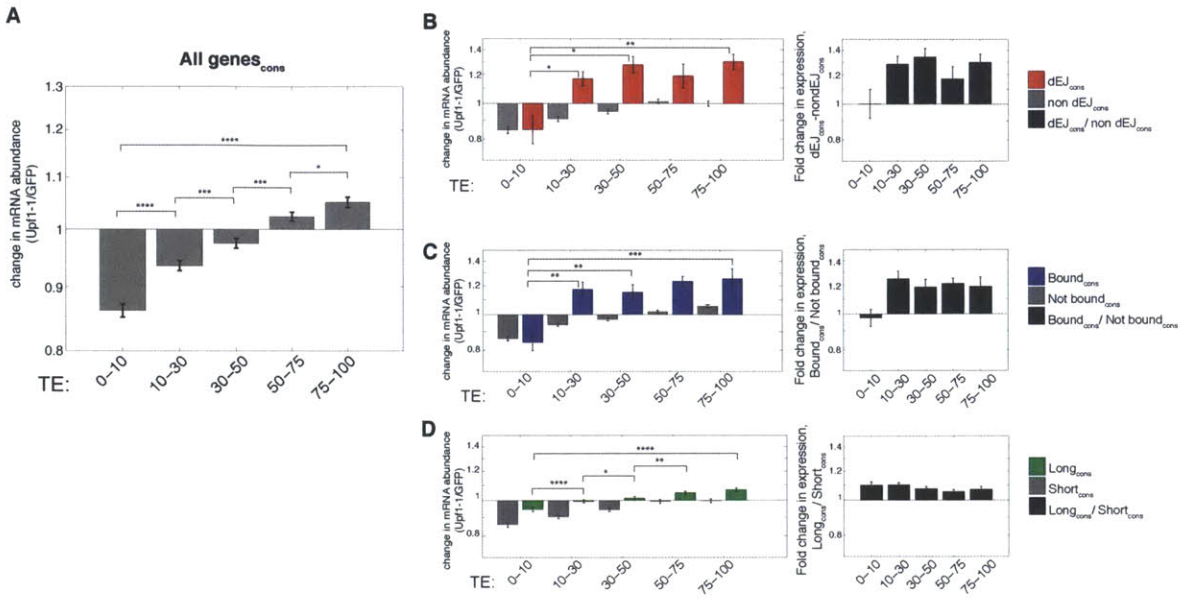



Figure 2-5: Relationship between TE and NMD-triggering gene features.

Figure 2-5 (A) Median fold change of mRNA abundance (log2) following UPF1 depletion (shRNA Upf1-1) for consistently behaving genes by varying translational efficiency (TE), grouped by percentile rank. Translational efficiency of each gene's ORF was determined in wildtype cells as ribosome footprint density/mRNA-Seq read density. P values calculated using Wilcoxon rank sum test. Results were similar using shRNA Upf1-2 and CHX treatment (not shown). (B) As in (A) except for non-dEJ (grey) and dEJ (red) mRNAs. Difference in median fold change (log2) in expression following UPF1 depletion between dEJ and non-dEJ mRNAs with given TE is shown at right (black). For dEJ calculation, expression changes were calculated on an isoform level and TE was assigned based on the TE of the full ORF. (C,D) As in (B), except for UPF1 binding in 3' UTRs (C, blue) and 3' UTR length (D, green). Long and short 3' UTRs were defined as 1500-10,000 nt and 50-350 nt respectively. Significance of differences in expression changes between TE bins of non feature-containing genes (non-dEJ, not bound, and short 3' UTR) were similar to those of all genes in (A). Asterisks as in Figure 1.

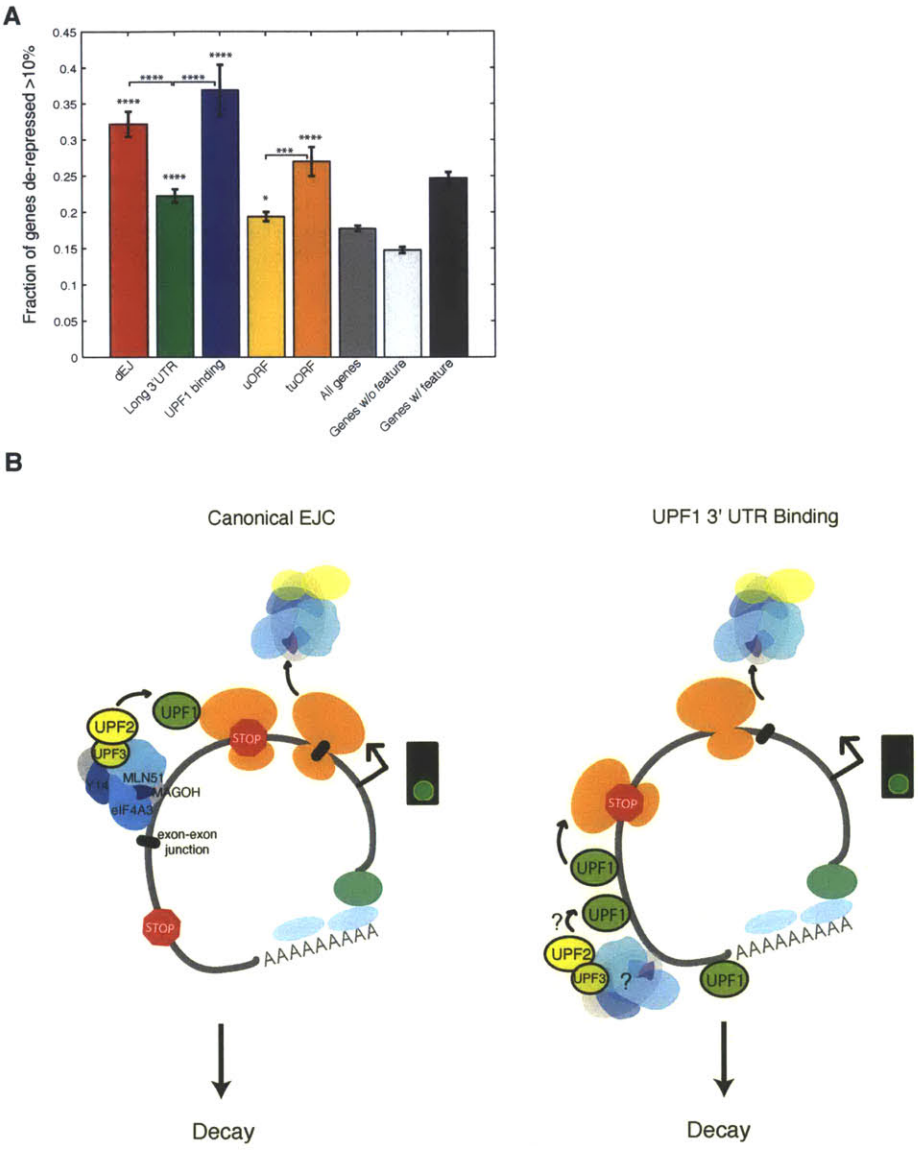



Figure 2-6: NMD features and models of UPF1-dependent mRNA repression.

Figure 2-6. (A) Predictive capacity of mRNA features for NMD-regulation. Fraction of expressed genes harboring a feature that were de-repressed consistently is shown. Fraction of genes without any NMD-inducing feature that were consistently de-repressed is also shown (white bar). P values above each feature indicate significance relative to all expressed genes; brackets indicate significant comparisons between features (hypergeometric test). See also Fig. S5A. Asterisks as in Figure 1. (B) Left: Canonical dEJ- mediated regulation. Transient recruitment of UPF1 to ribosomes terminating upstream of an exon junction complex (EJC). EJC components (blue and grey) are deposited as a consequence of splicing in the nucleus ~24 nucleotides upstream of an exon-exon junction (black bar). Members of the EJC, including UPF2 and UPF3, help to stabilize the UPF1-ribosome interaction as well as to stimulate it's phosphorylation and helicase activity ultimately leading to decay of the message. Right: 3' UTR UPF1 binding-mediated regulation. UPF1 binds to mRNA 3' UTRs independent of the presence of an exon-exon junction. At some frequency UPF1 is activated by interaction with cytoplasmic EJC components. These factors may either be recently released from mRNAs due to translation or perhaps stably associated with message 3' UTRs independent of an exon-exon junction.

## 2.5 Supplemental Figures

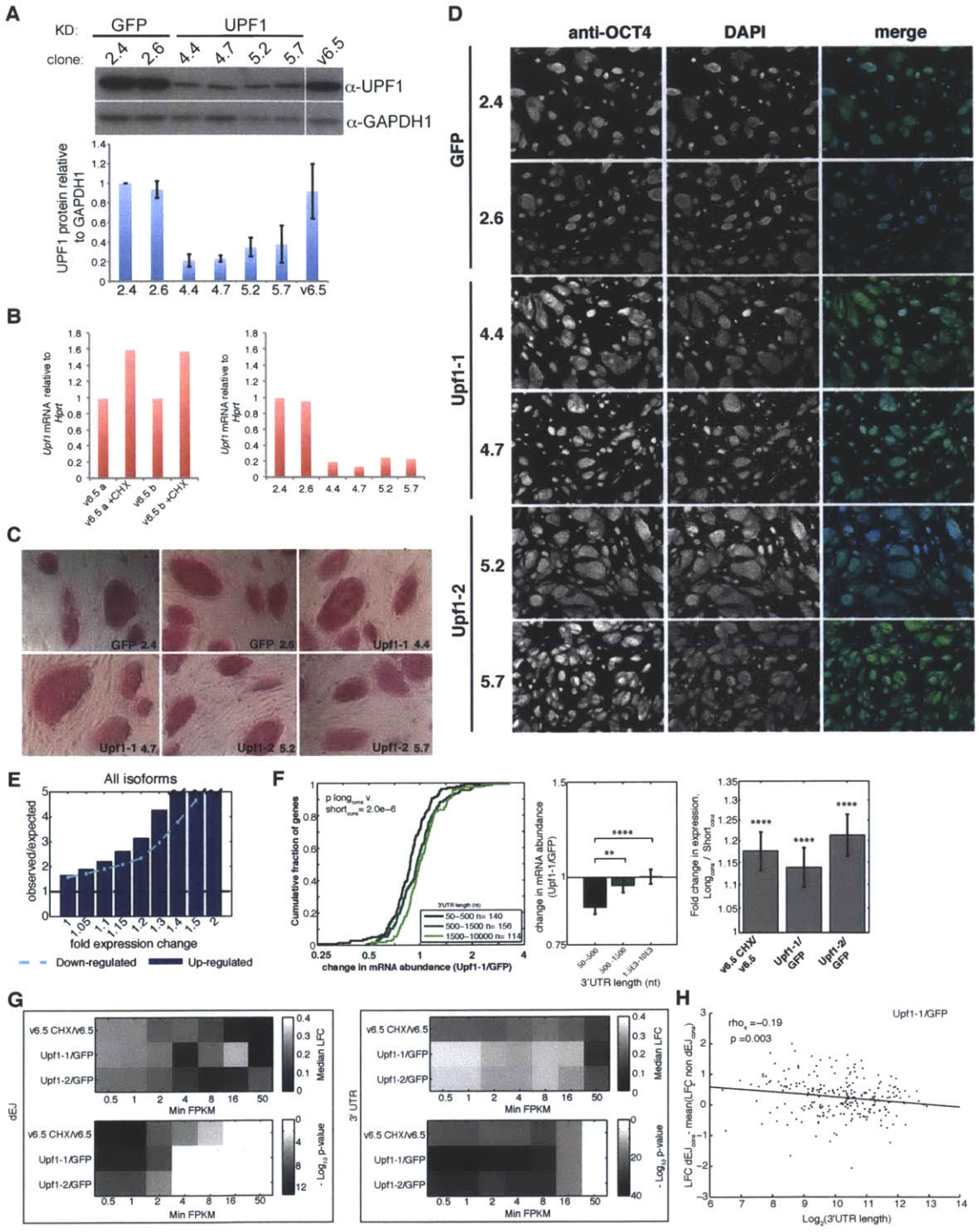



Figure 2-S1: Derivation of mESCs stably depleted of UPF1 and effect of constitutive, long 3' UTRs on gene expression upon NMD-inhibition, related to Fig. 1.

Figure 2-S1. (A) Stable depletion of UPF1 protein from mESCs. Two clones from each of three lentiviral infections (Upf1 shRNA-1, clones 4.4 and 4.7, Upf1 shRNA-2, clones 5.2 and 5.7, or GFP shRNA, clones 2.4 and 2.6) were isolated. Equivalent levels of total protein were loaded for Western blot analysis (representative blot shown above). Quantitation was performed using ImageJ software. Mean and standard error of three technical replicate Western blots are shown (below). (B) Assessment of Upf1 mRNA levels by real-time PCR in stable cell lines from (A) and biological replicates of wildtype v6.5 cells with and without 2 hours of treatment with cycloheximde (+CHX). Upf1 mRNA levels were normalized to those of Hprt and ratios are plotted relative to levels in line 2.4 or untreated cells. (C) Alkaline phosphatase staining of mESC clones. mESCs were cultured on MEFs and then fixed and stained for alkaline phosphatase levels after removing puromycin from growth media for 48 hours. Hairpins and clone names are indicated for each panel. Magnification = 10X. (D) Oct4 levels in mESC clones. Cells were fixed and stained using anti-OCT4 antibody (left) or DAPI (middle) and images were overlayed (right). Hairpins used in generating clones are indicated to the left. OCT4 is selectively expressed in mESCs and not in underlying MEFs. Magnification = 4X. (E) Enrichment of overlap in mRNAs between NMD inhibition experiments over what would be expected by chance, assuming independence of each experiment, for increasing fold change requirements (x-axis). Overlap of mRNAs that increased in expression, blue bars. Overlap of mRNAs that decreased in expression, cyan line. (F) Genes with constitutive 3' UTRs and without 3' UTR exon-exon junctions are regulated by NMD in a length-dependent manner. Only consistently behaving genes that had one annotated 3' UTR without any junctions were considered. Ribosomal genes, which tend to be highly expressed and to have shorter 3' UTRs were also removed to ensure there were no confounding effects of this gene class in the analysis. mRNA expression was summed over all isoforms per gene and used to estimate abundance in each experiment. Left: Cumulative distribution functions of change in abundance following UPF1 depletion (Upf1-1) of genes with different length 3' UTRs. Middle: Median LFC (log2) for each group of genes displayed in top as well as delta LFC for longer (>1500 nt) versus shorter (50-500 nt) UTRs. Error bars represent standard error of the population or populations considered. Right: Change in LFC of long versus short UTRs for different NMD-inhibition experiments. Error bars as above. P-values displayed as in Fig. 1. (G) Well-expressed dEJ isoforms and isoforms with long 3'UTRs are significantly de-repressed following NMD inhibition. P-values (as determined by Wilcoxon ranksum test) and median log2 fold changes between dEJcons and non-dEJ isoforms or between mRNAs with $\text{long}_{cons}$ (>1500 nt) versus $\text{short}_{cons}$ (50-500 nt) 3'UTRs are shown after requiring specific minimum expression levels for all isoforms considered (x-axis). In some cases, the fold changes observed appear to decrease from that observed at a lower threshold. This may indicate that most true NMD targets are lowly expressed, and those that are higher expressed have actually acquired features that protect them from NMD regulation. (H) Scatter plot of length of 3'UTR versus change in de-repression associated with dEJs following UPF1 depletion (Upf1-1/GFP). All mRNAs harboring a dEJ were matched with non-dEJ isoforms possessing 3' UTRs of similar lengths (10%). Y-axis plots the difference in expression change between the dEJ isoform and the mean of the matched non-dEJ isoforms. Spearman correlation and p-value are given. Line represents least squares fit of the data.

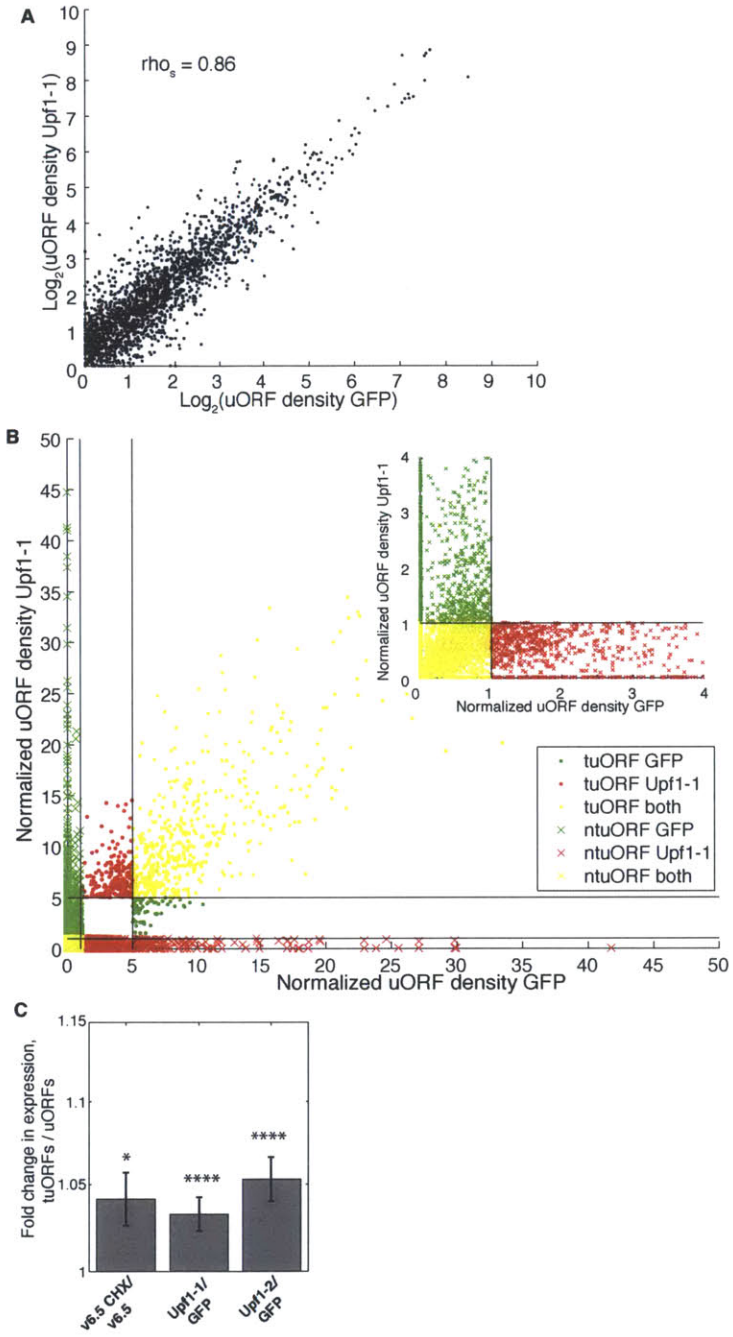



Figure 2-S2: Correlation of ribosome footprint densities of uORFs calculated in UPF1- and control-depleted mESCs, related to Fig. 2.

Figure 2-S2. (A) Density of ribosome footprints was determined in uORFs lying completely upstream of annotated translation start and Spearman correlation was determined after pseudo count of 0.0001 was applied to densities with value of 0. (B) As in (A) except that only uORFs that were called as translated or untranslated in either the UPF1-depleted or control cells are plotted and density values were normalized by the density of the surrounding background regions (for ntuORFs this is the higher of the footprint density in the flanking 60 nt or 1/60, for tuORFs this is the higher of the density in the flanking 60 nt or 2/3 see Methods). Black lines indicate normalized densities of 1 and 5 which were used as thresholds for classifying ntuORFs and tuORFs, respectively. Inset at top right is expanded view of low density region. (C) Bar graph illustrates difference in median log2 fold change (LFC) of abundance between tuORF genes and genes without any annotated uORFs for the three NMD inhibition experiments. Error bar represents standard error of the two populations compared. P-values displayed as in Fig. 1.

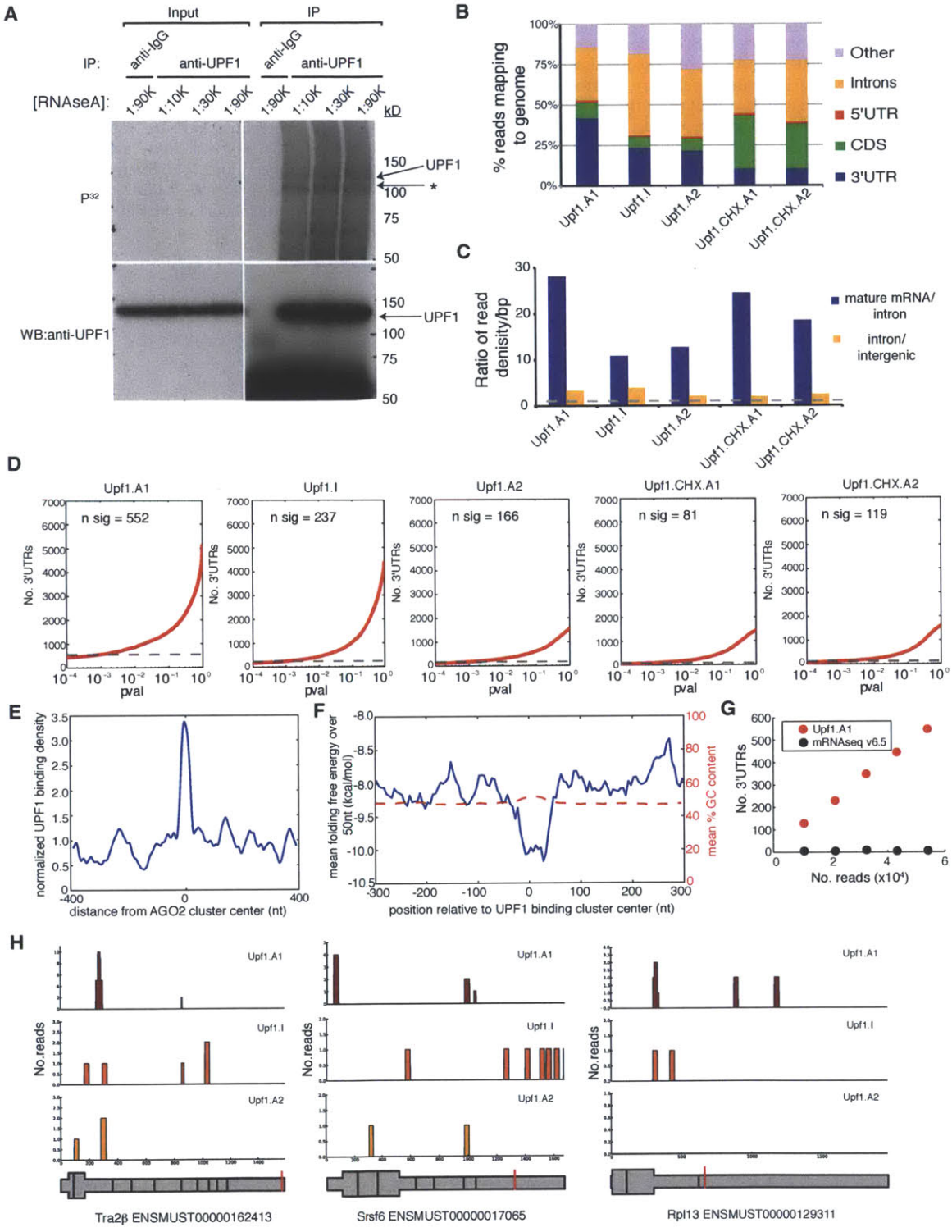



Figure 2-S3: UPF1 CLIP read densities and specificity of UPF1 binding, related to Fig. 3.

Figure 2-S3. (A) Isolation of UPF1 bound RNAs. UPF1-RNA complexes that migrated at 140-155 kDa (immediately above anti-UPF1 WB band and upper autoradiography band) were isolated from nitrocellulose. Nitrocellulose slices from 1:30 and 1:90K dilutions of RNAseA were combined. Nitrocellulose slices containing potential IgG complexes (treated with 1:90K RNAseA) were harvested in parallel from similar positions. * Indicates unknown RNA species migrating below UPF1 protein. (B) Locations of UPF1 CLIP reads. IgG subtracted, unique UPF1 sequences that mapped uniquely to the genome were further mapped to mRNA annotations. Order of calling to prevent duplicate calls: 1) CDS, 2) 3' UTR, 3) 5' UTR, 4) Intron, 5) Other. (C) UPF1 CLIP read densities. Density of reads per mature mRNA was calculated by summing coverage of reads mapping to 5'UTRs, CDS, and 3'UTRs and dividing by the summed total length of each region in the genome. Reads were counted as intronic if they mapped within a gene boundary but not within any exon and intergenic if they mapped outside of known gene boundaries. (D) P values calculated for binding to 3'UTRs in each UPF1 CLIP library. Hashed grey line marks the number of genes that has a P value of UPF1 binding in the 3' UTR $<0.001$ (y-intercept is indicated as number of significant calls, listed at top). (E) Density of UPF1 binding peaks at known AGO2 binding clusters in 3' UTRs. UPF1 CLIP read density was combined from multiple UPF1 CLIP libraries and compared to the location of known AGO2 binding within 3' UTRs, as determined in (Leung et al. 2011). Data was smoothed using a 10 nt Gaussian. (F) Prediction of the free energy of RNA folding within the 600 nt surrounding sites of UPF1 binding. Free energy predictions were calculated using a 50 nucleotide sliding window over these regions and were averaged over all UPF1 binding sites (Lorenz et al. 2011). Regions used in calculations were filtered to include only those that fall entirely within the annotated 3'UTR. This local change in free energy was not found to be significant when compared to that associated with regions of similar GC content nearby UPF1 binding sites. (G) Estimation of degree of saturation of 3' UTRs bound by UPF1 in Upf1.A1 CLIP library. The observed low degree of saturation may reflect transient nature of UPF1-RNA interactions. Upf1.A1 and mRNA-Seq reads were randomly sampled to reveal read coverage in 3'UTRs equivalent to that of 1x, 4/5x, 3/5x, 2/5x, and 1/5x the Upf1.A1 CLIP library. When mapping to 3' UTRs, uniqueness on positions was not enforced for mRNA-Seq reads (as it was for UPF1 CLIP samples, see Methods). This was reasoned to yield a more lenient estimate of the number of 3' UTRs that could be called bound by mRNA-Seq. Sampled reads were then used to call 3' UTRs bound with P-value $<0.001$. (H) dEJ mRNAs with UPF1 binding near PTC. Examples of dEJ isoforms that displayed UPF1 binding in one or more UPF1 CLIP libraries (CHX-) are shown. Width of grey bar illustrates CDS and UTR regions. Vertical black bars indicate exon-exon junctions and red line marks site corresponding to TC of non dEJ isoform for each respective gene. Note that for Tra2B, the dEJ isoform shown has an annotated cleavage site that is upstream of that which is annotated for the non dEJ isoform. Thus the position of the TC shown is very close to the end of the gene.

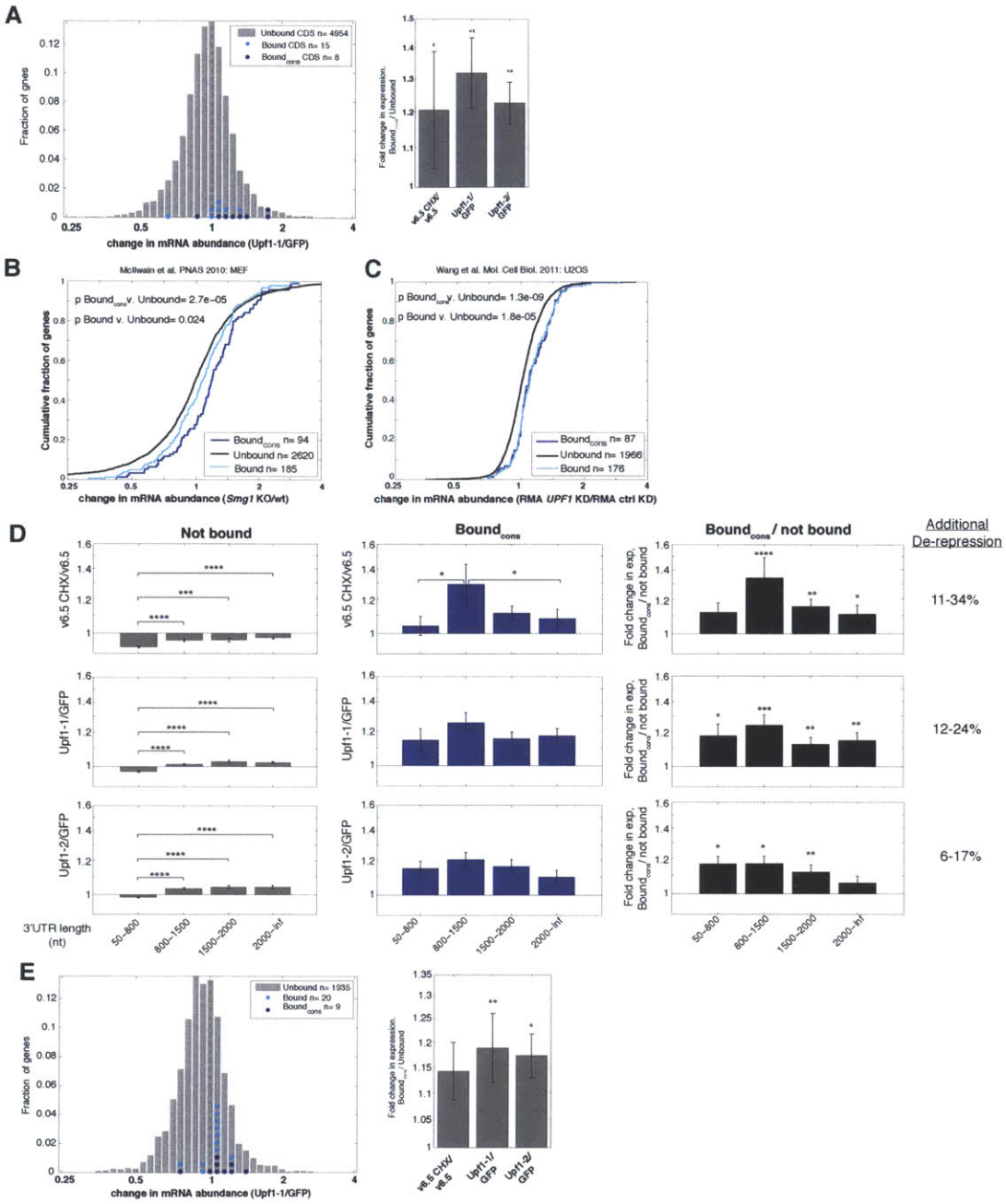



Figure 2-S4: UPF1 binding in CDS and 3' UTRs of varying lengths is associated with mRNA repression, related to Fig. 4.

Figure 2-S4. (A) Above, change in expression following UPF1 depletion (Upf1-1) for consistent genes with CDS bound by UPF1 (Boundcons, blue ), all genes with CDS bound by UPF1 (Bound, cyan), and genes with CDS not bound by UPF1 (gray bars). Size of dot reflects number of genes falling within this bin of expression change. Below, delta LFC in mRNA abundance between consistent genes with UPF1 binding in CDS and unbound genes. P values determined as in (A) and displayed as described in Fig. 1. (A) Cumulative distribution functions of gene expression changes between Smg1 KO MEFs and wildtype MEFs (McIlwain et al. 2010) for consistently behaving genes bound by UPF1 in the 3' UTR (blue), all genes bound by UPF1 in the 3' UTR (cyan), and unbound genes (black). (B) As in (A) except gene names were identified by homology to mouse genes either bound or unbound by UPF1 in their 3' UTR and mRNA abundance measurements were made in control- and UPF1-depleted U2OS cells (Wang et al. 2011). (C) Change in expression of consistently behaving genes bound by UPF1 in the 3' UTR compared to all genes not bound following NMD inhibition, binned by 3'UTR length. Left. Median LFC in expression and standard error of genes not bound by UPF1 with different 3' UTR lengths (x-axis). Error bars represent standard error of population. P values indicate significance of changes in expression between groups of genes with differing 3' UTR lengths. Middle: As for left, except for consistent genes bound by UPF1 in their 3' UTR. Right: Difference in median LFC of bound genes and unbound genes for different 3' UTR lengths. P values indicate significance of difference of expression changes between unbound and bound genes of indicated 3'UTR length by Wilcoxon rank sum test. In right column, range of differences in median fold change in expression across different length 3' UTR bins. (D) UPF1 binding to shorter 3' UTRs is associated with mRNA de-repression. Only genes with annotated 3' UTRs less than 800 nt were considered. Change in expression following UPF1 depletion (Upf1-1) for consistent genes with 3'UTRs bound by UPF1 (Boundcons, blue ), all genes with 3' UTRs bound by UPF1 (Bound, cyan), and genes not bound by UPF1 (gray bars). Size of dot reflects number of genes falling within this bin of expression change. (E) Delta LFC in mRNA abundance between consistent genes with UPF1 binding in 3' UTR and unbound genes. P values determined as in (A) and displayed as described in Fig. 1.

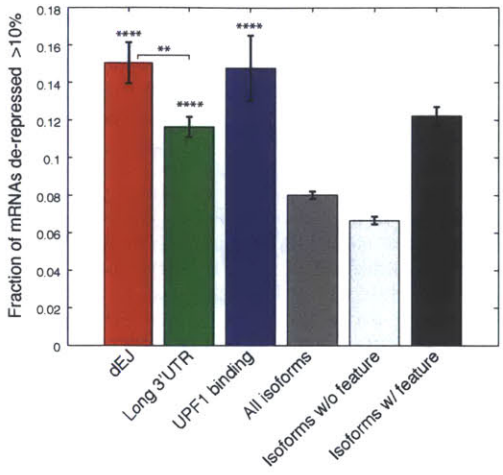



Figure 2-S5: NMD features that regulate isoform abundance, related to Fig. 6.

Figure 2-S5. Predictive capacity of mRNA features to trigger isoform-specific NMD. Calculations were performed as in Fig. 6A except that feature assignment and expression analysis was done on an isoform level. P values above individual bars represent significance of enrichment for de-repressed (>1.1-fold) isoforms within groups of isoforms with specific features as compared to fraction of all isoforms that were de-repressed (All isoforms). Fraction of isoforms not harboring any of the NMD-triggering features described that increased in expression is represented in white bar. P values calculated using hypergeometric test.

## 2.6 Methods

**Cell culture.** Mouse v6.5 (129SvJae x C57BL/6) ESCs were cultured on irradiated DR4 mouse embryonic fibroblasts (MEFs) (Applied Stem Cell) in KO DMEM (Gibco), Pen Strep, L-Glutamine, non-essential amino acids, LIF, and either 10% FBS (Hyclone) (for wildtype) or 15% FBS (for knockdown lines) in gelatinized culture dishes. Puromycin was added to media (1.5ug/ml) during selection as well as during routine culture of knockdown cells. Puromycin was removed from media for 48 hours prior to performing any analysis of knockdown lines. For translational inhibition, 100ug/ml cycloheximide was added to culture media 2 hours prior to harvest.

**Stable knockdown of UPF1.** Mouse ESCs (20,000 cells/well) were plated off of MEFs for 24 hours prior to infection with 40ul of 1.37E+08 titer virus particles (RNAi Consortium shRNA Library). shRNA sequences used were as follows:

Upf1 shRNA-1:
clone TRCN0000009664,
5'-CCGG-GCTGCCATGAACATCCCTATT-CTCGAG-
AATAGGGATGTTCATGGCAGC-TTTTT-3',
Upf1 shRNA-2: clone TRCN0000274486,
5'-CCGG-AGCTATGTGGCTTAGTCTATC-CTCGAG-
GATAGACTAAGCCACATAGCT-TTTTTG-3',

GFP shRNA: clone TRCN0000072181,
5'-CCGG-ACAACAGCCACAACGTCTATA-CTCGAG-
TATAGACGTTGTGGCTGTTGT-TTTTTG-3'.
After 24 hours, media was changed on all infections, and after 48 hours cells were replated with MEFs using media containing puromycin. Clonal populations were isolated and tested for Upf1 KD as well as expression of pluripotency factor Oct4/Pou5f1 by RT-PCR. Clones with minimal Oct4 expression variation from wildtype cells but significant change in Upf1 expression were chosen for further analysis; Upf1-1 KD (4.4, 4.7), Upf1-2 KD (5.2, 5.7), GFP-1,2 KD (2.4, 2.6).

**mRNA expression analysis.** mESCs were trypsinized and pre-plated on gelatinized plates for 30 minutes to remove MEFs prior to harvest in Trizol reagent. Total RNA was further purified following isopropanol precipitation, using RNAeasy columns and on-column DNAse digestion (Qiagen). Twice Poly-T selected RNA was isolated from 10ug of total RNA and used as starting material in paired-end, strand-specific dUTP (Parkhomchuk et al. 2009) library prep using the SPRIworks Fragment library system (Beckman Coulter). Final libraries were amplified using 14 PCR cycles, size selected by agarose gel for 290 base-pair fragments, and sequenced using either 2x80 nt (for knockdown cells) or 2x40 nt (for v6.5 and CHX v6.5 cells) reads on an Illumina HiSeq 2000. To maximize power of detection of lowly expressed isoforms, we combined sequencing data from clones of same hairpins prior to running Cufflinks gene expression analysis software.

**Determination of UPF1 binding.** UPF1 CLIP-Seq was perfomed as described previously (Wang, Kuyumcu-Martinez, Sarma, Mathur, Wehrens & Cooper 2009, Wang, Cody, Jog, Biancolella, Wang, Treacy, Luo, Schroth, Housman, Reddy, Lecuyer & Burge 2012) with the following modifications. mESCs were plated off of MEFs for

24 hours prior to 254 nm UV irradiation (400mJ/cm2) in 15cm plates, trypsinized, washed, and snap frozen. 3x15cm plates of cells were resuspended in 2ml of lysis buffer (50mM Tris 7.4, 100mM NaCl, 1% NP40, 0.5% NaDOC, and protease inhibitors), split into 8x250ul aliquots, and incubated for 30' on ice. After DNAse treatment (5ul TURBO DNAse, Ambion) for 10' at 37 C, dilutions of RNAseIf (NEB) or RNAseA (10mg/ml, Fermentas) were added to each sample to yield final RNAse concentrations of 1:1K-1:10K (for RNAseIf) or 1:10K-1:90K (for RNAseA). Samples were incubated at 37 C with shaking for 10' after which digestions were quenched with the addition of 2ul RNAseOUT (Invitrogen). Samples were spun twice at 4 C, 14K RPM, 10' and supernatant was recovered. Lysates were pre-cleared for 30' at 4deg using a 1:1 protein A:protein G DYNAbead (Invitrogen) slurry. Cleared lysate was recovered and 8ul (1.6ug) of rabbit anti-RENT1 (Cat. No. A301-902A Bethyl Laboratories Inc.) or 2ul (2ug) of rabbit IgG (Upstate Antibodies) was added to respective lysates and incubated for 2 hours at 4 C. Equivalent of 50ul beads was then added to each lysate and incubated for an additional 2 hours at 4 C. IPs were washed 3x in wash buffer (50mM Tris 7.4, 1M NaCl, 1% NP40, 0.1% SDS, 0.5% NaDOC, and protease inhibitors), followed by 2x in PNK buffer (50mM Tris7.4, 10mM MgCl2, 0.5% NP40). Like IPs were combined (2/sample) and CIP treated. Samples were then washed 3x in PNK+ buffer (50mM TrisHCl 7.4, 20mM EGTA, 0.5% NP40), followed by 3x in PNK buffer. 3'linker ligation was performed at 16 C overnight using either: /5Phos/TCGTATGCCGTCTTCTGCTTGT/ddC/ (libraries Upf1.A1, Upf1.I, IgG.A1, IgG.I) or /5Phos/TGGAATTCTCGGGTGCCAAGG/3ddC/ (for use with multi-plexing: libraries Upf1.A2, Upf1.CHX.A1, Upf1.CHX.A2, IgG.A2, IgG.CHX.A1, IgG.CHX.A2).

Both 3' linkers were pre-adenylated using in-house synthesized ImpA (Lau et al. 2001) to improve efficiency of ligations. Ligations were then washed 3X in PNK buffer, and radiolabeld using P32 -ATP. Samples were then washed a final 3X and

resuspended in PNK+ buffer plus LDS sample buffer supplemented with 0.1M DTT and were split among 4 lanes of 3-8% Tris-Acetate SDS-PAGE gels (Invitrogen). Input and supernatant samples were run simultaneously for comparison. An aliquot of each sample was run on a separate gel used strictly for western blotting analysis to verify immuoprecipitaiton of UPF1 protein from input extracts and location of migrating UPF1 protein. Proteins and RNA were transferred to nitrocellulose membrane and exposed to film overnight to visualize RNA species. Appropriately migrating UPF1-RNA complexes were identified and excised from nitrocellulose of both UPF1 CLIP and an equivalent band was excised from IgG CLIP lanes. Proteins were digested by proteinase K treatment, denatured with urea and phenol chloroform, precipitated, and resuspended in a small volume. 5' adapters were ligated at 16 C for 3-4 hours using /5AmMC6/GTT CAG AGT TCT ArCrA rGrUrC rCrGrA rCrGrA rUrCrN rN, where N represents any nucleotide and allows for assessment of PCR amplification bias in samples. Approximately 30% of each sample was not ligated to 5' linkers to enable comparison of the size of RNA species pre- post-ligation comparison by gel analysis. Precipiated RNA was run on 10% TBE-Urea gels and exposed to phosphoimager cassettes for 3hrs to inspect sizes. Gel slices corresponding to varying sizes were excised ( 80nt 150nt). Gel was crushed and RNA was eluted overnight in elution buffer (300mM NaCl and 1mM EDTA), precipitated, resuspended, and reverse transcribed using: 1. 5'-CAAGCAGAAGACGGCATACGA-3' (for non-multiplexed samples), or 2. 5'-GCCTTGGCACCCGAG AATTCCA-3' (for multiplexing). Libraries were then amplified using 25 or 30 cycles of PCR using:

1. 5'-CAAGCAGAAGACGGCATAGCA-3' and
5'-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3' (for non-multiplexed samples), or

2. 5'-AATGATACGGCGACCACCGAGAT-
CTACACGTTCAGAGTTCTACAGTCCGA-3' and, either
5' CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGG-
AGTTCCTTGGCACCCGAGAATTCCA-3',
5'-CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGG-
AGTTCCTTGGCACCCGAGAATTCCA-3',
5'-CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGG-
AGTTCCTTGGCACCCGAGAATTCCA-3', or
5'-CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGG-
AGTTCCTTGGCACCCGAGAATTCCA-3'.

Respective CLIP samples were size selected on 10% TBE gels (BioRad) and eluted, precipitated, and resuspended. Quality of UPF1 CLIP samples was validated by BioAnalyzer and libraries were sequenced on GAIIX (1x36 nt) or HiSeq2000 (1x40 nt) machines.

**Ribosome footprinting.** Footprinting was performed essentially as described by (Ingolia et al. 2009) except for the following modifications. Wildtype v6.5 cells and stable clones 2.6 (GFP shRNA) and 4.7 (Upf1-1 shRNA) were cultured and MEFs were removed as done for mRNAseq analysis. After removing MEFs, cells were washed quickly in PBS and snap frozen in liquid N2. Frozen cell pellets were directly lysed in lysis buffer (20mM HEPES pH7, 100mM KCl, 5mM MgCl2, 0.5% Na deoxycholate, 0.5% NP40, 1mM DTT, and protease inhibitors) on ice for 10'. Lysates were then treated with TURBO DNAse (10ul, Ambion) and RNAseI (2ul, NEB) treated at room temperature for 5'. Lysates were then spun for 10', 17K RPM, at 4 C and the soluble fraction further RNAse digested ( 0.625 ul RNAse/OD260

unit) at room temperature for 55'. Ribosome associated RNA fragments were then isolated by ultra-centrifugation (Beckman Coulter, Ti70) through a sucrose cushion for 1hr and 50' at 60K RPM. Pelleted species were resuspended in 8M Guanidinium HCl and RNA was isolated by acid phenol extraction. RNAs were subsequently processed as described by (Wang, Cody, Jog, Biancolella, Wang, Treacy, Luo, Schroth, Housman, Reddy, Lecuyer & Burge 2012).

**Computational Analysis** Computational analyses were performed using custom scripts in Python, Perl, MATLAB, or Matplotlib.

**Gene expression analysis.** For expression quantification, a custom gene annotation database was used consisting of combined 2011 Ensembl (Flicek et al. 2011) and UCSC FlicekFujita2011 annotations, with duplicate transcripts removed, and a handful well documented PTC+ isoforms that were not in these database releases (Bradley et al. 2012). The RNA-Seq and ribosome footprinting reads were mapped to the mouse genome (mm9) and a list of junctions enumerated from our annotation database using Tophat v.1.4.0 (Trapnell et al. 2009) allowing 2 mismatches but disallowing splice site mismatches and novel introns (–solexa1.3-quals –splice-mismatches 0 –min-intron-length 10 –max-intron-length 1000000 –min-isoform-fraction 0.0 –no-novel-juncs). Expression levels (FPKMs, fragments per kilobase per million mapped) for each library were calculated using Cufflinks v.1.3.0 (Trapnell et al. 2010), enabling compatible hits normalization and each pair of samples being compared were normalized against each other using a Loess-based normalization. mRNA-Seq FPKMs were calculated using annotations strictly from predicted protein coding genes. Ribosome footprint FPKMs were also calculated using Cufflinks with a custom annotation of only the protein coding sequence of these genes. For TE analysis, mRNA-Seq FPKMs were recalculated using the same annotations as for the footprints and then

Loess normalized paired with footprint FPKMs from the same cell type. In general, all analyses considered only isoforms (or in the case of genes, genes with a primary isoform) with 5' UTR >25 nt, CDS >200 nt, and 3' UTR >50 nt in order to avoid spurious annotations. Unless otherwise noted, a minimum expression threshold of 1 FPKM was required for all isoforms and genes (sum of expression levels of corresponding annotated isoforms). Since standard RNA-Seq data reveals relative rather than absolute changes, we generally compared gene expression changes in one group of isoforms or genes of interest relative to another (e.g., isoforms with dEJs or long 3' UTRs to those without).

**Overlap analyses and consistency filter.** For overlap analyses, genes or isoforms changing in expression above or below a given threshold in all three NMD-inhibiting experiments (CHX, Upf1-1, and Upf1-2) were identified (Fig. 1G and S1E). For all other expression analyses, we developed a consistency criterion. For an isoform or gene to pass this criteria, it must have either: 1) increased consistently (>1.1-fold increase in two or more experiments, and not decreased more than 1.1-fold in the third), 2) decreased consistently (>1.1-fold decrease in two or more experiments, and not increased more than 1.1-fold in the third), or 2) remained consistently unchanged (not changed more than 1.1-fold in either direction in all experiments) (Table S2; available online). Isoform and gene groups filtered to pass this filter are designated cons in all figures.

**Analysis of 3' UTR length-related NMD in alternative control cell lines.** The 3' UTR length-associated difference in expression between the two control clones (GFP 2.4/GFP 2.6) was smaller in magnitude (10%) than that seen in in any of the experimental treatments (CHX, Upf1-1, and Upf1-2), but reached statistical significance. Thus we confirmed the 3' UTR length results by comparing the magnitudes of

changes resultant in these cases (given by median LFC long (1500-10k nt ) isoforms short (50-350 nt) isoforms) to 2 additional control cell lines generated using unrelated hairpins (stable knockdown of RBMXL2 in v6.5 cells, generated in parallel with other clones used in this study). We did not find a significant expression change between genes with short 3' UTRs and genes with long 3' UTRs upon depletion of this factor nor when comparing either of these alternative control clones to GFP clone 2.4.

**Analysis of UPF1 binding.** Unique UPF1 CLIP sequences were first trimmed for adapter sequence and 5' randomized barcode (2 nucleotides) to yield fragments 22-38 nt in length and then mapped uniquely to the mouse genome and splice junction database allowing 2 nucleotide mismatches (-m 1 –best strata) using Bowtie (v.0.12.6) (Langmead et al. 2009). Sequencing reads from the IgG libraries were similarly processed, afterwhich read counts at a given position were amplified by an order of magnitude and subtracted from the respective UPF1 library, by iteratively canceling out any reads that overlapped. Regional distribution (exons, introns, intergenic/other) of unique UPF1 CLIP (IgG subtracted) sequences and unique mRNA-Seq reads were calculated by determining if a read mapped within any known coding region, any UTR, and finally within any intron. Remaining reads were assigned to the intergenic/other category. Lengths of all these regions were calculated based on the best isoform for each coding gene. For region specific calling of UPF1 bound mRNAs, UPF1 binding within a given region was compared to the number of reads that would be expected to fall within this region by chance given its expression level, mappability, and length as estimated by mRNA-Seq read coverage. To call mRNAs as bound via either 3' UTR regions or CDS regions UPF1 CLIP reads and mRNA-Seq reads were mapped to these regions. Prior to mapping, UPF1 CLIP reads (IgG subtracted) were collapsed based on position and barcode. Barcodes which were likely ($p >= 0.05$) misreads of other highly amplified barcoded reads at the same position were

predicted based on a Poisson model and removed. mRNA-Seq reads were filtered for those sequenced in the antisense direction of the transcript (in order to maintain even coverage throughout the ends of mRNAs) and for unique mapping within the genome (as was required for CLIP reads). For 3' UTR regions, a custom list of 3' UTR exons that lacked any sequence annotated as coding in any isoform was used. CLIP reads and mRNA-Seq reads (CHX+/-) were tallied per isoform that mapped within respective regions. A minimum expression threshold of 10 mRNA-Seq reads was required in a given region for an isoform to be considered for binding. A P value for binding to a given region of a gene or isoform was calculated for each CLIP library using a Poisson distribution according to the following:

Pval = 1 - Poisson(Cn-1,lambda)

Lambda = Ct*Nx/Nt Cn = number unique clip positions in region (CDS or 3' UTR) of gene/isoform x Ct = total number unique clip positions in region (CDS or 3' UTR) in all genes Nx = number unique mRNA-Seq reads in region (CDS or 3' UTR) of gene/isoform x Nt = total number of mRNA-Seq reads in region (CDS or 3' UTR) in all genes

For CHX- libraries, genes were called as bound if at least one of the three libraries had a P value $<0.001$, a second library had a P value $<0.05$, and library Upf1.A2 (library of lowest coverage) was at least represented with 1 read. For CHX libraries, genes and isoforms were called as bound if both libraries had P values $<0.05$. While CLIP binding was calculated on an isoform basis, binding was generally attributed to the entire gene (by utilizing binding P-values calculated for the bet isoform) for downstream analyses (all tables and figures except for Fig. S5A) since UPF1 binding information is difficult to attribute to specific isoforms. For UPF1 binding specificity analysis, reverse mRNA-Seq reads from CHX- cells (40nt) were mapped to 3' UTR exons and read coverage equivalent to that which mapped to 3' UTR exons for UPF1

CLIP-Seq library Upf1.A1 was randomly sampled from the mapping reads. For simplicity reads spanning exon-exon junctions were omitted (in both mRNA-Seq and CLIP-Seq samples) for this analysis. Genes were called as bound using a P value threshold of 0.001 as described above.

**UPF1 binding correlation.** Correlation coefficients between binding of UPF1, MBNL1 (Wang, Cody, Jog, Biancolella, Wang, Treacy, Luo, Schroth, Housman, Reddy, Lecuyer & Burge 2012), and AGO2 (Leung et al. 2011) were calculated as the correlation of CLIP densities in each 3' UTR (Fig. 3C) or in 100 nt windows within each 3' UTR (Fig. 3D) for highly expressed genes (FPKM  50) similar to in >Wang2012.

Analysis of expression change following NMD inhibition of UPF1 bound genes. To assess for the degree that 3' UTR length explains the de-repression of UPF1 bound genes following NMD inhibition, we subsampled, with replacement, groups of genes that were of similar 3' UTR lengths and initial expression levels as those consistent genes that were bound by UPF1 (inclusive of the bound set) and measured the degree to which they increased expression in each NMD inhibition experiment. The average cumulative distribution of expression changes for the subsampled populations is shown in Fig. 4F for Upf1-1, but were similar for other experiments. To estimate the degree to which UPF1 binding increases de-repression over what would be expected for length alone, we calculated the difference between the median LFC of the bound genes and the median of the median LFCs of the subsampled populations (results ranged between 1.1- to 1.16-fold for the three experiments). The degree to which UPF1 3' UTR binding increases de-repression was also estimated in Fig. S4, wherein expression changes were assessed for genes that were bound and not bound by UPF1 with varying 3' UTR lengths.

**Comparison with previously published data.** Upf2 KO and control mouse liver and BMM data sets were downloaded from Gene Expression Omnibus (GSE26561) (Weischenfeldt et al. 2012) and Smg1 KO and control MEF data was obtained from the lab of Benjamin Blencowe (University of Toronto) (McIlwain et al. 2010). All RNA-Seq data was processed as described for data generated in this study. Homologs of UPF1 bound and unbound genes were determined using the BioMart tool (Vilella et al. 2009). To assess gene expression changes for homologs of bound genes, microarray data was downloaded from Gene Expression Omnibus and the following parameters were used: for GSE30499, a minimum expression threshold of 64 was required of the RMA processed data for inclusion in analyses (Wang, Wengrod & Gardner 2011), for GSE26781, a minimum expression threshold of 8 of the log transformed, quantile normalized data was required for inclusion in analyses (Cho et al. 2012). For cases in which a gene was represented by multiple probesets, the mean value of all the corresponding probesets was used.

**Genome-wide survey of NMD.** Gene annotations and expression analysis in v6.5 mESCs were used to calculate the number of expressed genes (FPKM>1) harboring different NMD-inducing features. dEJ genes were defined as those that harbor at least one expressed annotated dEJ isoform (isoform level must be >1FPKM and account for at least 10% of the expression expression level of the entire gene). Long 3' UTR genes were defined as those whose primary annotation harbors a 3' UTR >1500nt. uORF genes were defined as those with at least one annotated uORF. tuORF genes and CLIP genes were defined as described in tuORF methods and UPF1 3' UTR binding methods. Predictive capacity of NMD features was determined by calculating the fraction of genes harboring a given feature that were up-regulated by a certain threshold in at least 2 of 3 NMD inhibition experiments (without significant down-regulation in the third) compared to all expressed genes in the transcriptome

that harbored this feature (significance calculated by two-tailed Fisher's exact test). In order to avoid double-counting of genes, features were called for genes using the following hierarchy: 1) UPF1 3' UTR binding, 2) Presence of dEJ, 3) Presence of a long 3' UTR; and 4) Presence of a tuORF.

**uORF classification.** For classification of uORF translation status, ribosome footprinting reads were reduced to the codon occupied by the A site of the ribosome calibrated based on the pile up at known stop codons, similar to (Ingolia et al. 2011). For each uORF, densities of mapped A sites were calculated within the uORF and the background untranslated region consisting of the 10 codons upstream and downstream of the uORF. uORFs were required to be covered by at least one mRNA-Seq read to ensure they were spliced into the message. uORFs were called as translated when there was at least 5 fold greater density above the higher of either: the footprint density in the flanking 60 nt or a minimum threshold coverage of 2/3. If a uORF was not called as translated, it could be called as confidently untranslated (an ntuORF) if the footprint density within the uORF was less than the higher of either: the footprint density in the flanking 60 nt or one read per 60 nt. Genes were classified as tuORF containing if they harbored a transcript with one or more tuORFs or as ntuORF containing if they harbored one or more ntuORFs and no tuORFs as called in data from clone 4.7 (shRNA Upf1-1).

## 2.7 Author Contributions

JAH and CBB designed the study. JAH performed the experiments. JAH and ADR analyzed the data. JAH ADR and CBB wrote the manuscript.

# Chapter 3

# Quantitative Analysis of Protein-RNA Binding Reveals Novel Functional Motifs and Impact of RNA Structure

## 3.1 Introduction

RBPs bind sequence and/or structural motifs in nuclear pre-mRNAs to direct their processing, and bind 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, Cody, Jog, Biancolella, Wang, Treacy, Luo, Schroth, Housman, Reddy, Lecuyer & Burge 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. Only about one third to one half of RBFOX2 binding sites identified in vivo contain these canonical motifs (Yeo et al. 2009), 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 bound with 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 time points or in certain cell types where the RBP accumulates to high levels.

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 proteins decrease and MBNL proteins accumulate over time, this antagonism may help to confer sharper developmental splicing transitions (Kalsotra et al. 2008). In the muscle wasting disease myotonic dystrophy

type 1 (DM1), expanded CUG repeats in the 3' UTR of DMPK mRNAs reduce available cellular levels of MBNL proteins by sequestration (Mankodi et al. 2005, Taneja et al. 1995), and CELF1 proteins are stabilized by hyperphosphorylation (Kuyumcu-Martinez et al. 2007). CELF1 has three RNA recognition motifs (RRMs) that bind motifs with consensus UGU (Ladd et al. 2001, Marquis et al. 2006). MBNL1 has two pairs of zinc fingers that are reported to bind preferentially to YGCY (Y = C or U) motifs (Ho et al. 2004). To date, it has remained unclear whether MBNL1 recognizes secondary structural motifs or single-stranded RNA elements. CUG repeat RNA crystallizes as an A-form helix (Mooers et al. 2005), with C and G bases paired and Us unpaired, and biochemical evidence showed that a mismatched RNA hairpin structure is important for recognition by MBNL1 (Warf & Berglund 2007). However, structures of MBNL1 zinc fingers co-crystallized with CGCUGU RNA suggested that MBNL1 recognizes single-stranded RNA (Teplova & Patel 2008). MBNL and CELF proteins each bind strongly to sequences containing repeats of their preferred motifs, but the presence of multiple RNA binding domains in each protein complicates analysis of the effects of the number of motifs, the degeneracy of intervening bases and motif spacing on binding strength.

Widely used methods for mapping protein-RNA interactions 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. 2012). SELEX identifies RNA motifs bound with high affinity in vitro, but is not quantitative and may miss lower affinity motifs. RNAcompete uses in vitro RNA-protein binding followed by microarray analysis, allow-

ing the high-throughput identification of RNA binding motifs citeRay2009,Ray2013. 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 including $K_d$ values can be obtained from methods such as electrophoretic mobility shift assays (EMSA) or surface plasmon resonance (SPR), but their throughput is quite low.

To better characterize the functions of biologically important RBPs, 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 (Zykovich et al. 2009, Jolma et al. 2010), 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 these 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 CLIP (or "iCLIP") data for these factors showed that in vitro RNA binding affinity largely drives binding in vivo. Analysis of data from systems in which these RBPs are depleted or inducibly over-expressed provides evidence of function for both non-canonical and canonical binding motifs

identified in vitro. This analysis also shows that motifs enriched by CLIP but not RBNS are not associated with regulatory activity, showing that RBNS provides information that is complementary to CLIP.

## 3.2 Results

### 3.2.1 Design considerations for RNA Bind-n-Seq experiments

RBNS is designed to dissect the sequence and RNA structural preferences of RBPs. A recombinantly expressed and purified RBP is incubated with a pool of randomized RNAs 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 of length $\lambda = 40$ nt flanked by short primers used to add the adapters needed 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 unusual case where the RBP had significant affinity to primer RNA, 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 measurements 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 on binding such as the RNA secondary structural effects described below that cannot be assessed using 10mers. 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).

### 3.2.2 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 (for k = 5, 6, 7) as the ratio of the frequency of the kmer in the selected pool to the frequency in the input RNA library. In the 0 RBP libraries, 99% of 6mers had R values of less than 1.16 and the maximal R values were 1.2, 1.6 and 1.3 for RBFOX2, CELF1 and MBNL1 respectively.

For RBFOX2, at all concentrations $\geq$14 nM the 6mer UGCAUG had the highest R value (Figure 1B and below), confirming this well-known motif as the 6mer with highest affinity. 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 (Chapter 4, equation 31). With R = 22,

k = 6 and $\lambda$ = 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, indicating that this 5mer represents an alternate RBFOX2 binding motif. Extensive analysis of the in vivo binding and regulatory 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 (Marquis et al. 2006, Timchenko et al. 1996)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 (see chapter 4), 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 affinity 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 is somewhat complex. Inspection of the top 50 CELF1 7mers (Supp. Fig. 1C) suggested that they can be clustered into 4 groups depending on the spacing of GU motifs, with classes matching $GUN_xUGU$ for x =

0,1,2 and a fourth class matching UNUGU, with each class having different degree of preference for U (or occasionally G or A) at the remaining positions (Figure 1E). This representation emphasizes the complexity of CELF1's recognition of RNA, likely deriving from the presence of multiple RRM domains in the protein.

MBNL1 is known to favor binding to YGCY motifs by SELEX in vitro (Goers et al. 2010, Wang, Cody, Jog, Biancolella, Wang, Treacy, Luo, Schroth, Housman, Reddy, Lecuyer & Burge 2012), and GCUU and UGCU were the most enriched 4mers by CLIP-Seq (Wang, Cody, Jog, Biancolella, Wang, Treacy, Luo, Schroth, Housman, Reddy, Lecuyer & Burge 2012). The most enriched 7mers for MBNL1 contained either YGCU or GCUU, often supplemented by a second GC, and the 7mer with highest R value, GCUUGCU, contained both of these 4mers (Figure 1D). This 7mer 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 better summarized as YGCU + GCUU rather than YGCY. Focusing on the core GC dinucleotide, MBNL1 prefers 7mers containing two GCs, with a slight preference for spacing of 2 or 3 Us (Figure 1F), 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 some variation in the extent of preferences for U (or occasionally C) at the remaining positions (Figure 1F).

At the level of 6mers and 7mers, we observed no evidence for cooperative binding to RNA for any of the proteins analyzed here (Methods).

### 3.2.3 Relative dissociation constants are accurately estimated from RBNS

To better understand the dependence of R values on RBP concentration and to assess the extent of experimental noise, we modeled RBNS 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 nM and several moderate affinity motifs Y each with $K_d$ = 30 nM (assuming binding with 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 half maximal binding to the motif occurring at a free protein concentration near the $K_d$ value (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).

The modeled enrichment profiles (Figure 2B) show that under all conditions R values of high affinity motifs decrease at higher RBP concentrations. This effect is readily understood by considering that at high RBP concentrations high affinity motifs will be saturated, resulting in increased binding to lower affinity RNA 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 a wide range of assumptions about affinities (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 to 40) nanomolar range, decreasing to near 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 (Supp. Table 2; available online). 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 and compared them to measurements obtained using SPR. 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 $\mu$M). The concentration of kmer in complex with RBP was then calculated from the total concentration of RBP-RNA complex as measured by fluorescence (Methods). The fraction of bound RNA attributable to binding at each specific kmer was then estimated using a novel "streaming kmer assignment" (SKA) algorithm (Chapter 4). SKA generalizes the analytical approach described in Chapter 4 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 a specific kmer in the sequence probabilistically, based on continually updated estimates of relative binding preferences, using 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 introduced in the recently described eXpress algorithm (Roberts & Pachter 2012). Using simulated read data, we observed that assignments of binding locations within reads are more accurate when using SKA than when using raw R values, or B values inferred using equation 4.18.

SKA is particularly helpful in distinguishing truly 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 for 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, and "learns" to assign lower probabilities (usually near background levels) to overlapping motifs (Supp. Fig. S5, S6). 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 technical factors that may systematically affect absolute $K_d$s and may vary between experiments (e.g., salt concentration, proportion of properly folded protein, etc.). 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 = 0.935, 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.

### 3.2.4 Secondary structure inhibits binding of RBFOX and CELF proteins to RNA

In addition to characterizing sequence preferences, RBNS can assess effects of RNA structure on binding of RBPs. Application of the thermodynamically-based Vienna RNAfold algorithm (Hofacker 2003) to sequence reads enabled assessment of 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). 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 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 recogni

### 3.2.5 MBNL1 binding tolerates pairing of GCs but prefers 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 better understand the impact of 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 3B). A similar tolerance for pairing of the central GC dinucleotide and preference for unstructured flanking pyrimidine bases was observed for all high affinity MBNL1 motifs tested, including UGCUU, GCUUGC, CGCUU and GCUGCU. This apparent tolerance for GC base-pairing does not result simply from MBNL1's preference for binding multiple nearby GpC dinucleotides (which might base-pair with one another), as similar structural preferences remained when GpC content was controlled for (Supp. Info.). 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).

### 3.2.6 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 (Merkin et al. 2012). In that study, we 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. Curiously, we found that MBNL1 binding to these introns (assayed by CLIP-Seq) exceeded that expected based on motif enrichment by several fold, implying that these introns possess additional contextual feature(s) that favor binding of MBNL proteins. To ask whether the RNA structural preferences observed above might contribute to this preferential binding, we performed RNA folding analysis of these ancient alternative exons, and also 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 (which were not as enriched for MBNL1 binding) (Figure 3C). These observations suggest that contexts in which MBNL motifs have unpaired Us represent at least one of the features that have evolved in ancient alternative exons to facilitate binding of MBNL proteins.

### 3.2.7 Motifs identified in vitro are almost invariably bound in vivo

RBNS data resolve a spectrum of high, moderate and low affinity motifs bound in vitro. To assess the extent to which these motifs are bound in vivo, we compared to CLIP-Seq data. For RBFOX, 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 stem cell biology (Jangi et al. in submission). 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 the Dyrk1a gene (Figure 4A). To systematically assess the in vivo binding specificity of RBFOX2, 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). 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 absence of Us from this motif and the somewhat higher background fluctuations present in these plots that results from the lower abundance of GCACG sites in the mouse transcriptome. This lower abundance likely results from 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 or alternate RBFOX2 motifs (Figure 4B, bottom).

Similar analyses of MBNL1 motifs using MBNL1 CLIP-Seq data from our previously published study using C2C12 mouse myoblasts (Wang, Cody, Jog, Biancolella, Wang, Treacy, Luo, Schroth, Housman, Reddy, Lecuyer & Burge 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), yielded a similar 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, which likely reflects the lower resolution of the standard CLIP-Seq protocol used for these factors relative to the iCLIP protocol used for RBFOX2. Again, these peaks were RBP-specific (not

shown).

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 at -80...-41 upstream and +41+80 downstream of the motif, representing the "background" level of CLIP density in motif-containing transcripts. Comparing CLIP S/B values to RBNS R values across all 6mers for RBFOX2, 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 (93%, 94% and 97% of 6mers for RBFOX2, CLEF1 and MBNL1), 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 for CELF1 and MBNL1 (data for 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.

### 3.2.8 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, Cody, Jog, Biancolella, Wang, Treacy, Luo, Schroth, Housman, Reddy, Lecuyer & Burge 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 levels of the inducer to an mESC line containing a tetracycline inducible version of RBFOX2 (Jangi et al. in submission). 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 (shFOX2, 0 $\mu$g/mL DOX) to 32 at the highest induced level (shGFP, 1 $\mu$g/mL Dox), a $\sim$3-fold dynamic range. Western analysis confirmed knockdown of endogenous RBFOX2 in shFOX2 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 increases. To systematically address the consistency of changes in splicing, we defined a "monotonicity Z-score" (MZ) for each exon whose PSI value changed significantly (Wang et al., in prep). 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 & Hynes 1993, Lim & 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 over-expression condition. Interestingly, inclusion of the alternative exon in UAP1 changes substrate specificity of this enzyme (Wang-Gillam et al. 1998). 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 almost all 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 an RNA-Seq time course analysis of splicing changes in mouse muscle and heart following induction of CELF1 expression (Wang et al, in preparation). CELF1 binding to upstream introns has been associated with repression of exon inclusion (Kalsotra et al. 2008, Dasgupta & 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).

### 3.2.9 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. 2012)However, the absence of alternative comprehensive high-resolution methods has made it challenging to critically assess the quality of CLIP data for systematic biases or sources of false positives and false negatives. Previous studies have shown that CLIP favors U-rich sequences, because uridines form RNA-protein crosslinks more readily than other bases (Sugimoto et al. 2012). 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 (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 high CLIP S/B but no significant RBNS enrichment. By contrast, the remainder of 6mers with high CLIP S/B also had significant positive 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 even essentially nonspecific (low specificity) protein-RNA interactions may be detected in contexts that are sufficiently U-rich.

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). 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 RBFOX2 binding to nearby 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.

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 6B). 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 explanation that can't be excluded is that these motifs 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. This analysis shows that RBNS can provide information useful 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.

### 3.2.10 RBNS motifs are conserved across mammals

Motifs that contribute to regulation of conserved alternative splicing events are expected to often be evolutionarily conserved, and the canonical binding motifs of RBFOX2, MBNL1 and CELF1 are highly conserved in introns flanking alternative exons and in 3' UTRs (Daughters et al. 2009, Sugnet et al. 2006, Wang, Cody, Jog, Biancolella, Wang, Treacy, Luo, Schroth, Housman, Reddy, Lecuyer & Burge 2012, Merkin et al. 2012, Wang, Sandberg, Luo, Khrebtukova, Zhang, Mayr, Kingsmore, Schroth & Burge 2008). 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 be more reliably aligned in most cases. For this analysis, we calculated for each 6mer the fraction of its occurrences in conserved introns 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 their function. No significant conservation was detected for the set of CLIP+/RBNS 6mers (Figure 6E), again suggesting that this set lacks regulatory activity. By contrast, the set of CLIP+/RBNS+ motifs matched for CLIP density showed significant conservation (Figure 6E).

## 3.3 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.

### 3.3.1 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 could form 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 some-

what 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.

### 3.3.2 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 tightly bound by MBNL proteins both in vitro and in vivo (Fardaei et al., 2001; Kino et al., 2004; Teplova and Patel, 2008a), crystallizes as a hairpin with paired GCs separated by unpaired U-U bulges (Mooers et al. 2005), 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 & 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. 2010). 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 (Oberstrass et al. 2005, Perez et al. 1997) and is also consistent with the crystal structure of MBNL1 zinc fingers 3 and 4 (Teplova & Patel 2008).

### 3.3.3 RBNS enhances interpretation of CLIP data

RBNS appears to yield a less biased portrait of the spectrum of RNA motifs bound by an RBP than do methods based on UV crosslinking, making it a useful complement to CLIP-based methods (including iCLIP and PAR-CLIP). For example, we observed a subset of CLIP-enriched motifs that were not detected by RBNS ("CLIP+/RBNS motifs") and found these motifs also lacked evidence of regulatory activity or sequence conservation, arguing that they do not reflect biologically relevant binding. In practice, when crosslinking to a CLIP+/RBNS motif that is located in close proximity to a CLIP+/RBNS+ motif is observed, our analyses imply that in most cases this binding should be attributed to the CLIP+/RBNS+ motif. Applying this sort of correction automatically could be explored as a means to improve the resolution of in vivo binding site mapping with CLIP data. When comparing the extent of binding to two or more different regions, we expect that RBNS affinities can be used to correct for the crosslinking bias inherent in CLIP and improve the accuracy of quantitation. In addition, 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 help to distinguish non-specific binding from cooperative binding through another RBP.

## 3.4 Figures

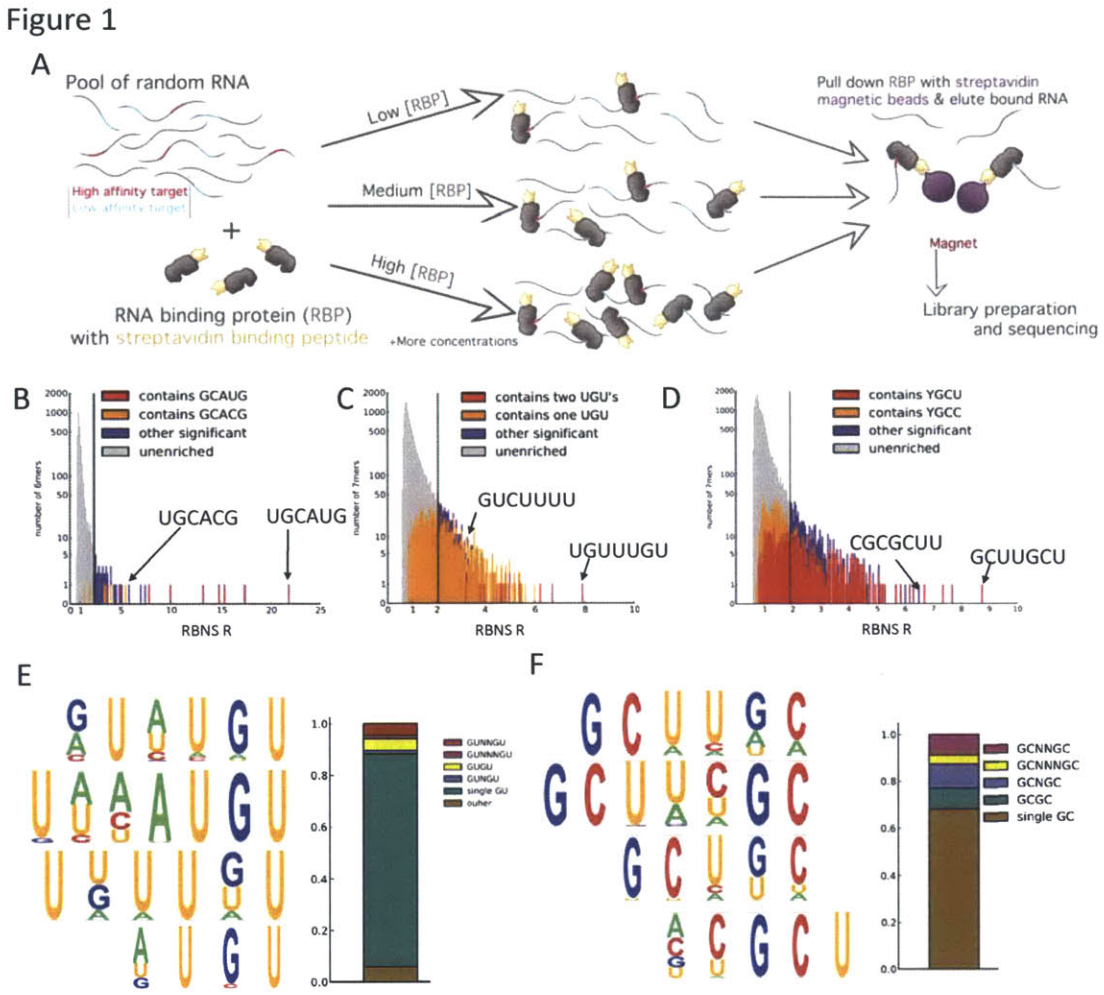



Figure 3-1: RNA Bind-n-Seq overview and motif enrichment analysis.

Figure 3-1. (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 $1\mu$ M. 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) Visualization of the CELF1 binding preferences. The top 50 motifs were grouped by submotif (GU) spacing and pictographs were generated from the grouped sequences. (F) F. Visualization of the Mbnl1 binding preferences. The top 50 motifs were grouped by submotif (GC) spacing and pictographs were generated from the grouped sequences.

Figure 2

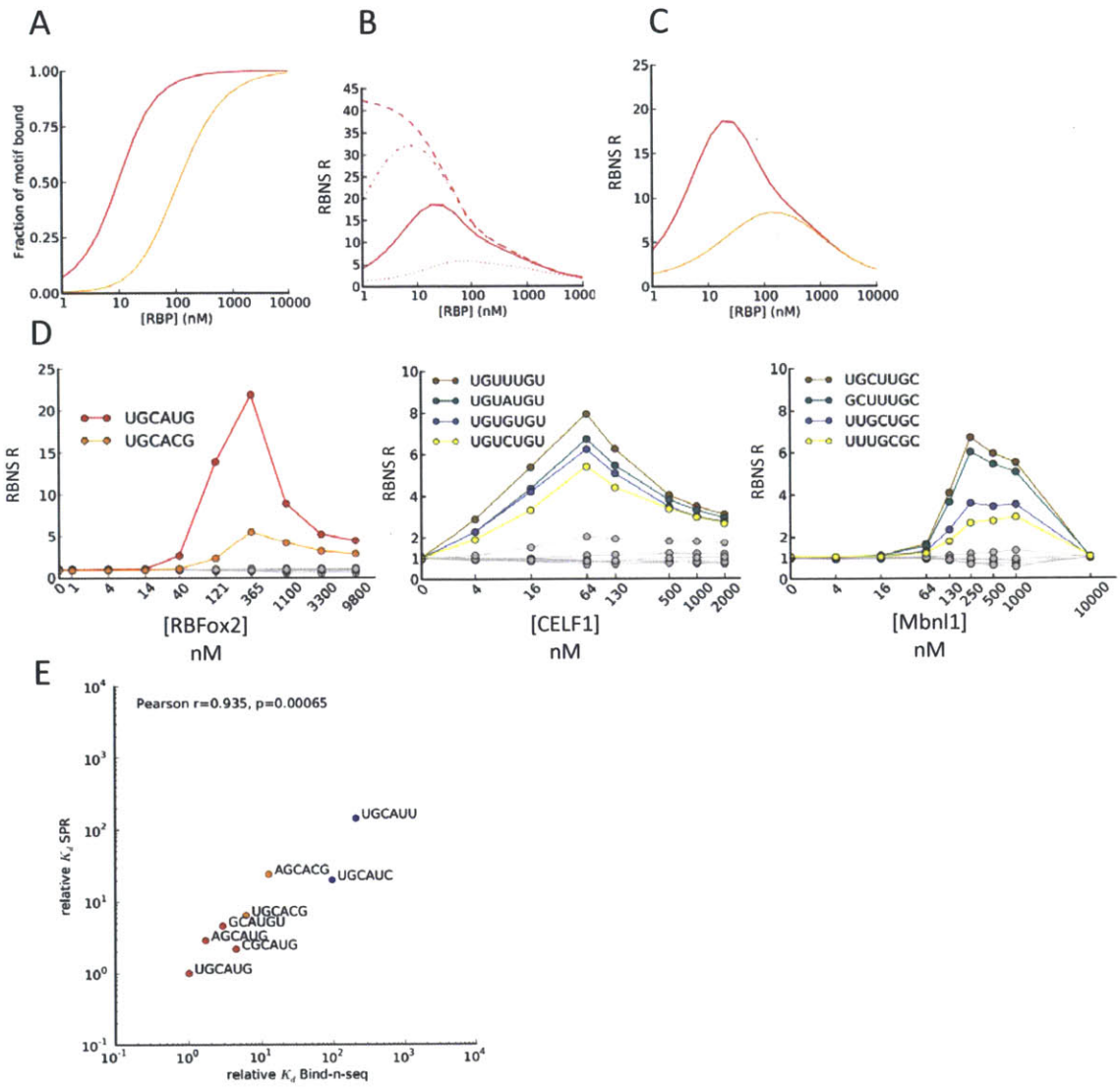



Figure 3-2: Estimation of dissociation constants using RBNS data.

Figure 3-2. (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) 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 UGUNUGU. 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 colors as in Figure 1B.

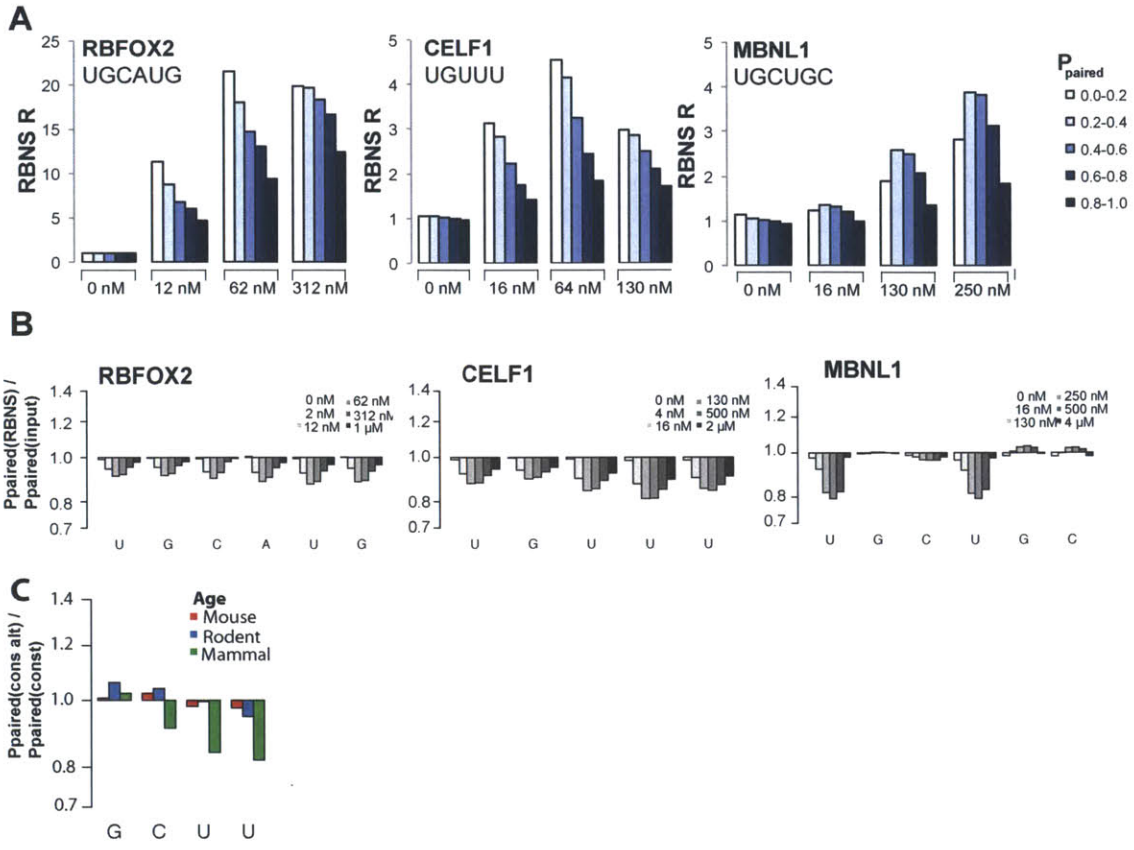



Figure 3-3: mRNA structure

Figure 3-3. (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) C. This bar plot shows the ratios of base-pairing probabilities as in (B) for motifs located in alternative exons based. Alternative exons are binned by how deeply conserved the alternative exon is as determined by (Merkin et al. 2012).

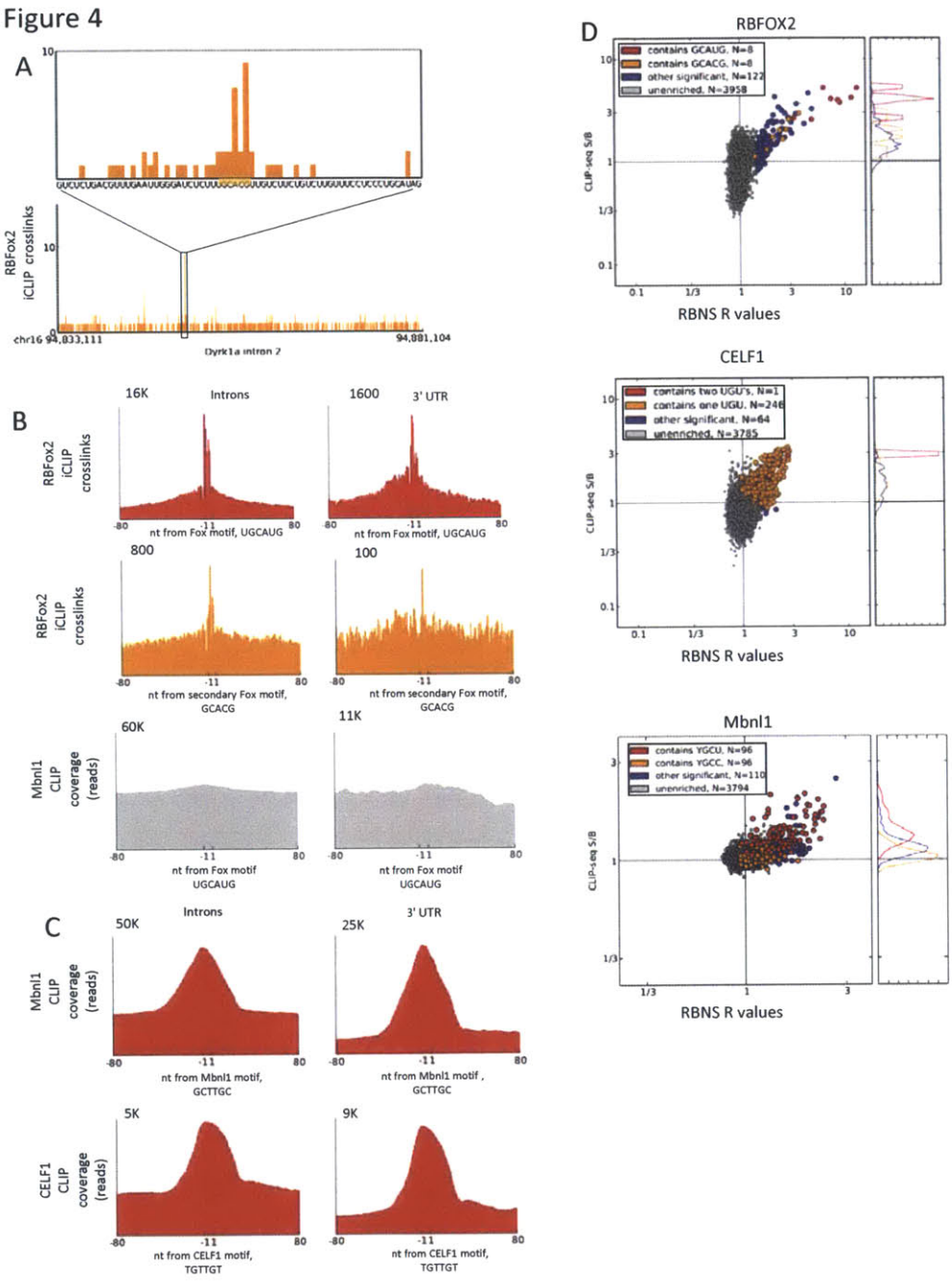
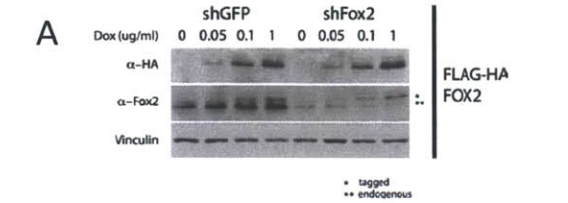
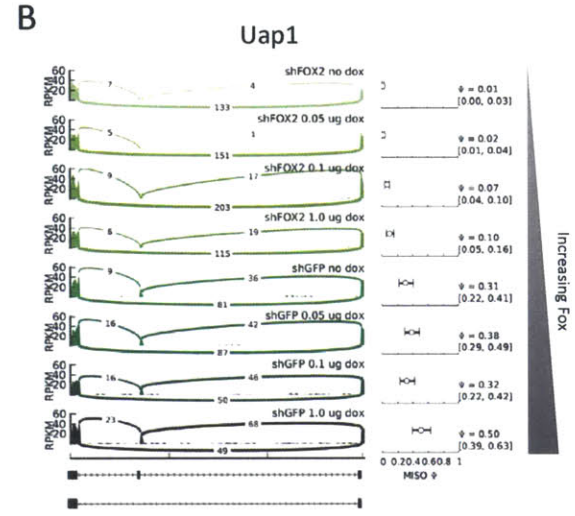


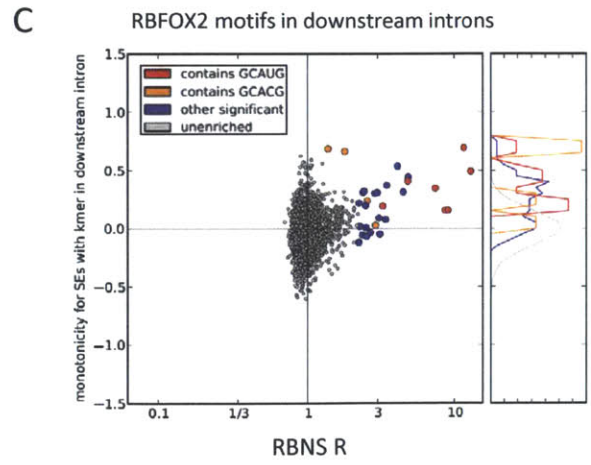

Figure 3-4: Correlation of RBNS with in vivo binding.

Figure 3-4. (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 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. Top 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. Bottom 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.

Figure 5



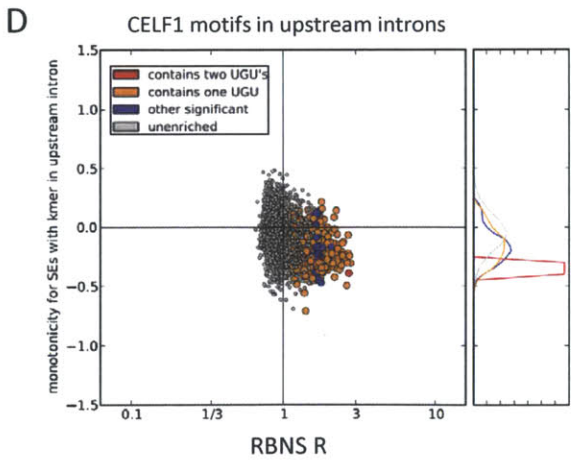




Figure 3-5: In vivo function.

Figure 3-5. (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 $\mu$ g/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).

Figure 6

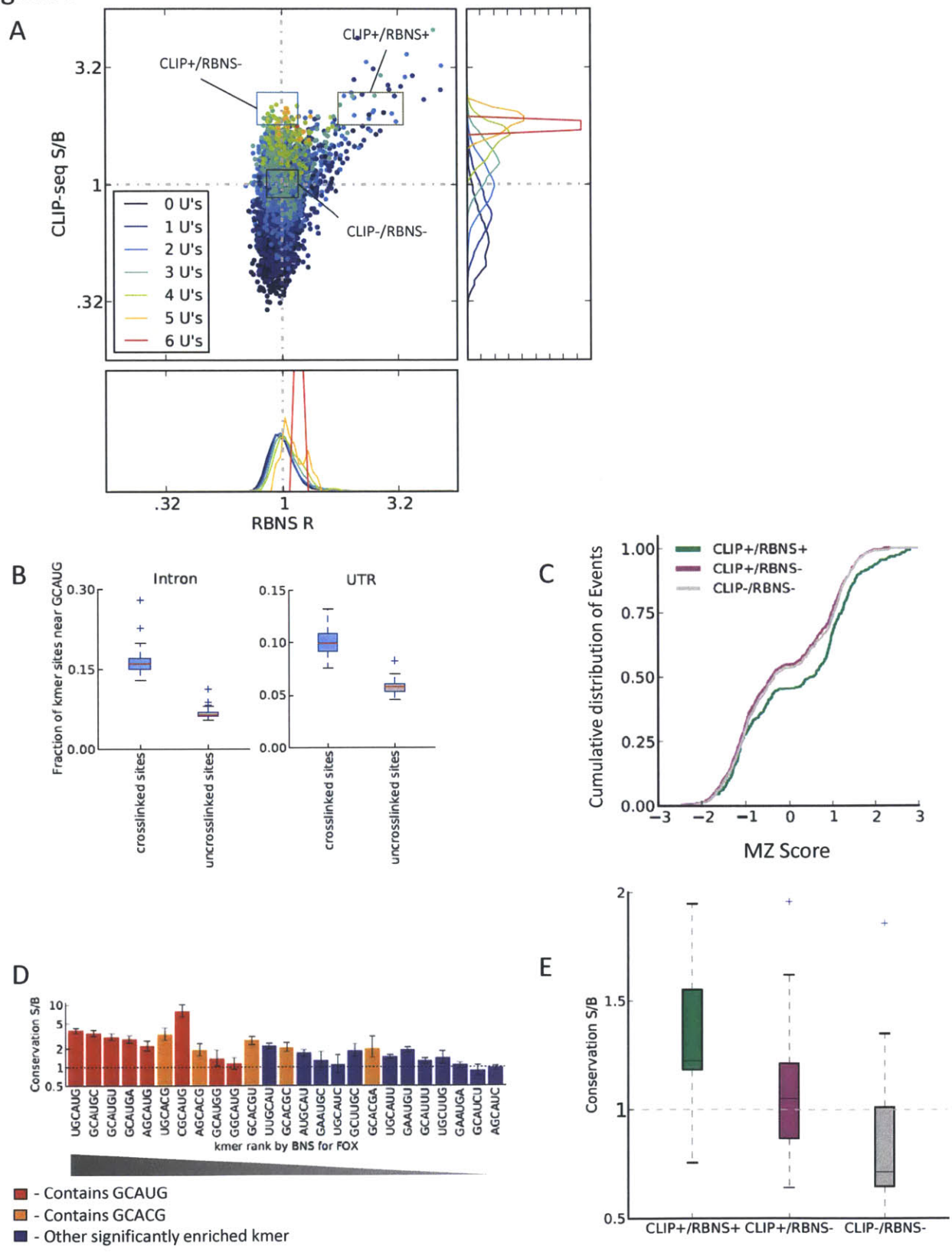



Figure 3-6: Conservation of identified motifs

Figure 3-6. (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 Us 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 (see online methods) for the CLIP/RBNS kmer sets as defined in Figure 6A.

## 3.5 Supplemental Figures

Figure S1

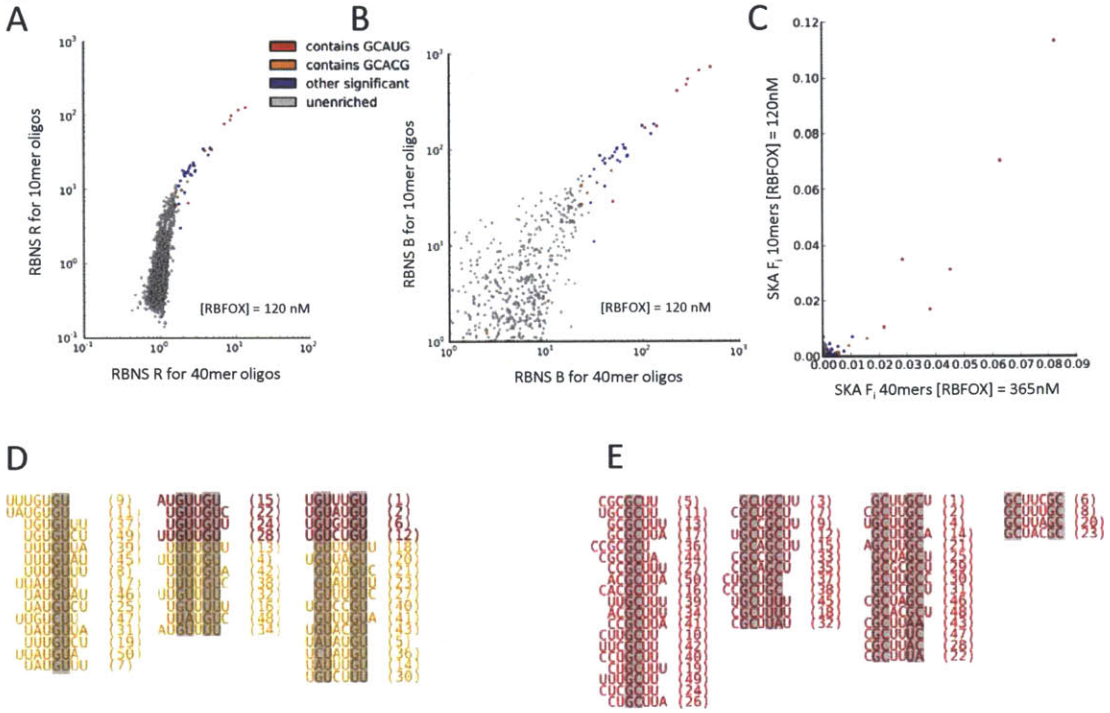



Figure 3-S1

Figure 3-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. (D) Aligned enriched CELF1 motifs clustered by separation of GUs. Red sequences indicate two UGUs are present; orange indicates one is present. Used to generate pictographs in 1E. Shading indicates submotifs used for alignment. (E) Aligned enriched MBNL1 motifs clustered by separation of GCs. Orange indicates GCUU is present; gray indicates no GCUU is present. Used to generate pictographs in 1F. Shading indicates submotifs used for alignment.

Figure S2

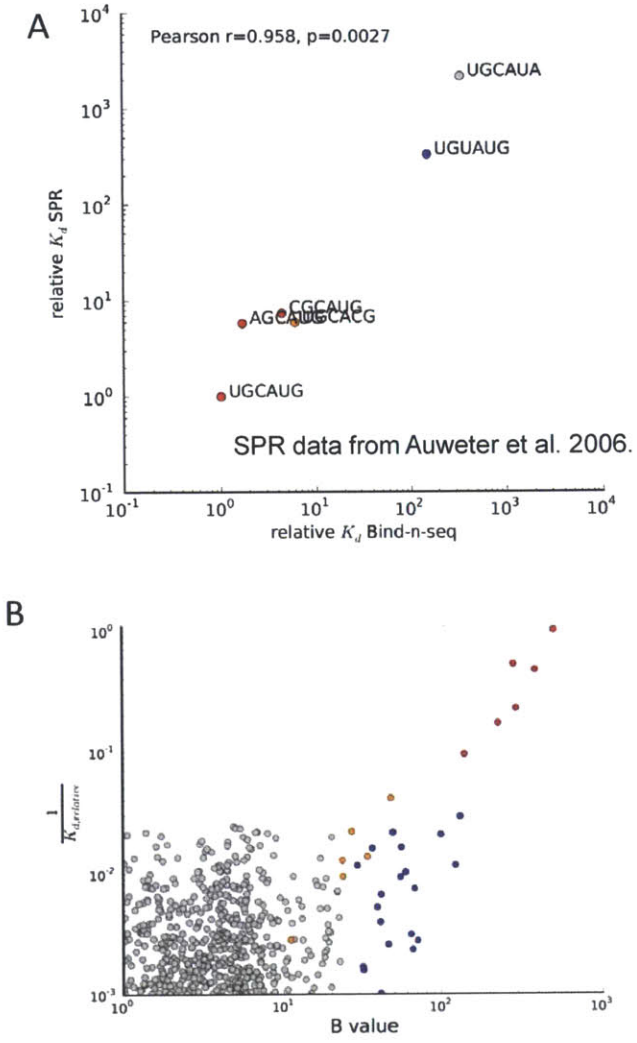



Figure 3-S2

Figure 3-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). (B) Correlation of R values and $1/K_{d,relative}$ values for RBFOX2 6mers.

Figure S3

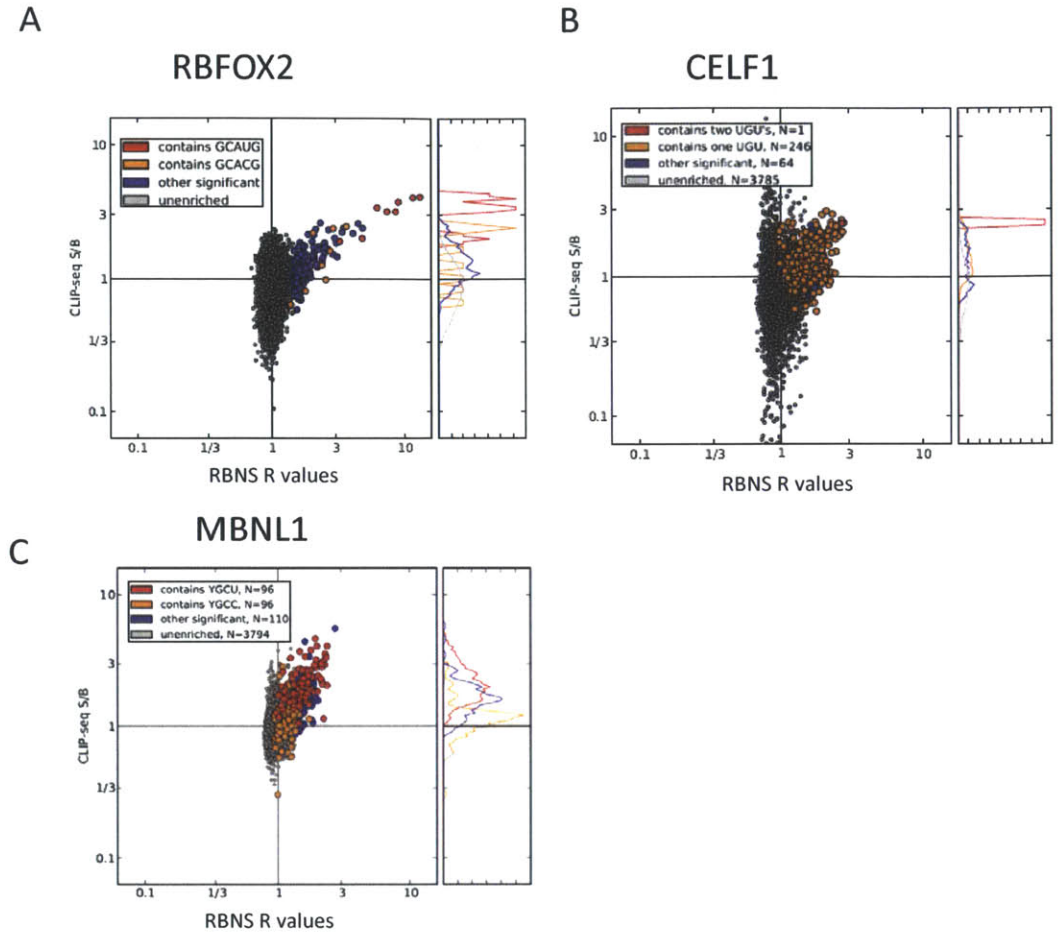



Figure 3-S3

Figure 3-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.

## Figure S4

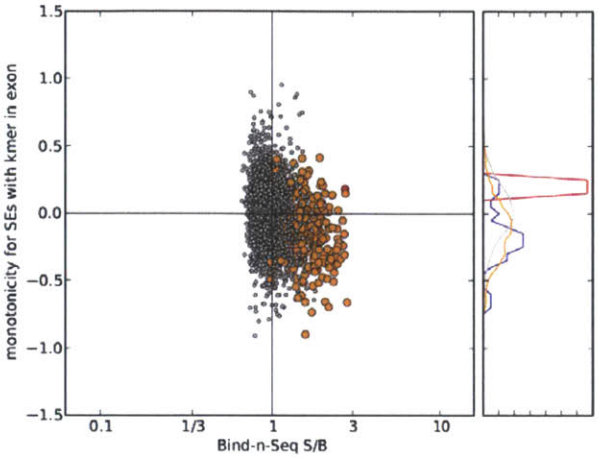



Figure 3-S4

Figure 3-S4. (A) Average CELF1 MZ scores for SEs which contain each of the 6mers within the exon.

Figure S5

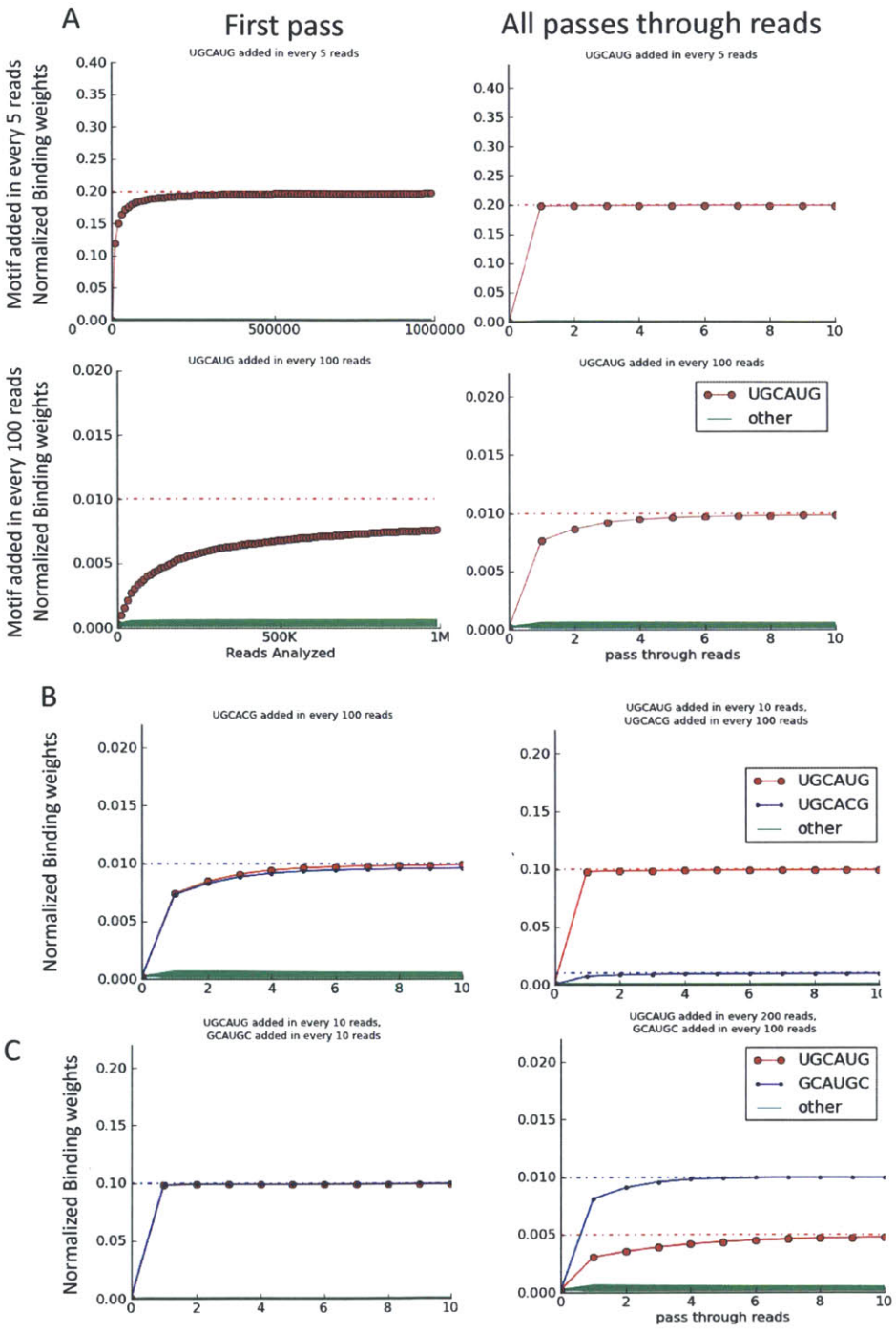



Figure 3-S5

Figure 3-S5. (A) Convergence of the SKA algorithm (k=6) over increasing amounts of data for simulated RBNS data with either 5% or 1% additional motif "spiked in". Left plots show the first pass through a 1 million read simulated data set. Right plots show subsequent passes through these data sets. (B) As in A right but for simulated data with secondary motifs added as well at 1% or 1%. (C) As in A but for simulated data with secondary motifs which overlap the primary motif added as well at 0.5% or 10%.

Figure S6

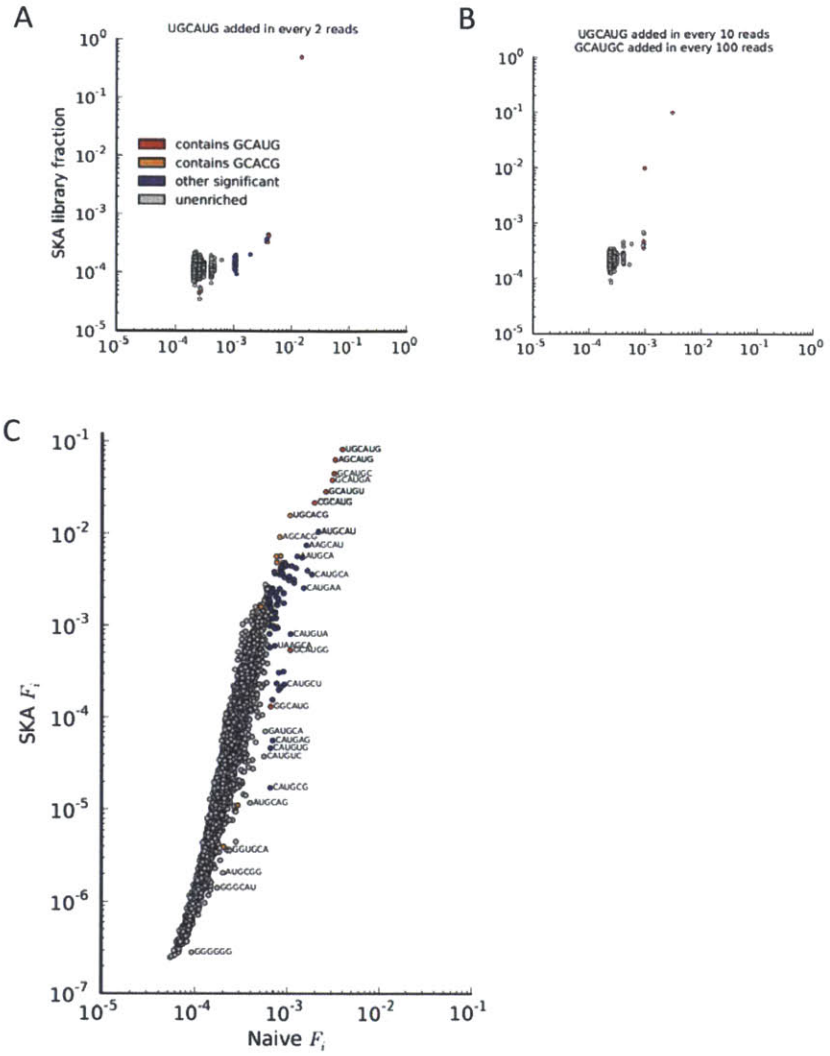



Figure 3-S6

Figure 3-S6. (A) A comparison of analysis of simulated data by either SKA (y axis) or the nave calculation of R values (x axis). Primary motif is spiked in at 50%. (B) As in S6A but with an overlapping secondary overlapping motif as well at 1% (primary motif at 10%). (C) Comparison of SKA and nave estimates of Fi (library fraction) for 6mers in experimental data from RBFOX2 RBNS ([RBFOX] = 365nM).

## 3.6 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** 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 Superasin (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.

**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 by Wang et al. (in preparation). 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 Fox concentration showed significantly more inclusion and the number where the lower Fox 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 RBNS-/CLIP+** 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.

**Simulated Bind-n-seq 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, ranging the protein concentration $[R]_T$ 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] would either be completely absent, 1 nM ("low"), 10 nM ("medium"), or 100 nM ("high"). The concentration of specific motifs $[L1]_T$ and $[L2]_T$ 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, with the remaining ligand pool with neither motif denoted as $[L]_T$. Setting the dissociation constant for the "strong" motif $K_d$-L1 = 5 nM, and the moderate motif $K_d$-L2 = 30 nM, we constructed the linear system of equations relating the concentrations of the free and bound states of R, B, L1, L2, and L 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 (specifically and non-specifically) 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 (specifically and non-specifically) of a particular motif is first divided by the total bound concentration of all ligands multiplied by 35, the number of motifs of length 6 within a ligand of length 40. This value is then divided by $\frac{1}{4^6}$, the background expectation of any particular motif of length 6 in a library without enrichment for any specific motif.

## 3.7 Author Contributions

NJL and CBB designed the RNA Bind-n-Seq experiments. NJL performed the RNA Bind-n-Seq experiments. NJL and ADR analyzed the data. ADR SEM and CBB developed the theory. MJ and PAS provided the CLIP data. ADR, NJL and CBB wrote the manuscript.

# Chapter 4

# Framework for Understanding Deep Binding Affinity Data

## 4.1 Introduction

This document describes the methods used to analyze RNA Bind-n-Seq (RBNS) experiments and the necessary theoretical background. Section 1 reintroduces the method and briefly summarizes the challenges. In section 2 we describe the experimental method in more detail and explain the adjustable experimental parameters. In section 3 we define the biochemical model and its variables. Section 4 presents a simplified model which gives insight and intuition for understanding the model. Section 5 describes a novel motif quantitation algorithm, its validation and the reasons for its necessity. Section 6 derives the equations for calculating relative $K_d$s from RBNS data.

RNA Bind-n-Seq generates deep sequence data for estimating the *in vitro* binding affinities of an RNA binding protein (RBP) for RNA sequences by selecting from a

large pool of random RNA oligos. Analyzing and interpreting the results in order to calculate relative affinities is non-trivial and requires novel analytic tools.

There are challenges in understanding RBNS data. Firstly, RBPs bind sequences of 5-8 nucleotides, but to study the effects of context and RNA structure on binding one must use longer RNA oligos. Determining which motifs are bound within a larger sequence requires novel methods. Secondly, RBNS data are normalized to the total number of reads in the library. Determining the total extent of binding in a library can addressed in a number of ways.

## 4.2 Method Overview

In the experiment, protein is incubated with random RNA oligos (typically 40mers, optionally with sequencing adapters attached) at each of several concentrations of protein with a fixed concentration of RNA. The RBP is selected and a deep-sequencing library is generated from the associated RNA for deep sequencing. The counts of sequences in this library are used to estimate concentrations of bound RNA molecules. The major adjustable experimental parameters are the concentration of RBP and the length and concentrations of the RNA oligos.

$$[R] = \text{Total concentration of RBP} \tag{4.1}$$

$$[O_{total}] = \text{Total concentration of RNA oligo library of random sequences} \tag{4.2}$$

$$\lambda = \text{Length of RNA oligonucleotide (excluding adapters)} \tag{4.3}$$

In addition to the adjustable experimental parameters, our model of the experiment includes observable variables we can measure from the data and and the underlying hidden variables.

## 4.3 Model Definitions

In our model, we consider RNA oligos of length $\lambda$ in solution with RBP. The RBP may bind one of the $k$mers within an oligo, forming a complex. We assume that all of the RBP-oligo complexes are pulled down, bringing with it all complexed RNA. This is reasonable because we use an excess of magnetic beads with very high affinity for the RBPs' tags.

$$[L_i] = \text{Total concentration of } k\text{mer i} \tag{4.4}$$

$$[RL_i] = \text{Concentration of oligos with protein bound to } k\text{mer i} \tag{4.5}$$

$$[L_{i,free}] = \text{Concentration of free } k\text{mer i, i.e. not bound by RBP} \tag{4.6}$$

Defined in terms of the total concentration of protein-RNA complex, [RL$_{total}$] the library fraction of a $k$mer is the fraction of the bound RBPs which are bound to that $k$mer, at the RBP concentration, C:

$$F_{i,C} = \frac{[RL_i]}{[RL_{total}]} \tag{4.7}$$

### 4.3.1 Relative $K_d$ definition

The goal of these analyses will be to calculate the relative $K_d$ for all $k$mer sequences which bind the RBP. The relative $K_d$ for a given $k$mer, $k$mer i is defined as the ratio of the absolute $K_d$ of $k$mer i for the RBP to the absolute $K_d$ for the strongest binding $k$mer, $K_d$.

$$K_{d,i,relative} = \frac{K_{d,i}}{K_{d,best}} = \frac{\frac{[L_{i,free}] \cdot [R_{free}]}{[RL_i]}}{\frac{[L_{best,free}] \cdot [R_{free}]}{[RL_{best}]}} \tag{4.8}$$

We would like to calculate these underlying values from the observable data.

## 4.4 Simple Binding Model

We first present a simplified model of protein RNA binding, which makes several simplifying assumptions in order to provide analytical insight into the system before our most general model.

Consider the situation where an RBP at concentration C binds to a motif, X of length $k$ with a $K_d$ B times lower than any other $k$mer. The protein is incubated in a pool of random RNA oligos of length $\lambda$ with all sequences equally present. The frequency of each $k$mer in this input is thus:

$$F_{input} = \frac{1}{4^k} \tag{4.9}$$

Let us define M to be an arbitrary background motif with no specific binding

affinity for the RBP. Then take the $K_d$ of X relative to this background motif.

$$K_{d,X,relative} = \frac{\frac{[R_{free}] \cdot [X_{free}]}{[RX]}}{\frac{[R_{free}] \cdot [M_{free}]}{[RM]}} \tag{4.10}$$

For small protein concentrations $[X_{free}]$ is approximately equal to $[M_{free}]$. Thus, the formula simplifies to the ratio of bound M to bound X

$$K_{d,X,relative} = \frac{[RM]}{[RX]} \tag{4.11}$$

### 4.4.1 Relation to observable data

In this section for simplicity we work with B rather than $K_d$.

For a given protein, the proprotion that it binds X is a function of B (the preference of the RBP for X) and the number of other sequences that are competing with X to bind B. X is B times more likely to be bound than any other sequence, which all have equal affinities for the RBP. We define $Pr_X$ to be the proportion of protein that is bound to X.

$$Pr_X = \frac{B}{B + 4^k - 1} \tag{4.12}$$

In the oligos bound by RBP, X will be present one time at the bound site as well as at the input frequency at all the other $\lambda - k + 1$ sites in the oligo. Thus the overall frequency of X in the set of oligos bound by RBP at X is:

$$F_{X,bound@X} = \frac{1}{\lambda - k + 1} + \frac{\lambda - k}{\lambda - k + 1} \cdot F_{input} \tag{4.13}$$

The frequency of X occurring in oligos which are not bound at an X position, $F_{X,unbound@X}$ is equal to the input frequency.

We define the Raw Reads Ratio, or R of a motif to be the frequency of that motif overall normalized by the input frequency.

$$R_X = \frac{F_X}{F_{input}} \tag{4.14}$$

R is the observable quantity from an RBNS experiment. By weighting the frequency of X in the pools of oligos bound and unbound at X by the respective proportions (from equation 4.12) we can derive $R_X$.

$$R_x = \frac{F_{X,bound@X} \cdot Pr_X + F_{X,unbound@X} \cdot (1 - Pr_X)}{F_{input}} \tag{4.15}$$

Substituting equation 4.12 into 4.15 yields an equation for $R_X$ in terms of B, $\lambda$ and k.

$$R_x = \frac{\frac{B}{B + 4^k - 1} \cdot \left( \frac{1}{\lambda - k + 1} + \frac{\lambda - k}{\lambda - k + 1} \cdot \frac{1}{4^k} \right) + \frac{4^k - 1}{B + 4^k - 1} \cdot \frac{1}{4^k}}{\frac{1}{4^k}} \tag{4.16}$$

This equation expresses the observable signal, R in terms of the unobservable affinity, B and the experimental parameters $\lambda$ and k. Under normal conditions we can make the approximation that B+ $4^k$ is much greater than one, $4^k$ is much greater than $\lambda$ and that $4^k - 1$ is approximately equal to $B + 4^k$. Applying these gives the

approximate formula of R.

$$R \approx 1 + \frac{4^k \cdot \frac{B}{B + 4^k}}{\lambda - k + 1} \tag{4.17}$$

The original, exact equation (4.16) can be solved analytically for B.

$$B = \frac{(R_x - 1) \cdot (4^k - 1) \cdot (\lambda - k + 1)}{4^k + \lambda - k - R_x \cdot (\lambda - k + 1)} \tag{4.18}$$

### 4.4.2 Modeling Nonspecific Binding

Equation 4.18 assumes that there is no noise in the experiment. However, in reality there is RNA pulled down nonspecifically. Of the RNA pulled down in the experiment a certain amount, C is bound to the protein (ideally, the majority of RNA) and a certain amount, N is nonspecifically pulled down. In the low protein case, it is fair to assume that all protein is bound. Thus the ratio of protein selected RNA to non protein selected RNA is C to C + N.

The frequency of X in nonspecific RNA is $F_{random}$ and the frequency of X in selected RNA is given by equation 4.16 as before. Thus equation 4.16 can be corrected to reflect the effects of nonspecific binding as follows.

$$R_{X,ns} = R_X \cdot \frac{C}{N + C} + \frac{N}{N + C} \tag{4.19}$$

Expanded this takes the form:

$$R_{X,ns} = \frac{\left(\frac{B}{B+4^k-1} \cdot \left(\frac{1}{\lambda-k+1} + \frac{\lambda-k}{\lambda-k+1} \cdot \frac{1}{4^k}\right) + \frac{4^k-1}{B+4^k-1} \cdot \frac{1}{4^k}\right) \cdot \frac{C}{N+C} + \frac{1}{4^k} \cdot \frac{N}{N+C}}{\frac{1}{4^k}} \tag{4.20}$$

Thus the amount of nonspecific RNA pulled down by the apparatus will dampen $R_X$ in proportion to how much of the total RNA is nonspecific. Since N is constant with protein concentration, this effect is very weak when C is high.

Since $R_{X,ns}$ is known and B and N can be solved simultaneously by doing the experiment with multiple protein concentrations. This allows the nonspecific RNA to be estimated as well as the B value which takes into account the presence of nonspecific RNA to be estimated. This value can be compared to a value calculated from fluorescence measurements.

### 4.4.3 Insight from simplified model

The simplified model thus gives us several insights into the broad behavior of the system. First it shows that while increasing the length of the oligo introduces noise in the form of additional sequences brought down with the bound sequence, this happens in a predictable way as a function of $\lambda$. Second it shows that the if there is consistent nonspecific RNA selected to the experimental apparatus then doing RBNS experiments at multiple concentrations of RBP yields a system of equations allowing the concentration of nonspecific RNA to be estimated. We solved this for each pair of concentrations from 12 nM, 60 nM and 312nM, yielding a B value of 800 and N of 20 nM. This is very close to measured value of 20 nM indicating that

the multiple equations can be used in lieu of doing a fluorescent measurement of nonspecific RNA concentration. Lastly, the simplified analysis shows that the relative $K_d$ of a given motif is approximately inversely proportional to the concentration of that motif bound.

## 4.5 Detailed Analysis

### 4.5.1 Estimating *k*mer library fractions, $F_{i,C}$

It is necessary to estimate $F_{i,C}$ since it is not observable in RBNS data. Here we consider two general methods for estimating these. The more naive method is to assign equal weight to every $k$mer present within the oligo. The algorithm we introduce here, SKA, offers several advantages over the naive method and other existing algorithms for finding motifs in sequence data.

#### Raw Counts method

Library fractions estimated from the raw $k$mer counts are designated, $CF_{i,C}$. For each protein concentration these are calculated as the raw counts of $k$mer i in the selected library divided by the total number of all $k$mers in the sequenced library. $CF_{i,C}$ is defined to be the number of occurrences of $k$mer i in library C normalized by the sum of occurences of all $k$mers.

As an illustration of the limitation of this method consider the case of a very strong binder, $k\text{mer}_s$, that is bound in every single read in the selected library. Assume there are D oligo reads in the sequence data from this library. There will therefore be D occurrences of $k\text{mer}_s$. Each oligo has a total of $\lambda - k + 1$ $k$mers. If the remaining

$k$mers are perfectly random then there will be $D \cdot (\lambda - k + 1)$ total $k$mers present. The estimated library fraction of $k\text{mer}_s$ is thus:

$$CF_{i,C} = \frac{D}{D \cdot (\lambda - k + 1)} = \frac{1}{\lambda - k + 1} \tag{4.21}$$

This imposes a theoretical upper limit on the enrichment for $k\text{mer}_s$, when a single binding site is sufficient for binding. In the limit of high specificity for a single $k$mer, $k\text{mer}_s$ will be present in every single oligo. Assuming that the other $\lambda - k$ $k$mers in each oligo are unbiased, the upper limit, UL is a simple function of k and $\lambda$:

$$UL = \frac{4^k}{\lambda - k + 1} \tag{4.22}$$

For a $\lambda$ of 40 and k of 6 the upper limit is 120 which is about six fold over the maximum enrichment observed empirically for RBFOX2.

Another issue with the raw counting method is that $k$mers which have an overlapping sequence with the true motif are artificially inflated. For example for every 1024 occurrences of a bound TGCATG there will be 256 occurrences of GCATGC and 64 occurrences CATGCT. For any given motif (non-homopolymeric) there are 8 sequences that are just shifted by one base. Not being able to distinguish sequences which are truly bound from those that merely overlap with sequences that are truly bound dampens the signal of the truly bound motif.

### 4.5.2 Streaming *k*mer assignment: SKA

This problem is analogous to the problem of determining which genomic region to assign a read to when a read may map ambiguously. This has been addressed by

described by (Roberts & Pachter 2012) with a "streaming" algorithm for read assignment, called eXpress. SKA works analogously to eXpress, with several modifications since the problem are not identical. For example eXpress uses a "forgetting mass" whereas SKA uses multiple passes through the data. Estimates of library fractions from SKA are designated $SF_{i,C}$.

SKA is more appropriate for this problem than other existing motif finding algorithms such as MEME, because the goal is to quantitatively estimate the number of reads attributable to each motif rather than identify significantly enriched motifs present in a set of sequences.

### SKA: Step-by-Step description

The three general steps of this method are:

1) Initialization, 2) First pass through the reads 3) Subsequent passes through reads.

**Initialization** The initialization step is to assign pseudocount binding weight of 1.0 to all $k$mers.

**First Pass** In the first pass through the reads, each read starts with a weight of 1.0. Reads are then sequentially passed through, with a read's weight fractionally assigned to the $k$mers present in its sequence in proportion to their binding weight. For example, on the first read, the binding weights are equal from the initialization pseudocounts so the read's weight will be split equally among all the $k$mers present. After each read, the binding weights are incremented by the fraction of the read assigned to each $k$mer so now all the $k$mers which are present in the first read will have slightly higher binding weight. This process continues over the rest of the reads

in the library. As strong binding $k$mers are present in a larger fraction of the reads than weak binders, their binding weights will increase more rapidly the weak binders which are present in fewer reads. As their binding weights get larger the algorithm learns to assign more and more of each read's weight to them. Thus if there is a very strong binder present in a read (a $k$mer with a high binding weight) the read will be mostly assigned to that $k$mer. If there are two $k$mers present each with a strong binding site then a large fraction of the read's weight will be assigned to each. In reads where the $k$mers all have similar binding weights, the algorithm will split the weight more or less evenly among the $k$mers.

After completing a pass through the reads the algorithm has learned a good approximation of the distribution of $k$mers in the library. However, the distribution may not have converged yet, especially if the distribution is not strongly biased. To address this SKA takes one or more subsequent passes through the reads using the binding weights calculated by the previous pass.

**Subsequent Passes** After completing the First pass through the reads, the algorithm takes the binding weights from the first pass and uses them to reassign all the reads in the library. In this pass a new set of weights are calculated, starting with 0 weights for each $k$mer, by iterating through the reads a second time and assigning them fractionally to each $k$mer based on the binding weights from the previous pass. The key difference between the first pass and subsequent passes is that the reads are assigned based on a set of weights calculated during the previous pass and the new set of binding weights is not used for read assignment (until the next pass).

After a set number of passes through the reads have been completed (anywhere from 2-10), SKA returns the binding weights as an approximate measure of the fraction of the complexed RNA is attributable to binding to each $k$mer.

### 4.5.3 Streaming *k*mer assignment method: Validation

To verify that SKA accurately measures the library fractions, $F_i$, the algorithm was tested extensively on simulated data. A diverse set of simulated data were generated to test how well SKA would work under various circumstances. The data were generated under the assumption that a certain fraction of the reads in a given library were pulled down due to the presence of a given strong binding *k*mer and the rest were pulled down nonspecifically and are therefore random. The specifically pulled down reads had the *k*mer in a random location within a random oligo sequence. This was done for the case of having a single strong binder bound at several different fractions of the library. Note that the *k*mer is present in the random sequence a predictable amount of the time in addition to the amount of time it is bound. Multiple binders were tested by taking two binding sequences (TGCATG and TGCACG) and adding them in at different levels. Another major concern with the naive counting method is quantitatively distinguishing cases where there are two overlapping sequences both confer binding specificity from overlapping sequences where one confers binding specificity and the other is enriched only by virtue of overlapping with the strong binder. The overlapping sequences TGCATG and GCATGC were added into simulated data at different frequencies, to test if SKA can quantitatively distinguish the enrichment attributable to each.

In all of the tested cases SKA calculated binding weights converged rapidly to the correct bound library fractions. Tests where the true library fraction is small converge more slowly than when the true library fraction is high. Supplemental Figure 5A illustrates this. Whereas in the case where the true library fraction is 0.20 converges after a few hundred thousand reads in the first pass, the simulated data with a true library fraction of 0.01 takes several passes through the library to converge, illustrating the utility of subsequent passes. SKA also converges in the

case of multiple motifs (two of the tested simulated data sets shown in Supplemental Figure 5B) with the same effect that the higher library fraction reads takes longer to converge. This is also the case with overlapping motifs (Supplemental Figure 5C).

To illustrate the advantages of SKA over the naive method of calculating enrichment we compare them for simulated and real data. The naive algorithm was used to calculate library fractions for the same simulated data as with SKA. The most salient difference is that whereas the naive method does not correctly calculate the correct library fractions, SKA does (Supplemental Figure 6A and 6B). A second related point is that SKA correctly assigns library fractions for relative affinities spanning multiple orders of magnitude, while naive enrichment values are much less sensitive to these large relative affinity differences. Thus SKA is needed to quantitatively distinguish the strongest binders. This reflects the case in experimental data where strong binding sequences bind RBPs with many orders of magnitude more affinity (Supplemental Figure 6C). The relationship between the Naive results and SKA appears to be sigmoid (on a log-log scale). However, the SKA normalized library fractions are spread over many more orders of magnitude.

Since SKA converges well for simulated data under the described assumptions it is assumed for the purpose of calculating relative $K_d$s that the library fractions of the $k$mers are equal to the normalized binding weights given by SKA.

### 4.5.4 Biochemical Assumptions

This method makes several assumptions about the underlying chemistry in the experiment.

### Equilibrium

All binding reactions are assumed to be equilibrated at the time of the pull-down. This is necessary for the dissociation constant equation 4.8 to hold and is usually the case during an incubation.

### The concentration of each *k*mer is constant in the input to each experiment

The pool of random input RNA may have slight sequence biases. It is assumed that any such biases are represented to the same extent in the sequenced input pool as in the pools used for pull-down with RBP present. Since they each come from the same stock this is reasonable. When multiplexed in an Illumina Hi-Seq lane, each library has sufficient coverage to estimate the abundance of $k$mers up to length 10.

### The experiment does not cause biases in nonspecific RNA

Not all the RNA pulled down from the incubating pool is from bound complex with an RBP. There is nonspecific RNA pulled down by other aspects of the apparatus (beads, etc.). When RNA Bind-n-Seq is performed in the absence of RBP much less RNA is recovered than when RBP concentration is high as measured by fluorescence, indicating that the amount of nonspecific RNA makes up an insignificant portion of the high protein concentration libraries. Additionally, sequencing reveals the sequence content of libraries done in the absence of RBP is highly similar to the input pool. Together these indicate that nonspecific binding to the protocol apparatus is expected to dampen the signal to a small degree rather than systematically bias it.

### Total concentration of *k*mer

One difficulty when dealing with equations designed for small molecule ligands is that there are many $k$mers tiled across each oligo. Here, the $k$mer count is considered to be the sum of each time a given $k$mer occurs, including overlapping $k$mers. This means that overall the summed concentrations of the $k$mers will exceed the concentration of oligo by a factor of $\lambda - k + 1$. One concern is that two overlapping sites are considered to be independent, but in reality two overlapping sites cannot be simultaneously bound by RBP. At low RBP concentration this approximation is acceptable as likelihood of a single oligo being bound by multiple RBPs is very small.

$$[L_{total}] = [O_{total}] \cdot (\lambda - k + 1) \tag{4.23}$$

### No Protein Sponges

It is assumed that all protein is either free in solution or bound to an RNA oligo and none is bound to the apparatus or other object. This means that:

$$[R_{total}] = [R_{free}] + \sum_i [RL_i] \tag{4.24}$$

### All RNA in library was bound by protein

It is assumed that the concentration of protein-RNA complex for each $k$mer is proportional to the total amount of complex and the $k$mer fraction of the library.

$$[RL_{total}] = \sum_i [RL_i] \tag{4.25}$$

A corollary of this assumption is that at equilibrium all the oligo is either free in solution or bound by protein. Likewise each instance of each $k$mer is either bound directly by protein, free in solution or part of an oligo which is bound elsewhere. Since SKA is designed to ignore $k$mers that have been "brought along for the ride," $k$mers which are part of an oligo bound elsewhere are counted as being in the unbound, free $k$mer and only $k$mers which are bound at that position are counted in $[RL_i]$.

$$[L_i] = [L_{i,free}] + [RL_i] \tag{4.26}$$

## 4.6 Estimating relative $K_d$s

Given the above assumption (4.26), we can simplify the formula for $K_{d,i,relative}$ given in 4.8 to this form:

$$K_{d,i,relative} = \frac{([L_i] - [RL_i]) \cdot [RL_{best}]}{([L_{best}] - [RL_{best}]) \cdot [RL_i]} \tag{4.27}$$

### 4.6.1 Estimating $[RL_i]$ and $[RL_{best}]$

From 4.7) and the assumption that SKA estimates the fraction of the bound pool that is bound at $k$mer i, $[RL_i]$ and $[RL_{best}]$ can be estimated. $[RL_{total}]$ can be estimated from a fluorescence measurement.

$$[RL_i] = SF_i \cdot [RL_{total}] \tag{4.28}$$

Note that these library fractions and complex concentrations can be estimated

independently for each library. The shortened names are shown for simplicity.

### 4.6.2 Estimating $[L_{i,free}]$ and $[L_{best,free}]$

The amount of unbound $k$mer i is the total minus the bound. The concentration of bound $k$mer i is given by 4.28 and the total amount of $k$mer i present in solution can be estimated from the sequenced input library. Since it is assumed that each oligo is bound by at most one protein, additional occurrences of a $k$mer within an oligo that already has one occurrence does not count. Therefore, for this purpose the concentration of $k$mer i is defined to be the concentration of RNA oligos which have $k$mer i present in their sequence. Therefore, $[L_i]$ is equal to the total RNA oligo concentration times the fraction of input oligos which have that sequence. We define this as the presence fraction of $k$mer i, $F_{presence,i,C}$ to be the fraction of oligos in library C which contain $k$mer i. Given this definition the total concentration of $k$mer i in solution can be calculated.

$$[L_{i,free}] = [O_{total}] \cdot F_{presence,i} - SF_i \cdot [RL_{total}] \tag{4.29}$$

### 4.6.3 Derived $K_d$ equation

By substitution the derived values from 4.28 and 4.29 into equation 4.27 we can derive a formula for the relative $K_d$'s.

$$K_{d,i,relative} = \frac{([O_{total}] \cdot F_{presence,i} - SF_i \cdot [RL_{total}]) \cdot SF_{best}}{([O_{total}] \cdot F_{presence,best} - SF_{best} \cdot [RL_{total}]) \cdot SF_i} \tag{4.30}$$

Since the SF and [RL] vary depending on the concentration of RBP in the library,

the relative $K_d$ can be calculated for each library separately. We observe that relative $K_d$'s calculated at different protein concentrations are highly correlated to each other. However, the strongest correlation is observed when the signal (maximum enrichment of the strongest binding $k$mer) is strongest. This is likely due to there being the best signal to noise ratio in this equilibrium.

### 4.6.4 Comparison to the Results of the Simplified Model

The detailed model and the simplified model each yield quantitative measures of how strongly each sequence binds to the RBP. While the simplified model has the advantage of simplicity, the SKA algorithm and the full formula for the relative $K_d$ has the advantages of taking into account the biased distribution of sequence in the input pool and correcting for sequences which overlap strong-binding motifs but confer no specificity themselves. There is a high degree of correlation between the inverse relative $K_d$ and B values, especially among the strongest binding motifs (Figure S2B).

## 4.7 Conclusion

This document overviews the theoretical framework and methodology for determining relative $K_d$'s from RNA Bind-n-Seq data. It also introduces SKA, an algorithm for quantitative motif analysis of very deep data sets. Overall, the methods detailed here allow a thorough analysis of the binding preferences of an RBP from RNA Bind-n-Seq data. In addition, SKA is likely to have many uses for quantitating motif contributions to binding beyond the analysis of RNA Bind-n-Seq data, such as for ChIP-seq.

# Chapter 5

# Conclusion

## 5.1 Summary

This thesis has presented an integrative study of NMD as well and a novel experimental method for measuring protein-RNA binding affinities. Chapter 2 discusses analyses which attempt to provide a more complete answer to the question: what triggers NMD. RNA-seq, Ribo-seq and CLIP-seq data are used to get a broad, transcriptome-wide look at determinants of NMD in mouse embryonic stem cells. In addition to the known determinants of downstream exon junctions and elongated 3' UTRs, these data give evidence that direct binding by Upf1 or translation of a uORF trigger NMD, while conversely low translation of a message may protect it from NMD. Of particular interest was the observation that Upf1 bound genes were regulated in human cells as well as in our experiment, indicating conservation of function. Chapter 3 presents the novel in vitro method RNA Bind-n-seq which takes advantage of the technology in genome sequencers. We observe that there are secondary motifs to RNA binding proteins which direct binding to a lesser extent than previously identified motifs. We see that there are strong, complex effects of structure on RBP binding, which vary

from protein to protein. Also, we were somewhat surprised by the degree to which the inherent binding affinities of RBPs were correlated with the in vivo binding patterns. We also noted a potentially synergistic relationship between CLIP and RBNS data. In chapter 4 I describe the mathematical underpinnings of the RBNS analyses, describe a simple model which can yield important insight for experiment design, and describe a novel algorithm for motif finding.

## 5.2 Future directions

### 5.2.1 RBNS in conjunction with CLIP-seq

RBNS and CLIP-seq yield very different data and by combining them one can improve understanding of protein-RNA interactions. CLIP-seq, and especially the iCLIP variant, is very powerful method for detecting in vivo binding sites with nucleotide precision. However, it is specific to the particular cell type and the genotype chosen for a particular experiment. For the proteins studied so far we observe that there are motifs which are enriched in CLIP-seq but not RBNS (possibly due to differences in cross-linking efficiency between nucleotides), but there do not seem to be motifs which are enriched in RBNS but not CLIP-seq. Since RBNS is unbiased by cross-linking and covers all sequence space up to a certain length, it can be used to filter and improve CLIP-seq. It can also be used to identify cases where binding is directed by a second factor, since those will not be enriched in an isolated RBNS experiment.

### 5.2.2 Potential experimental extensions of RBNS

There are many potential variants of RBNS that could be done in order to get study different aspects of protein-RNA interaction. To study structure in more depth, one

could generate libraries with longer read lengths to allow more scope for structure to develop. To study certain aspects of cooperativity in protein-RNA binding, one could do the experiment on full-length proteins, including their protein-protein interaction domains. Oftentimes in splicing, outcomes are determined by competition between factors. This can be studied via RBNS by pulling down one tagged factor in the presence of another factor. The concentrations of each can be varied in order to generate results for different relative expression patterns. In order to hone in on particularly interesting genomic sequences, they can be spiked into an RBNS experiment as long as the input is sequenced as well. Each of these variations on RBNS addresses a slightly different question in protein-RNA binding.

### 5.2.3 Potential other applications of SKA algorithm

In order to properly analyze the data from RBNS experiments we developed an SKA, an algorithm which identifies and quantifies motif presence in a library, given a set of assumptions. SKA is sufficiently efficient to analyze data sets which exceed the genome in size, an important consideration in high throughput in vitro studies. SKA operates under the assumption that for every sequence in a library, there is a motif within that sequence which is responsible for it being selected. The goal is to generate aggregate statistics about how much each possible motif contributes to the library. In RBNS the motifs represent binding to RBPs, but the algorithm may be applied to any other situation where the same assumptions hold true (or approximately true). Thus, for example, in ChIP-seq it could quantify contributions of transcription factor binding sites.

### 5.2.4 RNA Bind-n-seq and human genetics

As sequencing of human genomes and transcriptomes becomes more common, the understanding the consequences of genetic polymorphisms between individuals becomes more and more important. At this point there is sufficient data available to identify hundreds of genetic polymorphisms which alter splicing patterns, called splicing quantitative trait loci or sQTLs (Pickrell et al. 2010). Many sQTLs disrupt splicing by breaking the splice sites, however, many do not. With RBNS data for a sufficiently comprehensive set of RBPs, one could predict how protein-RNA interactions differ between the different genotypes as well as the downstream effects of such differences in protein binding.

# Bibliography


Adereth Y, Dammai V, Kose N, Li R & Hsu T 2005 *Nat Cell Biol* **7**(12), 1240–7.

Altschul S F, Gish W, Miller W, Myers E W & Lipman D J 1990 *Journal of Molecular Biology* **215**(3), 403–410.

Amrani N, Ganesan R, Kervestin S, Mangus D A, Ghosh S & Jacobson A 2004 *Nature* **432**(7013), 112–8.

Anastasaki C, Longman D, Capper A, Patton E E & Caceres J F 2011 *Nucleic acids research* **39**(9), 3686–94.

Arciga-Reyes L, Wootton L, Kieffer M & Davies B 2006 *The Plant journal : for cell and molecular biology* **47**(3), 480–9.

Auweter S D, Fasan R, Reymond L, Underwood J G, Black D L, Pitsch S & Allain F H 2006 *EMBO J* **25**(1), 163–73.

Avery P, Vicente-Crespo M, Francis D, Nashchekina O, Alonso C R & Palacios I M 2011 *RNA* **17**(4), 624–38.

Azzalin C M & Lingner J 2006 *Current biology : CB* **16**(4), 433–9.

Bailey T & Elkan C 1994 *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* pp. 28–36.

Bailey T L, Boden M, Buske F A, Frith M, Grant C E, Clementi L, Ren J Y, Li W W & Noble W S 2009 *Nucleic Acids Research* **37**, W202–W208.

Baraniak A P, Chen J R & Garcia-Blanco M A 2006 *Mol Cell Biol* **26**(4), 1209–22.

Barash Y, Calarco J A, Gao W, Pan Q, Wang X, Shai O, Blencowe B J & Frey B J 2010 *Nature* **465**(7294), 53–9.

Behm-Ansmant I, Gatfield D, Rehwinkel J, Hilgers V & Izaurralde E 2007 *The EMBO journal* **26**(6), 1591–601.

Bentley D R, Balasubramanian S, Swerdlow H P, Smith G P, Milton J, Brown C G, Hall K P, Evers D J, Barnes C L, Bignell H R, Boutell J M, Bryant J, Carter R J, Cheetham R K, Cox A J, Ellis D J, Flatbush M R, Gormley N A, Humphray S J, Irving L J, Karbelashvili M S, Kirk S M, Li H, Liu X H, Maisinger K S, Murray L J, Obradovic B, Ost T, Parkinson M L, Pratt M R, Rasolonjatovo I M J, Reed M T, Rigatti R, Rodighiero C, Ross M T, Sabot A, Sankar S V, Scally A, Schroth G P, Smith M E, Smith V P, Spiridou A, Torrance P E, Tzonev S S, Vermaas E H, Walter K, Wu X L, Zhang L, Alam M D, Anastasi C, Aniebo I C, Bailey D M D, Bancarz I R, Banerjee S, Barbour S G, Baybayan P A, Benoit V A, Benson K F, Bevis C, Black P J, Boodhun A, Brennan J S, Bridgham J A, Brown R C, Brown A A, Buermann D H, Bundu A A, Burrows J C, Carter N P, Castillo N, Catenazzi M C E, Chang S, Cooley R N, Crake N R, Dada O O, Diakoumakos K D, Dominguez-Fernandez B, Earnshaw D J, Egbujor U C, Elmore D W, Etchin S S, Ewan M R, Fedurco M, Fraser L J, Fajardo K V F, Furey W S, George D, Gietzen K J, Goddard C P, Golda G S, Granieri P A, Green D E, Gustafson D L, Hansen N F, Harnish K, Haudenschild C D, Heyer N I, Hims M M, Ho J T, Horgan A M et al. 2008 *Nature* **456**(7218), 53–59.

Bhattacharya A, Czaplinski K, Trifillis P, He F, Jacobson A & Peltz S W 2000 *RNA* **6**(9), 1226–35.

Black D L 2003 *Annual Review of Biochemistry* **72**, 291–336.

Blanchette M & Chabot B 1999 *Embo Journal* **18**(7), 1939–1952.

Blanchette M, Green R E, MacArthur S, Brooks A N, Brenner S E, Eisen M B & Rio D C 2009 *Molecular Cell* **33**(4), 438–449.

Bohnsack M T, Martin R, Granneman S, Ruprecht M, Schleiff E & Tollervey D 2009 *Molecular cell* **36**(4), 583–92.

Bradley R K, Merkin J, Lambert N J & Burge C B 2012 *PLoS biology* **10**(1), e1001229.

Breitbart R E, Andreadis A & Nadalginard B 1987 *Annual Review of Biochemistry* **56**, 467–495.

Buhler M, Steiner S, Mohn F, Paillusson A & Muhlemann O 2006 *Nature structural & molecular biology* **13**(5), 462–4.

Buratti E & Baralle F E 2004 *Molecular and Cellular Biology* **24**(24), 10505–10514.

Burrows M & Wheeler D 1994 *Technical Reports, Digital Equipment Corporation* **124**.

Calfon M, Zeng H, Urano F, Till J, Hubbard S, Harding H, Clark S & Ron D 2002 *Nature* **415**, 92–96.

Calvo S E, Pagliarini D J & Mootha V K 2009 *Proceedings of the National Academy of Sciences of the United States of America* **106**(18), 7507–12.

Carothers J, Oestreich S & Szostak J 2006 *J Am Chem Soc* **128**, 7929–37.

Cartegni L, Chew S L & Krainer A R 2002 *Nature Reviews Genetics* **3**(4), 285–298.

Carter M S, Doskow J, Morris P, Li S, Nhim R P, Sandstedt S & Wilkinson M F 1995 *The Journal of biological chemistry* **270**(48), 28995–9003.

Cass D, Hotchko R, Barber P, Jones K, Gates D P & Berglund J A 2011 *BMC Mol Biol* **12**, 20.

Castle J C, Zhang C, Shah J K, Kulkarni A V, Kalsotra A, Cooper T A & Johnson J M 2008 *Nat Genet* **40**(12), 1416–25.

Chakrabarti S, Jayachandran U, Bonneau F, Fiorini F, Basquin C, Domcke S, Le Hir H & Conti E 2011 *Molecular cell* **41**(6), 693–703.

Chamieh H, Ballut L, Bonneau F & Le Hir H 2008 *Nature structural & molecular biology* **15**(1), 85–93.

Chang Y F, Imam J S & Wilkinson M F 2007 *Annual review of biochemistry* **76**, 51–74.

Chen C D, Kobayashi R & Helfman D M 1999 *Genes & Development* **13**(5), 593–606.

Chen M & Manley J L 2009 *Nature Reviews Molecular Cell Biology* **10**(11), 741–754.

Cheng J & Maquat L E 1993 *Molecular and Cellular Biology* **13**(3), 1892–1902.

Chester A, Somasekaram A, Tzimina M, Jarmuz A, Gisbourne J, O'Keefe R, Scott J & Navaratnam N 2003 *The EMBO journal* **22**(15), 3971–82.

Chi S W, Zang J B, Mele A & Darnell R B 2009 *Nature* **460**(7254), 479–486.

Cho H, Kim K M, Han S, Choe J, Park S G, Choi S S & Kim Y K 2012 *Molecular cell* **46**(4), 495–506.

Choe J, Cho H, Lee H C & Kim Y K 2010 *EMBO reports* **11**(5), 380–6.

Clark T A, Sugnet C W & Ares M 2002 *Science* **296**(5569), 907–910.

Conti E & Izaurralde E 2005 *Current opinion in cell biology* **17**(3), 316–25.

Culler S J, Hoff K G, Voelker R B, Berglund J A & Smolke C D 2010 *Nucleic Acids Research* **38**(15), 5152–5165.

Czaplinski K, Weng Y, Hagan K W & Peltz S W 1995 *RNA* **1**(6), 610–23.

Das R, Yu J, Zhang Z, Gygi M P, Krainer A R, Gygi S P & Reed R 2007 *Molecular Cell* **26**(6), 867–881.

Dasgupta T & Ladd A N 2012 *Wiley Interdiscip Rev RNA* **3**(1), 104–21.

Daughters R, Tuttle D, Gao W, Ikeda Y, Moseley M, Ebner T, Swanson M, & Ranum L 2009 *PloS Genetics* **5**, e1000600.

Davenport T G, Jerome-Majewska L A & Papaioannou V E 2003 *Development* **130**(10), 2263–73.

Dreyfuss G, Kim V N & Kataoka N 2002 *Nature Reviews Molecular Cell Biology* **3**(3), 195–205.

Eberle A B, Stalder L, Mathys H, Orozco R Z & Muhlemann O 2008 *PLoS biology* **6**(4), e92.

Ellington A D & Szostak J W 1990 *Nature* **346**(6287), 818–822.

Fairbrother W G, Yeh R F, Sharp P A & Burge C B 2002 *Science* **297**(5583), 1007–1013.

Feng B, Jiang J, Kraus P, Ng J H, Heng J C, Chan Y S, Yaw L P, Zhang W, Loh Y H, Han J, Vega V B, Cacheux-Rataboul V, Lim B, Lufkin T & Ng H H 2009 *Nature cell biology* **11**(2), 197–203.

Flicek P, Amode M R, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari A,

Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat H S, Rios D, Ritchie G R, Ruffier M, Schuster M, Sobral D, Spudich G, Tang Y A, Trevanion S, Vandrovcova J, Vilella A J, White S, Wilder S P, Zadissa A, Zamora J, Aken B L, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez X M, Herrero J, Hubbard T J, Parker A, Proctor G, Vogel J & Searle S M 2011 *Nucleic acids research* **39**(Database issue), D800–6.

Fox-Walsh K, Davis-Turak J, Zhou Y, Li H R & Fu X D 2011 *Genomics* **98**(4), 266–271.

Franks T M, Singh G & Lykke-Andersen J 2010 *Cell* **143**(6), 938–50.

Fried M & Crothers D M 1981 *Nucleic Acids Research* **9**(23), 6505–6525.

Friedman R C, Farh K K, Burge C B & Bartel D P 2009 *Genome Res* **19**(1), 92–105.

Frizzell K A, Rynearson S G & Metzstein M M 2012 *RNA* **18**(8), 1475–86.

Fu X D 1995 *Rna-a Publication of the Rna Society* **1**(7), 663–680.

Gaba A, Wang Z, Krishnamoorthy T, Hinnebusch A G & Sachs M S 2001 *The EMBO journal* **20**(22), 6453–63.

Garner M M & Revzin A 1981 *Nucleic Acids Research* **9**(13), 3047–3060.

Gatfield D & Izaurralde E 2004 *Nature* **429**(6991), 575–8.

Gatfield D, Unterholzner L, Ciccarelli F D, Bork P & Izaurralde E 2003 *The EMBO journal* **22**(15), 3960–70.

Gehman L T, Meera P, Stoilov P, Shiue L, O'Brien J E, Meisler M H, Ares, M. J, Otis T S & Black D L 2012 *Genes Dev* **26**(5), 445–60.

Giorgi C, Yeo G W, Stone M E, Katz D B, Burge C, Turrigiano G & Moore M J 2007 *Cell* **130**(1), 179–91.

Goers E S, Purcell J, Voelker R B, Gates D P & Berglund J A 2010 *Nucleic Acids Res* **38**(7), 2467–84.

Gong C & Maquat L E 2011 *Nature* **470**(7333), 284–8.

Graveley B R, Hertel K J & Maniatis T 1998 *Embo Journal* **17**(22), 6747–6756.

Grimson A, Farh K K, Johnston W K, Garrett-Engele P, Lim L P & Bartel D P 2007 *Molecular cell* **27**(1), 91–105.

Han J, Ding J H, Byeon C W, Kim J H, Hertel K J, Jeong S & Fu X D 2011 *Mol Cell Biol* **31**(4), 793–802.

Hansen K D, Lareau L F, Blanchette M, Green R E, Meng Q, Rehwinkel J, Gallusser F L, Izaurralde E, Rio D C, Dudoit S & Brenner S E 2009 *PLoS genetics* **5**(6), e1000525.

He F & Jacobson A 2001 *Molecular and cellular biology* **21**(5), 1515–30.

He F, Li X, Spatrick P, Casillo R, Dong S & Jacobson A 2003 *Molecular cell* **12**(6), 1439–52.

He F, Peltz S W, Donahue J L, Rosbash M & Jacobson A 1993 *Proceedings of the National Academy of Sciences of the United States of America* **90**(15), 7034–8.

Ho T H, Charlet B N, Poulos M G, Singh G, Swanson M S & Cooper T A 2004 *EMBO J* **23**(15), 3103–12.

Hodgkin J, Papp A, Pulak R, Ambros V & Anderson P 1989 *Genetics* **123**(2), 301–13.

Hofacker I L 2003 *Nucleic Acids Res* **31**(13), 3429–31.

Hogg J R & Goff S P 2010 *Cell* **143**(3), 379–89.

Holbrook J A, Neu-Yilik G, Hentze M W & Kulozik A E 2004 *Nature genetics* **36**(8), 801–8.

Hosoda N, Kim Y K, Lejeune F & Maquat L E 2005 *Nature structural & molecular biology* **12**(10), 893–901.

Huang L, Lou C H, Chan W, Shum E Y, Shao A, Stone E, Karam R, Song H W & Wilkinson M F 2011 *Molecular cell* **43**(6), 950–61.

Huelga S C, Vu A Q, Arnold J D, Liang T Y, Liu P P, Yan B Y, Donohue J P, Shiue L, Hoon S, Brenner S, Ares M & Yeo G W 2012 *Cell Reports* **1**(2), 167–178.

Huh G & Hynes R 1993 *Mol cell biol* **13(9)**, 5301–5314.

Hwang J & Maquat L E 2011 *Current opinion in genetics & development* **21**(4), 422–30.

Hwang J, Sato H, Tang Y, Matsuda D & Maquat L E 2010 *Molecular cell* **39**(3), 396–409.

Imayoshi I, Sakamoto M, Yamaguchi M, Mori K & Kageyama R 2010 *The Journal of neuroscience : the official journal of the Society for Neuroscience* **30**(9), 3489–98.

Ingolia N T, Ghaemmaghami S, Newman J R & Weissman J S 2009 *Science* **324**(5924), 218–23.

Ingolia N T, Lareau L F & Weissman J S 2011 *Cell* **147**(4), 789–802.

Ishigaki Y, Li X, Serin G & Maquat L E 2001 *Cell* **106**(5), 607–17.

Isken O & Maquat L E 2007 *Genes & development* **21**(15), 1833–56.

Ivanov P V, Gehring N H, Kunz J B, Hentze M W & Kulozik A E 2008 *The EMBO journal* **27**(5), 736–47.

Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, DeCoste C, Schafer X, Lun Y & Lemischka I R 2006 *Nature* **442**(7102), 533–8.

Jean-Philippe J, Pasz S & Caputi M 2013 *Int. J. Mol. Sci.* **14**, 18999–19024.

Jepsen K, Hermanson O, Onami T M, Gleiberman A S, Lunyak V, McEvilly R J, Kurokawa R, Kumar V, Liu F, Seto E, Hedrick S M, Mandel G, Glass C K, Rose D W & Rosenfeld M G 2000 *Cell* **102**(6), 753–63.

Jiang L C, Schlesinger F, Davis C A, Zhang Y, Li R H, Salit M, Gingeras T R & Oliver B 2011 *Genome Research* **21**(9), 1543–1551.

Jin Y, Suzuki H, Maegawa S, Endo H, Sugano S, Hashimoto K, Yasuda K & Inoue K 2003 *EMBO J* **22**(4), 905–12.

Johansson M J, He F, Spatrick P, Li C & Jacobson A 2007 *Proceedings of the National Academy of Sciences of the United States of America* **104**(52), 20872–7.

Johnson J M, Castle J, Garrett-Engele P, Kan Z Y, Loerch P M, Armour C D, Santos R, Schadt E E, Stoughton R & Shoemaker D D 2003 *Science* **302**(5653), 2141–2144.

Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas J M, Yan J, Sillanpaa M J, Bonke M, Palin K, Talukder S, Hughes T R, Luscombe N M, Ukkonen E & Taipale J 2010 *Genome Res* **20**(6), 861–73.

Kalsotra A, Xiao X, Ward A J, Castle J C, Johnson J M, Burge C B & Cooper T A 2008 *Proc Natl Acad Sci U S A* **105**(51), 20333–8.

Kashima I, Yamashita A, Izumi N, Kataoka N, Morishita R, Hoshino S, Ohno M, Dreyfuss G & Ohno S 2006 *Genes & development* **20**(3), 355–67.

Kent W J 2002 *Genome Research* **12**(4), 656–664.

Kervestin S & Jacobson A 2012 *Nature reviews. Molecular cell biology* **13**(11), 700–12.

Kim J, Chu J, Shen X, Wang J & Orkin S H 2008 *Cell* **132**(6), 1049–61.

Kim Y K, Furic L, Desgroseillers L & Maquat L E 2005 *Cell* **120**(2), 195–208.

Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M & Zavolan M 2011 *Nature methods* **8**(7), 559–64.

Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner D, Luscombe N & Ule J 2010 *Nat Struct Mol Biol.* **17(7)**, 909–15.

Kopp J L, Ormsbee B D, Desler M & Rizzino A 2008 *Stem cells* **26**(4), 903–11.

Kornblihtt A R, Schor I E, Allo M, Dujardin G, Petrillo E & Munoz M J 2013 *Nature Reviews Molecular Cell Biology* **14**(3), 153–165.

Kreahling J M & Graveley B R 2005 *Molecular and Cellular Biology* **25**(23), 10251–10260.

Krecic A M & Swanson M S 1999 *Current Opinion in Cell Biology* **11**(3), 363–371.

Kurosaki T & Maquat L E 2013 *Proceedings of the National Academy of Sciences of the United States of America* **110**(9), 3357–3362.

Kuyumcu-Martinez N M, Wang G S & Cooper T A 2007 *Mol Cell* **28**(1), 68–78.

Kuzmiak H A & Maquat L E 2006 *Trends in molecular medicine* **12**(7), 306–16.

Ladd A N, Charlet N & Cooper T A 2001 *Mol Cell Biol* **21**(4), 1285–96.

Lallena M, Chalmers K, Llamazares S & Valcarcel J 2002 *Cell* **Volume 109, Issue 3**, 285–296.

Langmead B, Trapnell C, Pop M & Salzberg S L 2009 *Genome biology* **10**(3), R25.

Lareau L F, Brooks A N, Soergel D A, Meng Q & Brenner S E 2007 *Advances in experimental medicine and biology* **623**, 190–211.

Lareau L F, Inada M, Green R E, Wengrod J C & Brenner S E 2007 *Nature* **446**(7138), 926–9.

Larsson O, Sonenberg N & Nadon R 2010 *Proceedings of the National Academy of Sciences of the United States of America* **107**(50), 21487–92.

Lau N C, Lim L P, Weinstein E G & Bartel D P 2001 *Science* **294**(5543), 858–62.

Le Hir H, Gatfield D, Braun I C, Forler D & Izaurralde E 2001 *EMBO reports* **2**(12), 1119–24.

Le Hir H, Gatfield D, Izaurralde E & Moore M J 2001 *The EMBO journal* **20**(17), 4987–97.

Le Hir H, Izaurralde E, Maquat L E & Moore M J 2000 *The EMBO journal* **19**(24), 6860–9.

Le Hir H, Moore M J & Maquat L E 2000 *Genes & development* **14**(9), 1098–108.

Leeds P, Peltz S W, Jacobson A & Culbertson M R 1991 *Genes & development* **5**(12A), 2303–14.

Lejeune F, Ishigaki Y, Li X & Maquat L E 2002 *The EMBO journal* **21**(13), 3536–45.

Leung A K, Young A G, Bhutkar A, Zheng G X, Bosson A D, Nielsen C B & Sharp P A 2011 *Nature structural & molecular biology* **18**(2), 237–44.

Lewis B P, Green R E & Brenner S E 2003 *Proceedings of the National Academy of Sciences of the United States of America* **100**(1), 189–92.

Li H & Durbin R 2009 *Bioinformatics* **25**(14), 1754–1760.

Licatalosi D, Mele A, Fak J, Ule J, Kayikci M, Chi S, Clark T, Schweitzer A, Blume J, Wang X, Darnell J & Darnell R B 2008 *Nature* **456**, 464–469.

Lim L & Sharp 1998 *Mol Cell Biol* **18(7)**, 3900–3906.

Linz B, Koloteva N, Vasilescu S & McCarthy J E 1997 *The Journal of biological chemistry* **272**(14), 9131–40.

Liu H X, Chew S L, Cartegni L, Zhang M Q & Krainer A R 2000 *Molecular and Cellular Biology* **20**(3), 1063–1071.

Liu H X, Zhang M & Krainer A R 1998 *Genes & Development* **12**(13), 1998–2012.

Lu R, Yang A & Jin Y 2011 *The Journal of biological chemistry* **286**(10), 8425–36.

Mankodi A, Lin X, Blaxall B C, Swanson M S & Thornton C A 2005 *Circ Res* **97**(11), 1152–5.

Marquis J, Paillard L, Audic Y, Cosson B, Danos O, Le Bec C & Osborne H B 2006 *Biochem J* **400**(2), 291–301.

Matlin A J, Clark F & Smith C W 2005 *Nature reviews. Molecular cell biology* **6**(5), 386–98.

McCullough A J & Berget S M 1997 *Molecular and Cellular Biology* **17**(8), 4562–4571.

McIlwain D R, Pan Q, Reilly P T, Elia A J, McCracken S, Wakeham A C, Itie-Youten A, Blencowe B J & Mak T W 2010 *Proceedings of the National Academy of Sciences of the United States of America* **107**(27), 12186–91.

Medghalchi S M, Frischmeyer P A, Mendell J T, Kelly A G, Lawler A M & Dietz H C 2001 *Human molecular genetics* **10**(2), 99–105.

Melero R, Buchwald G, Castano R, Raabe M, Gil D, Lazaro M, Urlaub H, Conti E & Llorca O 2012 *Nature structural & molecular biology* **19**(5), 498–505, S1–2.

Mendell J T, Sharifi N A, Meyers J L, Martinez-Murillo F & Dietz H C 2004 *Nature genetics* **36**(10), 1073–8.

Merkin J, Russell C, Chen P & Burge C B 2012 *Science* **338**(6114), 1593–9.

Metzstein M M & Krasnow M A 2006 *PLoS genetics* **2**(12), e180.

Mooers B H, Logue J S & Berglund J A 2005 *Proc Natl Acad Sci U S A* **102**(46), 16626–31.

Moraes K C, Wilusz C J & Wilusz J 2006 *RNA* **12**(6), 1084–91.

Mort M, Ivanov D, Cooper D N & Chuzhanova N A 2008 *Human mutation* **29**(8), 1037–47.

Mortazavi A, Williams B A, Mccue K, Schaeffer L & Wold B 2008 *Nature Methods* **5**(7), 621–628.

Muhlrad D & Parker R 1999 *RNA* **5**(10), 1299–307.

Nagy E & Maquat L E 1998 *Trends in biochemical sciences* **23**(6), 198–9.

Nasim F U H, Hutchison S, Cordeau M & Chabot B 2002 *Rna-a Publication of the Rna Society* **8**(8), 1078–1089.

Ni J Z, Grate L, Donohue J P, Preston C, Nobida N, O'Brien G, Shiue L, Clark T A, Blume J E & Ares, M. J 2007 *Genes & development* **21**(6), 708–18.

Nichols J, Zevnik B, Anastassiadis K, Niwa H, Klewe-Nebenius D, Chambers I, Scholer H & Smith A 1998 *Cell* **95**(3), 379–91.

Nicholson P, Yepiskoposyan H, Metze S, Zamudio Orozco R, Kleinschmidt N & Muhlemann O 2010 *Cellular and molecular life sciences : CMLS* **67**(5), 677–700.

Niwa H, Miyazaki J & Smith A G 2000 *Nature genetics* **24**(4), 372–6.

Noensie E N & Dietz H C 2001 *Nature biotechnology* **19**(5), 434–9.

Nutiu R, Friedman R C, Luo S, Khrebtukova I, Silva D, Li R, Zhang L, Schroth G P & Burge C B 2011 *Nat Biotechnol* **29**(7), 659–64.

Oberstrass F, Auweter S, Erat M, Hargous Y, Henning A, Wenter P, Reymond L, Amir-Ahmady B, Pitsch S, Black D & F.H. A 2005 *Science* **309**, 2054–2057.

Ogawa N & Biggin M D 2012 *Gene Regulatory Networks: Methods and Protocols* **786**, 51–63.

Okunola H L & Krainer A R 2009 *Molecular and Cellular Biology* **29**(20), 5620–5631.

Orengo J P, Ward A J & Cooper T A 2011 *Ann Neurol* **69**(4), 681–90.

Page M F, Carr B, Anders K R, Grimson A & Anderson P 1999 *Molecular and cellular biology* **19**(9), 5943–51.

Pan Q, Saltzman A L, Kim Y K, Misquitta C, Shai O, Maquat L E, Frey B J & Blencowe B J 2006 *Genes & Development* **20**(2), 153–158.

Pan Q, Shai O, Lee L J, Frey J & Blencowe B J 2008 *Nature Genetics* **40**(12), 1413–1415.

Park E, Gleghorn M L & Maquat L E 2013 *Proceedings of the National Academy of Sciences of the United States of America* **110**(2), 405–412.

Park E & Maquat L E 2013 *Wiley Interdisciplinary Reviews-Rna* **4**(4), 423–435.

Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H & Soldatov A 2009 *Nucleic acids research* **37**(18), e123.

Perez I, McAfee J & Patton J 1997 *Biochemistr* **36**, 11881–11890.

Pickrell J, Marioni J, Pai A, Degner J, Engelhardt B, Nkadori E, Veyrieras J, Stephens M, Gilad Y & Pritchard J 2010 *Nature* **464**, 768–772.

Ponthier J, Schluepen C, Chen W, Lersch R, Gee S, Hou V, Lo A, Short, S.A. Chasis J, Winkelmann J & Conboy J 2006 *J Biol* **281(18)**, 12468–74.

Purcell J, Oddo J C, Wang E T & Berglund J A 2012 *Mol Cell Biol* **32**(20), 4155–67.

Ramani A K, Nelson A C, Kapranov P, Bell I, Gingeras T R & Fraser A G 2009 *Genome biology* **10**(9), R101.

Ray D, Kazan H, Chan E T, Pena Castillo L, Chaudhry S, Talukder S, Blencowe B J, Morris Q & Hughes T R 2009 *Nat Biotechnol* **27**(7), 667–70.

Ray D, Kazan H, Cook K B, Weirauch M T, Najafabadi H S, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat L H, Dale R K, Smith S A, Yarosh C A, Kelly S M, Nabet B, Mecenas D, Li W, Laishram R S, Qiao M, Lipshitz H D, Piano F, Corbett A H, Carstens R P, Frey B J, Anderson R A, Lynch K W, Penalva L O, Lei E P, Fraser A G, Blencowe B J, Morris Q D & Hughes T R 2013 *Nature* **499**(7457), 172–7.

Rehwinkel J, Raes J & Izaurralde E 2006 *Trends in biochemical sciences* **31**(11), 639–46.

Roberts A & Pachter L 2012 *Nature Methods* .

Rufener S C & Muhlemann O 2013 *Nature Structural & Molecular Biology* **20**(6), 710–+.

Saltzman A L, Kim Y K, Pan Q, Fagnani M M, Maquat L E & Blencowe B J 2008 *Molecular and cellular biology* **28**(13), 4320–30.

Sanford J R, Ellis J & Caceres J F 2005 *Biochemical Society Transactions* **33**, 443–446.

Sanford J R, Wang X, Mort M, VanDuyn N, Cooper D N, Mooney S D, Edenberg H J & Liu Y L 2009 *Genome Research* **19**(3), 381–394.

Sauliere J, Murigneux V, Wang Z, Marquenet E, Barbosa I, Le Tonqueze O, Audic Y, Paillard L, Roest Crollius H & Le Hir H 2012 *Nature structural & molecular biology* **19**(11), 1124–31.

Schwarzbauer J E, Patel R S, Fonda D & Hynes R O 1987 *Embo Journal* **6**(9), 2573–2580.

Shalgi R, Hurt J A, Krykbaeva I, Taipale M, Lindquist S & Burge C B 2013 *Molecular Cell* **49**(3), 439–452.

Shamoo Y, Abdulmanan N & Williams K R 1995 *Nucleic Acids Research* **23**(5), 725–728.

Shi H, Hoffman B E & Lis J T 1997 *Molecular and Cellular Biology* **17**(5), 2649–2657.

Shin C, Feng Y & Manley J L 2004 *Nature* **427**(6974), 553–558.

Sievers C, Schlumpf T, Sawarkar R, Comoglio F & Paro R 2012 *Nucleic acids research* .

Singh G, Kucukural A, Cenik C, Leszyk J D, Shaffer S A, Weng Z & Moore M J 2012 *Cell* **151**(4), 750–64.

Singh G, Rebbapragada I & Lykke-Andersen J 2008 *PLoS biology* **6**(4), e111.

Sorek R & Ast G 2003 *Genome Research* **13**(7), 1631–1637.

Stockklausner C, Breit S, Neu-Yilik G, Echner N, Hentze M W, Kulozik A E & Gehring N H 2006 *Nucleic acids research* **34**(8), 2355–63.

Sugimoto Y, Konig J, Hussain S, Zupan B, Curk T, Frye M & Ule J 2012 *Genome Biol* **13**(8), R67.

Sugnet C W, Srinivasan K, Clark T A, O'Brien G, Cline M S, Wang H, Williams A, Kulp D, Blume J E, Haussler D & Ares M 2006 *Plos Computational Biology* **2**(1), 22–35.

Sun X, Perlick H A, Dietz H C & Maquat L E 1998 *Proceedings of the National Academy of Sciences of the United States of America* **95**(17), 10009–14.

Sureau A, Gattoni R, Dooghe Y, Stevenin J & Soret J 2001 *Embo Journal* **20**(7), 1785–1796.

Suyama M, Harrington E D, Vinokourova S, Doeberitz M V, Ohara O & Bork P 2010 *Nucleic Acids Research* **38**(22), 7916–7926.

Takagaki Y, Seipelt R L, Peterson M L & Manley J L 1996 *Cell* **87**(5), 941–952.

Takahashi K & Yamanaka S 2006 *Cell* **126**(4), 663–76.

Taneja K L, McCurrach M, Schalling M, Housman D & Singer R H 1995 *J Cell Biol* **128**(6), 995–1002.

Teplova M & Patel D J 2008 *Nat Struct Mol Biol* **15**(12), 1343–51.

Tilgner H, Knowles D G, Johnson R, Davis C A, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras T R & Guigo R 2012 *Genome Research* **22**(9), 1616–1625.

Timchenko L, Miller J, Timchenko N, DeVore D, Datar K, Lin L, Roberts R, Caskey C & Swanson M 1996 *Nucleic acids research* **24**, 4407–4414.

Trapnell C, Pachter L & Salzberg S L 2009 *Bioinformatics* **25**(9), 1105–11.

Trapnell C, Williams B A, Pertea G, Mortazavi A, Kwan G, van Baren M J, Salzberg S L, Wold B J & Pachter L 2010 *Nature biotechnology* **28**(5), 511–5.

Tu M, Tong W, Perkins R & Valentine C R 2000 *Mutation Research-Genomics* **432**(1-2), 15–32.

Ule J, Jensen K B, Ruggiu M, Mele A, Ule A & Darnell R B 2003 *Science* **302**(5648), 1212–5.

Ule J, Jensen K, Mele A & Darnell R B 2005 *Methods* **37**(4), 376–86.

Underwood J G, Boutz P L, Dougherty J D, Stoilov P & Black D L 2005 *Mol Cell Biol* **25**(22), 10005–16.

Valcarcel J, Fortes P & Ortin J 1993 *Journal of General Virology* **74**, 1317–1326.

Varsally W & Brogna S 2012 *Biochemical Society transactions* **40**(4), 778–83.

Vattem K M & Wek R C 2004 *Proceedings of the National Academy of Sciences of the United States of America* **101**(31), 11269–74.

Vilella A J, Severin J, Ureta-Vidal A, Heng L, Durbin R & Birney E 2009 *Genome research* **19**(2), 327–35.

Vlasova I A, Tahoe N M, Fan D, Larsson O, Rattenbacher B, Sternjohn J R, Vasdewani J, Karypis G, Reilly C S, Bitterman P B & Bohjanen P R 2008 *Mol Cell* **29**(2), 263–70.

Wang D, Wengrod J & Gardner L B 2011 *The Journal of biological chemistry* **286**(46), 40038–43.

Wang D, Zavadil J, Martin L, Parisi F, Friedman E, Levy D, Harding H, Ron D & Gardner L B 2011 *Molecular and cellular biology* **31**(17), 3670–80.

Wang E T, Cody N A, Jog S, Biancolella M, Wang T T, Treacy D J, Luo S, Schroth G P, Housman D E, Reddy S, Lecuyer E & Burge C B 2012 *Cell* **150**(4), 710–24.

Wang E T, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore S F, Schroth G P & Burge C B 2008 *Nature* **456**(7221), 470–6.

Wang G S, Kuyumcu-Martinez M N, Sarma S, Mathur N, Wehrens X H & Cooper T A 2009 *J Clin Invest* **119**(12), 3797–806.

Wang-Gillam A, Pastuszak I & Elbein A 1998 *The Journal of biological chemistry* **273**, 27055–27057.

Wang J H, Gostissa M, Yan C T, Goff P, Hickernell T, Hansen E, Difilippantonio S, Wesemann D R, Zarrin A A, Rajewsky K, Nussenzweig A & Alt F W 2009 *Nature* **460**(7252), 231–6.

Wang R H, Sengupta K, Li C, Kim H S, Cao L, Xiao C, Kim S, Xu X, Zheng Y, Chilton B, Jia R, Zheng Z M, Appella E, Wang X W, Ried T & Deng C X 2008 *Cancer cell* **14**(4), 312–23.

Wang Y, Ma M, Xiao X S & Wang Z F 2012 *Nature Structural & Molecular Biology* **19**(10), 1044–U104.

Wang Y, Xiao X S, Zhang J M, Choudhury R, Robertson A, Li K, Ma M, Burge C B & Wang Z F 2013 *Nature Structural & Molecular Biology* **20**(1), 36–U54.

Wang Z F, Rolish M E, Yeo G, Tung V, Mawson M & Burge C B 2004 *Cell* **119**(6), 831–845.

Warf M B & Berglund J A 2007 *RNA* **13**(12), 2238–51.

Weischenfeldt J, Damgaard I, Bryder D, Theilgaard-Monch K, Thoren L A, Nielsen F C, Jacobsen S E, Nerlov C & Porse B T 2008 *Genes & development* **22**(10), 1381–96.

Weischenfeldt J, Waage J, Tian G, Zhao J, Damgaard I, Jakobsen J S, Kristiansen K, Krogh A, Wang J & Porse B T 2012 *Genome biology* **13**(5), R35.

Wen J & Brogna S 2008 *Biochemical Society transactions* **36**(Pt 3), 514–6.

Wittkopp N, Huntzinger E, Weiler C, Sauliere J, Schmidt S, Sonawane M & Izaurralde E 2009 *Molecular and cellular biology* **29**(13), 3517–28.

Wu J Y & Maniatis T 1993 *Cell* **75**(6), 1061–1070.

Yamashita A, Ohnishi T, Kashima I, Taya Y & Ohno S 2001 *Genes & development* **15**(17), 2215–28.

Yeo G W, Coufal N G, Liang T Y, Peng G E, Fu X D & Gage F H 2009 *Nat Struct Mol Biol* **16**(2), 130–7.

Yeo G W, Van Nostrand E L & Liang T Y 2007 *Plos Genetics* **3**(5), 814–829.

Yepiskoposyan H, Aeschimann F, Nilsson D, Okoniewski M & Muhlemann O 2011 *RNA* **17**(12), 2108–18.

Young R A 2011 *Cell* **144**(6), 940–54.

Zahler A M, Damgaard C K, Kjems J & Caputi M 2004 *Journal of Biological Chemistry* **279**(11), 10077–10084.

Zahler A M, Lane W S, Stolk J A & Roth M B 1992 *Genes & Development* **6**(5), 837–847.

Zahler A M, Neugebauer K M, Stolk J A & Roth M B 1993 *Molecular and Cellular Biology* **13**(7), 4023–4028.

Zeng S W, Yong K T, Roy I, Dinh X Q, Yu X & Luan F 2011 *Plasmonics* **6**(3), 491–506.

Zhang X, Zhang J, Wang T, Esteban M A & Pei D 2008 *The Journal of biological chemistry* **283**(51), 35825–33.

Zhang Z & Krainer A R 2004 *Molecular cell* **16**(4), 597–607.

Zhao C, Datta S, Mandal P, Xu S & Hamilton T 2010 *The Journal of biological chemistry* **285**(12), 8552–62.

Zykovich A, Korf I & Segal D J 2009 *Nucleic Acids Res* **37**(22), e151.