UNIVERSITY OF CALIFORNIA, SAN DIEGO

Comprehensive discovery and analysis of RNA binding protein-dependent post-transcriptional events in mammalian systems

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

in

Bioinformatics and Systems Biology

by

Stephanie C. Huelga

Committee in charge:

Professor Gene Yeo, Chair
Professor Christopher Woelk, Co-Chair
Professor Jens Lykke Andersen
Professor Kelly Frazer
Professor Bing Ren

2014

The Dissertation of Stephanie C. Huelga is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2014

Dedication

For Gene Yeo, you've been an incredible mentor.

I hope I've made you proud.

Epigraph

It always seems impossible until it’s done.

Nelson Mandela

Table of Contents

List of Figures

List of Tables

Acknowledgements

Along my graduate school journey I have encountered many people whom I owe many thanks. I owe the biggest thank you to my mentor and dissertation chair, Gene Yeo. Gene's energy and passion for science has inspired me to always work "hard and fast." He has always challenged me, trusted in me, and given me the freedom to grow. Gene has provided me with many fantastic opportunities and invaluable advice when I needed it. I would not be the scientist I am today without his guidance, and for that I am so grateful. In addition to Gene, I would like to thank my entire dissertation committee for their support and critique: Christopher Woelk (Co-Chair), Bing Ren, Kelly Frazer and Jens Lykke-Andersen.

There are also a few professors, who may not know, but have greatly influenced the decisions I've made in the early years of my graduate career. These professors have greatly influenced my graduate path: Dr. John Wooley for his initial guidance on rotations; Dr. James Halpert and Dr. Phil Bourne for providing me with my first UCSD research experience and proving to me that I am meant to be in graduate school; and Dr. Steve Briggs for supporting and encouraging my exploration of teaching. I'd also like to thank my students for showing me that I can be a great teacher, but for also sparking the realization in me that teaching at the college level is not for me.

I would also like to thank my friends who I was fortunate to meet throughout graduate school. Thanks to my UCSD AGEP Competitive Edge "family" for being my first friends in graduate school and providing me with constant grad-student life support, especially Isabel Canto. To the people in my Bioinformatics 2008 cohort who have experienced the same struggles and coped together, specifically Josh Lerman and Boyko Kakaradov. To my closest labmates and friends who have always kept my head on straight, accompanied me to lunch and coffee breaks, and provided me with constant

entertainment and laughs: Dr. Thomas Stark, Dr. Missy Wilbert, and Dr. Kasey Hutt. Also, Briana Boyd for being a good listener and great friend when grad school life got really tough.

Most importantly, I have to thank my family for their love and blind support of my crazy continuation of schooling. My sister Ashley and brother-in-law Brandon, my mom Dana and my dad Fernando for always being impressed with my seemingly strange accomplishments and for accepting that I am the smartest person in the family. To my Granny, my Mumzie, my Papa and my Nana for always making solid efforts to read my latest publications. And my boyfriend Brian for being supportive and always trying his best to speak my language: “RNA” and “transcription”. I could not have survived the past 6 years without all these people, whom were each crucial for my PhD success.

Finally, I must thank the many talented scientists from my own and other labs that I was able to work closely with on research projects, and eventually manuscripts. The invaluable interactions with these people have helped me grow immensely as a scientist.

Chapter 1 is, in part, published as a chapter entitled “Alternative Splicing in Stem Cells” in the electronic book *Computational Biology of Embryonic Stem Cells* (Huelga and Yeo, 2012). This dissertation author was the primary author who reviewed the literature and wrote text.

The Lin28 work described in Chapter 2 has been published as part of a manuscript entitled “LIN28 Binds Messenger RNAs at GGAGA Motifs and Regulates Splicing Factor Abundance” in *Molecular Cell* (Wilbert et al., 2012). This dissertation author was the secondary author who helped to write the manuscript, analyze the splicing array data, and direct experiments. The TDP-43 work described in Chapter 2 has been published as part of a manuscript entitled “ALS-linked TDP-43 mutations

produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43" in *PNAS* (Arnold et al., 2013). This dissertation author was a secondary author who helped to write the manuscript, analyze the splicing array data, and direct experiments.

Chapter 3 is in whole published as a manuscript entitled "Integrative Genome-wide Analysis Reveals Cooperative Regulation of Alternative Splicing by hnRNP Proteins" in *Cell Reports* (Huelga et al., 2012). This dissertation author was the primary author who wrote the manuscript, directed experiments, and analyzed the data.

Chapter 4 is in preparation to be submitted for publication in *Cell*. This dissertation author was the primary researcher/author who designed and directed the experiments, analyzed the data, and wrote the text. Co-authors who have also contributed to this work include Kristopher Brannan (Yeo Lab, UC San Diego), Mike Washburn (Stowers Institute), and Danette Daniels (Promega).

Chapter 5 is in preparation to be submitted for publication in *Nature*. This dissertation author was the primary researcher/author who designed and directed the experiments, analyzed the data, and wrote the text. Co-authors who have also contributed to this work include Eric Van Nostrand (Yeo Lab, UC San Diego), Patrick Liu (Yeo Lab, UC San Diego), and Shawn Hoon (A*STAR, Singapore).

Chapter 6 is, in part, published as a chapter entitled "Alternative Splicing in Stem Cells" in the electronic book *Computational Biology of Embryonic Stem Cells* (Huelga and Yeo, 2012). This dissertation author was the primary author who reviewed the literature and wrote text.

Vita

Education

2014 Doctor of Philosophy, Bioinformatics and Systems Biology
University of California, San Diego

2008 Bachelor of Science, Bioinformatics
University of California, Santa Cruz

Professional Experience

2007-2008 Biochemistry Research Assistant, Rubin Lab
University of California, Santa Cruz

2007 Crystallography Intern
Roche Pharmaceuticals, Palo Alto, CA

2007 Computational Genomics Research Assistant, Ares Lab
University of California, Santa Cruz

Teaching Experience

2011-2013 UCSC Genome Browser Tutorial Instructor
University of California, San Diego

2010-2011 Teaching Assistant, Department of Biology, Bioinformatics Lab
University of California, San Diego

Awards and Affiliations

2012-2013 ARCs Scholar

2010-2013 NSF Graduate Research Fellow

2010-2012 RNA Society Member

2009-2013 Graduates United in the Interests of Diversity and Excellence (GUIDE)
Member and Mentor

2009-2010 Cancer Cell Biology Training Grant Recipient

Publications

1. Arnold, E.S.*, Ling, S.C.*, **Huelga, S.C.**, Lagier-Tourenne, C., Polymenidou, M., Ditsworth, D., Kordasiewicz, H.B., McAlonis-Downes, M., Platoshyn, O., Parone, P.A., et al. (2013). ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. ***PNAS*** 110, E736-745. (*authors contributed equally)

2. Lagier-Tourenne, C.*, Polymenidou, M.*, Hutt, K.R.*, Vu, A.Q., Baughn, M., **Huelga, S.C.**, Clutario, K.M., Ling, S.C., Liang, T.Y., Mazur, C., et al. (2012). Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. ***Nature neuroscience*** 15, 1488-1497. (*authors contributed equally)

3. Wilbert, M.L., **Huelga, S.C.**, Kapeli, K., Stark, T.J., Liang, T.Y., Chen, S.X., Yan, B.Y., Nathanson, J.L., Hutt, K.R., Lovci, M.T., et al. (2012). LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. ***Molecular cell*** 48, 195-206.

4. **Huelga, S.C.**, and Yeo, G.W. (2012). Alternative Splicing in Stem Cells. In ***Computational Biology of Embryonic Stem Cells***, M. Zhan, ed. (Bentham Science Publishers), pp. 147-160.

5. **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., et al. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. ***Cell Reports*** 1, 167-178.

6. Polymenidou, M.*, Lagier-Tourenne, C.*, Hutt, K.R.*, **Huelga, S.C.**, Moran, J., Liang, T.Y., Ling, S.C., Sun, E., Wancewicz, E., Mazur, C., et al. (2011). Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. ***Nature neuroscience*** 14, 459-468. (*authors contributed equally)

7. Gay, S.C., Shah, M.B., Talakad, J.C., Maekawa, K., Roberts, A.G., Wilderman, P.R., Sun, L., Yang, J.Y., **Huelga, S.C.**, Hong, W.X., et al. (2010). Crystal structure of a cytochrome P450 2B6 genetic variant in complex with the inhibitor 4-(4-chlorophenyl)imidazole at 2.0-A resolution. ***Mol Pharmacol*** 77, 529-538.

Abstract of the Dissertation

Comprehensive discovery and analysis of RNA binding protein-dependent post-transcriptional events in mammalian systems

by

Stephanie C. Huelga

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2014

Professor Gene Yeo, Chair
Professor Christopher Woelk, Co-Chair

Nascent transcripts produced by RNA polymerase II in eukaryotic cells are subject to extensive processing prior to the generation of a functional messenger RNA (mRNA). These RNAs are generally found coated with RNA binding proteins (RBPs), which act in concert to regulate RNA processing events, including splicing, polyadenylation, and RNA stability. The aggregate effect of various RBPs on a given RNA transcript eventually dictates its fate, a phenomenon referred to as the RNP code. The precise control of RNA processing by RBPs is incredibly important for cellular homeostasis, defects of which lead to numerous accounts of human genetic diseases in

many tissues. With the emergence of genome-wide methods for detecting direct binding of RBPs on target RNAs, coupled with technologies to measure changes in various aspects of RNA processing, global rules and insights for individual RBPs have been revealed. However, few studies have combined regulatory changes for more than one RBP to better understand their combinatorial effects on RNA targets. Additionally, despite the importance of RBPs, we still do not have the complete repertoire of which proteins bind to RNA, and which of these bind simultaneously with other RBPs to constitute the “RBP-RNA interactome”. Here, I conduct genome-wide RNA processing analyses in mammalian cells, integrating regulation and binding information for multiple RBPs, including disease-related RBPs, the highly abundant heterogeneous nuclear ribonucleoparticle (hnRNP) proteins, and many previously uncharacterized RBPs. This battery of computational and experimental assays provides insight into the unique roles of hundreds of individual RBPs, as well as the extent of coordinated regulation between RBPs. I also describe a systematic approach to identifying the proteins that interact with RNA simultaneously with hnRNP proteins. Not limited to only mRNA-bound proteins, my strategy identifies thousands of hnRNP protein interactors, including putative novel proteins that interact with mRNA and pre-mRNA, and have biochemical and statistical attributes of known RBPs. My findings expand the repertoire of RNA-interacting RBPs and provide a resource for the study of human simultaneous RBP-RBP interactomes. This comprehensive analysis of RBPs investigates their specific roles in the regulation of RNA processing yielding interesting findings for the RNA biology field and insights into how misregulation can impact human disease.

1 Introduction

1.1 Post-transcriptional gene regulation

The central dogma of molecular biology simply states that the genome, DNA, is transcribed into the transcriptome, RNA, and then translated into the proteome, proteins. However, the regulation of gene expression, from DNA to RNA and then to protein, has been shown in recent years to be far more complicated than previously anticipated, requiring a complex network of regulation. Human gene expression is finely tuned for cell-, tissue-, and condition-specific functions via a variety of co- and post-transcriptional processes.

In eukaryotes, premature messenger RNA (pre-mRNA) molecules undergo several steps of post-transcriptional processing before generating final mature RNA (mRNA) transcripts. Transcribed RNA molecules are capped, poly-adenylated, edited, spliced, and localized to various cellular compartments. Depending on the regulatory decisions made, multiple isoforms of a single RNA molecule can be produced, each yielding different proteins with unique cellular functions. Ultimately, all mRNA molecules undergo quality control to be translated into protein or degraded and destroyed. Importantly, RNA transcripts are never naked, and throughout all steps of RNA processing RNAs are in complex with RNA binding proteins (RBPs), the combination and organization of which dictate the final fate of the RNA (Figure 1.1).

1.1.1 Alternative splicing

Alternative splicing is a major mechanism by which the complexity of the proteome is enhanced through the selective excision of non-coding intronic regions of a pre-mRNA and juxtaposition of exonic regions in different linear arrangements. From a

single pre-mRNA transcript, alternative splicing can generate a plethora of mRNA isoforms through varied use of exons (Figure 1.1). The large, RNA-protein complex that is essential for pre-mRNA splicing is called the “spliceosome.” This complex is made up of five small nuclear ribonucleoprotein particles (snRNPs), U1, U2, U4, U5, and U6, as well as auxiliary factors that cooperate to recognize splice sites and catalyze the splicing reaction (Figure 1.2). Interactions between the spliceosomal machinery, *cis*-regulatory elements and *trans*-regulatory splicing factors define a set of rules for accurate splice site selection. While this “splicing code” that dictates splice site selection is still being deciphered, some important splicing regulatory elements (SREs) have been identified. Exonic and intronic splicing enhancer (ESE, ISE) and splicing silencer (ESS, ISS) motifs act to enhance or inhibit adjacent splice site recognition. Trans-acting splicing factors such as hnRNP proteins, serine-arginine (SR) proteins, and other proteins that bind RNA are recruited to these elements to enhance or repress splicing (Martinez-Contreras et al., 2007; Mayeda et al., 1993; Wang and Burge, 2008). As these splicing factors are differentially expressed in various cell types, their combinatorial interaction regulates the expressed, cell-type specific alternative splicing patterns. Differential expression of these splicing factors, amongst others, allow for alternative splicing to be precisely regulated in various cells/cell states.

1.2 RNA binding proteins (RBPs)

Similar to histones on DNA, RNA binding proteins (RBPs) are arranged on pre-mRNAs and mRNAs as “beads on a string.” This property, first visible in cells in the 1970s (McKnight and Miller, 1976), intrigued researchers and led to the initial discovery of the most abundant set of RBPs in the 1980s, the heterogeneous nuclear ribonucleoparticle (hnRNP) proteins (Dreyfuss et al., 1984). HnRNP-RNA complexes contain at least 20 abundant and ubiquitously expressed hnRNP proteins named hnRNP

A1 through hnRNP U, with molecular weights ranging from 34 to 120 kDa, and a modular RNA-binding protein domain (RBD) structure (Dreyfuss et al., 2002; Dreyfuss et al., 1993) (Figure 1.4). Although the hnRNP family of RBPs were first discovered because of their abundance, other RBPs were quickly identified due to their multifunctional participation in all aspects of RNA processing in eukaryotic cells (Han et al., 2010).

RNA binding proteins are each composed of unique combinations of a limited set of RBDs. The most commonly studied RNA binding proteins contain canonical RNA binding domains that recognize RNA, such as the RNA recognition motif (RRM), the zinc finger, the K homology (KH) domain, or helicase domains. Varied arrangements of these domains allow RBPs to bind unstructured or structured RNAs that are single- or double-stranded. Some RNA binding proteins recognize specific *cis*-regulatory RNA elements or sequence motifs, found all across a gene, in the 5' and 3' untranslated regions (UTRs), introns and exons, while others recognize structural motifs or bind in a non-specific manner. Initial estimates considering proteins containing known RNA binding domains predict that close to 600 RBPs exist in the human genome. However recent genome-wide approaches have revealed putative RBPs with non-canonical or uncharacterized RNA interactions, suggest that there are likely many more RNA binding proteins than previously thought.

The proper control of RNA processing by RBPs is essential for cellular development and homeostasis. Consequently, misregulation of RNA processing events such as splicing has been shown to cause several human genetic diseases (Lukong et al., 2008) and cancer pathologies (Gardina et al., 2006; Lapuk et al., 2010; Misquitta-Ali et al., 2011). Despite their importance, the precise roles of each RBP have yet to be determined. There remain many open questions related to RBPs and the RNAs they

target. Namely, which RNAs are targets and how are they regulated? How do multiple RBPs coordinate the regulation of their target RNAs? And are there functional groups of RBPs? The identification of RBP-regulated RNA processing events, as well as understanding the rules that direct RBP regulation will substantially improve our understanding of RNA binding proteins in general RNA biology and human disease.

1.3 The growth of genome-wide methods for analysis of alternative splicing and regulatory RBPs

In recent years genome-wide approaches and computational approaches have been created to study alternative splicing and the various *cis*- and *trans*-factors that regulate this process. Prior to these global methods, discovering new alternative splicing events was limited to candidate genes. The availability of microarray and sequencing technologies have remarkably increased our throughput of identifying new alternative splicing events and providing systematic readouts of their biological implication.

Microarrays specifically designed for studying alternative splicing contain exonic oligonucleotide probes and probes spanning two annotated exons that may be spliced together in a specific transcript variant (Johnson et al., 2003). These “exon-junction arrays” can then be probed with cDNA (complimentary DNA) to measure the degree of alternative splicing (Figure 1.5A). This platform is a useful tool for uncovering tissue-specific splicing patterns (Pan et al., 2004), and has also revealed many RBP-dependent splicing events via comparison of cDNA from wildtype and splicing factor depleted cells (Blanchette et al., 2005).

While useful for studying splicing at known exon coordinates and junctions, exon-junction arrays are limited in their ability to detect novel splice sites and are thus biased in alternative splicing detection. Whole genome tiling arrays with overlapping or non-overlapping oligonucleotide probes that cover entire genomic regions of interest

circumvent these issues (Figure 1.5A). Unlike exon-junction arrays, these arrays can detect altered splice site usage, retained introns, and unannotated exons (Zhang et al., 2007). A cost-effective alternative to genome tiling arrays, which exclude intronic and intergenic regions, are exon-tiling arrays (Gardina et al., 2006) (Figure 1.5A).

Profiling methods reliant on microarrays have been developed to identify transcripts bound by specific splicing factors. Ribonucleoprotein (RNP) immunoprecipitation followed by microarray application (RIP-Chip) uses an antibody against a regulatory RBP of interest to pull down an endogenous RNP complex from cell extracts. The bound RNA is extracted from the eluate, purified, and identified by microarray analysis (Keene et al., 2006). This method has been successful in revealing sets of mRNA targets that are regulated by similar factors.

While useful for alternative splicing studies of annotated genes, various microarray technologies are limited by high background, cross-hybridization, low sensitivity for lowly expressed transcript variants, physical space, and reliance on previous gene and exon annotations. Tag-based profiling, such as serial analysis of gene expression (SAGE) (Velculescu et al., 1995), cap analysis of gene expression (CAGE) (Shiraki et al., 2003), and polony multiplex analysis of gene expression (PMAGE) (Kim et al., 2007) address these limitations by detecting lowly expressed and novel gene transcripts. However, these methods are also incredibly time-consuming and expensive.

More recently developed high-throughput sequencing technologies are being used as an alternative to microarrays due to the relatively low-background, high sensitivity, and lack of dependence on gene and exon annotations. Sequencing technologies can produce gigabytes of nucleotide resolution data through massively parallel sequencing within a few days and for a fraction of the cost compared to

conventional sequencing methods. Whole transcriptome sequencing, or RNA-seq, has become a standard technique for qualitative and quantitative identification of novel transcripts and splice variant detection more reliably than microarray technologies (Pan et al., 2008). RNA-seq involves sequencing of cDNA reverse transcribed from total RNA or poly-A selected mRNA of a cell population (Figure 1.5B). Wang et al. were among the first to demonstrate the power of this technique as they surveyed the transcriptome across ten diverse human tissues, revealing many tissue-specific alternative splicing events (Wang et al., 2008).

High-throughput sequencing based methods have also been developed to identify genome-wide RBP-RNA targets and their general mechanism of action in regulating splicing and other RNA processing events. The Darnell laboratory developed a method called cross-linking immunoprecipitation (CLIP), which exposes the cells to UV crosslinking to irreversibly cement physical RNA-protein interaction (Ule et al., 2003). We, and others, have optimized the CLIP method to be coupled with high-throughput sequencing of the bound RNA to allow for the large-scale identification of endogenous RBP targets *in vivo* (Licatalosi et al., 2008; Sanford et al., 2008; Sanford et al., 2009; Yeo et al., 2009) (Figure 1.5C). This high-throughput experimental approach, referred to as CLIP-seq, or HITS-CLIP, begins with the stabilization of *in vivo* protein-RNA interactions with UV irradiation. This is followed by immunoprecipitation of the protein of interest with a specific antibody, RNA trimming by a nuclease, application of a 3' linker and 5' radio-label, and isolation of the protein-mRNA complex by SDS-PAGE. Short RNA sequences 50-100 nucleotides long protected from nuclease digestion by the crosslinked protein of interest are then amplified by RT-PCR and sequenced on high throughput sequencing instruments.

This technique allows for a genome-wide analysis of targets for a specific RBP in the genome. Analysis of these targets can reveal sequence-level understanding of the specific roles of RBPs in splicing regulation. CLIP-seq has successfully provided RNA-maps of targets, and delineating alternative splicing events regulated by hundreds of RBPs. RNA-maps are a useful tool in the elucidation of the splicing code. A handful of RBPs have been strongly linked to alternative splicing, and continued efforts are rapidly identifying the other functional elements of this highly regulated process.

Additionally, these methods to identify the specific genome-wide targets of RBPs continue to be developed and improved. Variations of this technique, termed PAR-CLIP (Hafner et al., 2010) and iCLIP (Konig et al., 2010), claim to more specifically amplify regions closer to the RBP-RNA crosslink site. As these high-throughput sequencing methods become more accessible through the generation of kits and analysis pipelines, global targets for more and more RBPs are being mapped, helping to unravel the rules for RBP-regulated RNA processing.

1.4 The evolution of bioinformatic analysis for detecting alternative splicing and splicing regulatory elements

The vast amounts of data generated by high-throughput sequencing necessitate efficient computational approaches to analyze and interpret the data. Traditionally, alternative splicing events and splicing regulatory elements (SREs) within the transcriptome are studied using computational analyses of publicly available expressed sequence tag (EST) databases. However, ESTs are limited by their 3' end bias, their disproportionate weight of well-studied tissues and conditions, and their insensitivity among paralogs. SELEX (systematic evolution of ligands by exponential enrichment) is an experimental method that is often complimented with *in silico* approaches for indentifying SREs (Lorenz et al., 2006) and has been successful in identifying ESEs for

SR proteins (Cavaloc et al., 1999; Tacke and Manley, 1995). Bioinformatic analysis of sequence data for transcripts identified by microarray and sequencing experiments will be vital for detecting alternative splicing events and unraveling the SREs that mediate those events.

The utility of microarray data relies on computational post-processing to remove background and quantify differential expression. Algorithms such as SPLICE (Hu et al., 2001), and ANOSVA (Cline et al., 2005) have been developed to detect differentially spliced exons from sets of microarray data. These techniques rely on statistical tests of significance and have proven to be accurate and useful. However, they are not optimized for novel splice variant detection, nor can they accurately quantify expression of isoforms from a single pre-mRNA transcript due to low sensitivity of the statistics and inherent microarray limitations (Cuperlovic-Culf et al., 2006). In addition to locating differentially expressed exons, global bioinformatic approaches are necessary to examine the neighboring sequences of these exons for discovery of SREs. In our microarray-based study of alternative splicing, we uncovered a conserved 'GCAUG' *RBFOX1/2* binding motif near splice sites for exons with differential expression in neuronal differentiation (Yeo et al., 2007).

More recently, microarray technology has been overpowered by the advantages of high-throughput sequencing data, and thus many have focused their attention toward generating algorithms to analyze this new source of data. In order to render this wealth of new sequencing data useful, the resultant reads must first be pre-processed to remove PCR and sequencing artifacts and then mapped to a reference genome. Due to repetitive regions of the genome, and the small size and error rate of the reads output by high-throughput sequencing devices (~35-80 nucleotides), uniquely mapping the reads to known genes has become a challenge. The large structural variations caused by

alternative splicing events and other post-transcriptional modifications enormously complicate the sequence mapping problem, since reads are capable of mapping across novel and annotated exon-exon junctions.

Many tools are rapidly being developed to approach this bottleneck of accurately mapping short reads to the transcriptome. Tools such as BLAT (Kent, 2002), Bowtie (Langmead et al., 2009), Maq (Li et al., 2008b), ELAND (Illumina), and BFAST (Homer et al., 2009) use different algorithms to quickly align millions of reads to a reference genome. Algorithms allow for some mismatches, and have the option to consider the sequence quality values of each read for improved mapping. These techniques are useful but often leave many reads unmapped (Wang et al., 2008).

Mapping approaches specific for alternative splicing detection have also been developed. Since sequenced reads come from the highly-processed transcriptome, many reads do not map directly to the reference genome because many reads do not fall entirely within exons or introns, and instead span exon-exon junctions. Reads that do not map directly to the genome can be mapped using standard mapping algorithms to custom-generated exon-exon junction databases (Cloonan et al., 2008; Li et al., 2008a; Wang et al., 2008; Wang et al., 2010). In most cases, these approaches are known to still leave one-third of the reads unmappable (Wang et al., 2008). This may be due to non-unique repetitive regions in the genome, sequencing errors, or reads spanning novel splice junctions. More recent mapping tools, such as TOPHAT (Trapnell et al., 2009), GSNAP (Wu and Nacu, 2010), and RNA-STAR (Dobin et al., 2013) attempt to tackle the alternative splicing mapping problem directly, by mapping reads to known and novel splice junctions defined significant within the data. Once mapped across splice junctions other recently developed tools can be used to quantify exon-junction reads and predict alternatively spliced exons. One such algorithm, MISO (Katz et al., 2010), uses a

Bayesian inference model to identify isoform differences, including differential cassette exons, and differential end usage between two RNA-seq datasets, called “tandem 3′UTRs”**.** Continual improvement of these mapping tools will be necessary for accurate analysis of alternative splicing within high-throughput sequencing data.

Part of the power of algorithmic approaches is their ability to identify sequence patterns such as those due to conservation, evolution, and enrichment. Computational approaches have been surprisingly successful in defining *cis*-regulatory elements within the genome (Wang and Burge, 2008). For example, RESCUE-ESE (Relative Enhancer and Silencer Classification by Unanimous Enrichment of Exon Splicing Enhancers) uses statistical analysis of exon-intron and splice site composition to predict the ESE activity of sequences (Fairbrother et al., 2002). Another statistical method, called CoCOA (orthogonalized co-occurrence analysis) identifies enriched, co-occurring oligonucleotide sequence pairs in exon flanking regions that may cooperate during splicing (Friedman et al., 2008). With enough knowledge of key functional elements, purely computational methods can be accurately applied to predict alternative splicing events.

1.5 Introduction publication acknowledgment

Chapter 1 is, in part, published as a chapter entitled “Alternative Splicing in Stem Cells” in the electronic book *Computational Biology of Embryonic Stem Cells* (Huelga and Yeo, 2012). This dissertation author was the primary author who reviewed the literature and wrote text.

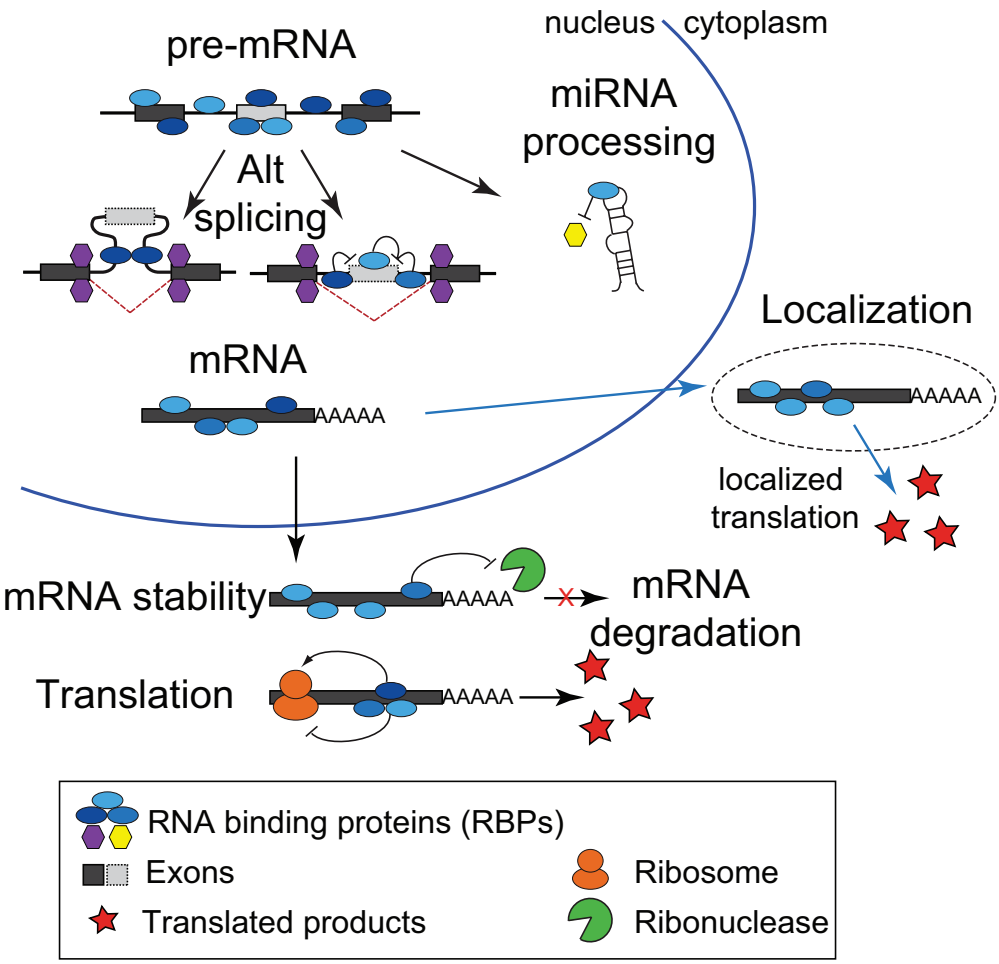


Figure 1.1: RNA binding proteins affect all RNA processing steps

RBPs coat pre-mRNA and mRNA and are thus involved in many RNA processing steps including alternative splicing, microRNA (miRNA) processing, mRNA stability, mRNA degradation, mRNA localization, and translation.

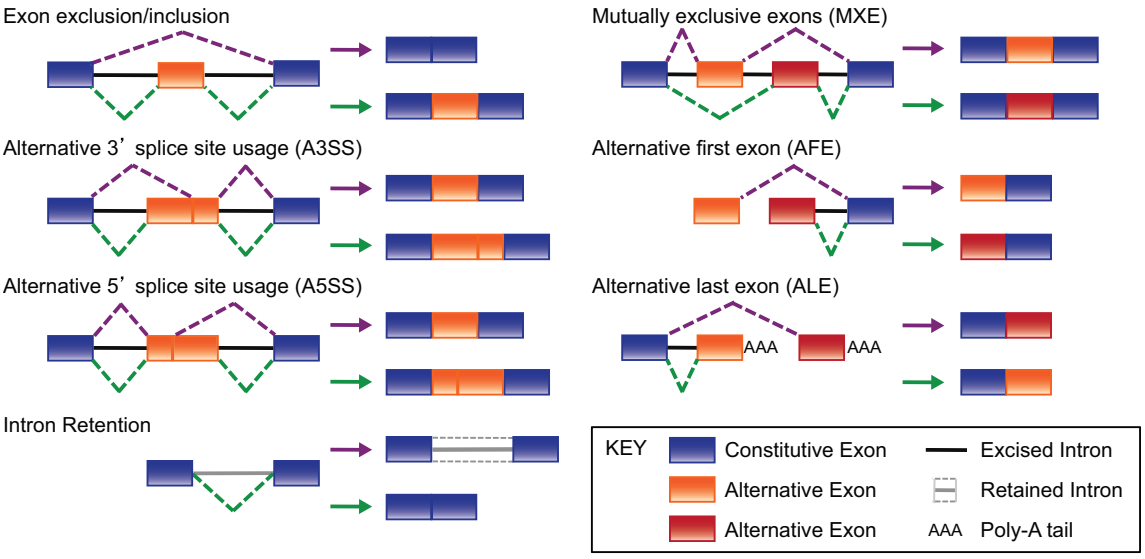


Figure 1.2: Alternative splicing produces multiple mRNA isoforms from a single pre-mRNA transcript

The events depicted include: exon exclusion/inclusion, alternative 3' splice site usage (A3SS), alternative 5' splice site (A5SS), intron retention, mutually exclusive exons (MXE), alternative first exon (AFE), and alternative last exon (ALE).

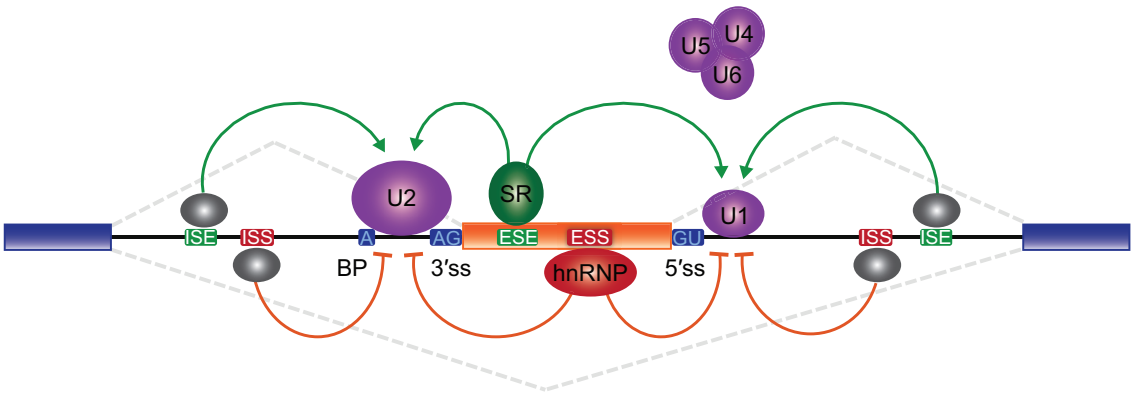


Figure 1.3: The splicesome complex and regulatory splicing factors surrounding an alternatively spliced exon

Spliceosomal assembly is initiated by the recognition of the 5' GU by the U1 snRNP. The U2 snRNP is then recruited to the branch-point (BP) in an ATP-dependent manner. Subsequent recruitment of the U4/U6-U5 tri-snRNP is then followed by a series of structural rearrangements leading to the formation of a catalytically active spliceosome complex to remove the alternative exon (orange) and introns (black). Other auxiliary factors (gray) bind to cis-regulatory elements, such as SR proteins (green) and the hnRNP proteins (red) that bind to ESE and ESS to enhance or repress splicing, respectively.

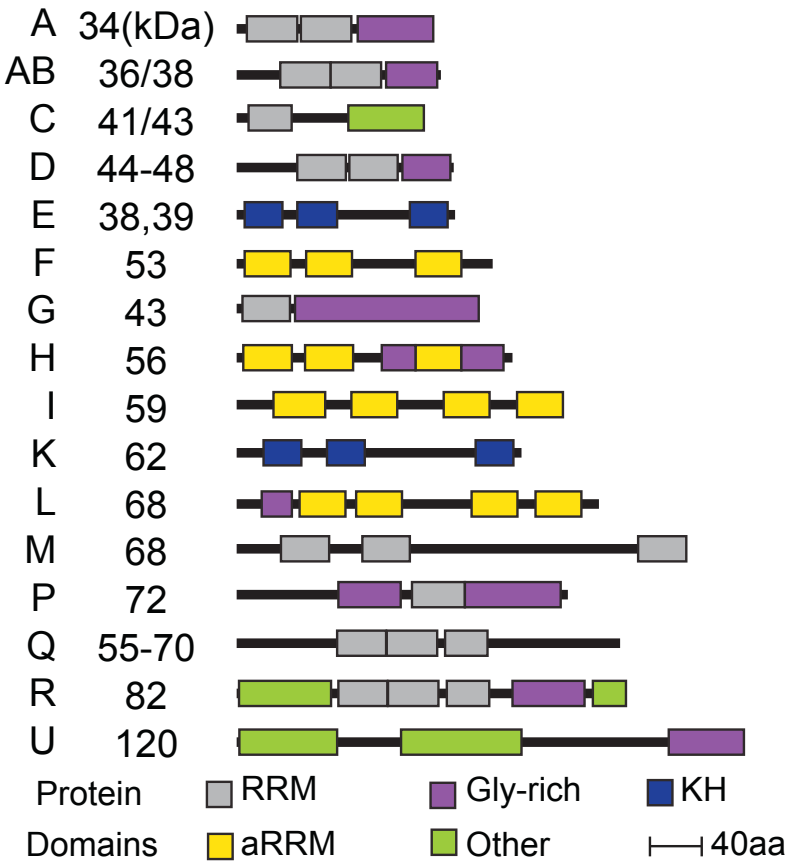


Figure 1.4: HnRNP protein family domain structure

HnRNP proteins are a structurally diverse family of RNA binding proteins containing various organizations of classical RNA binding protein domains. The domains listed are RNA recognition motif (RRM), Glycine-rich (Gly-rich), hnRNPK homology (KH), atypical RRM (aRRM), and other.

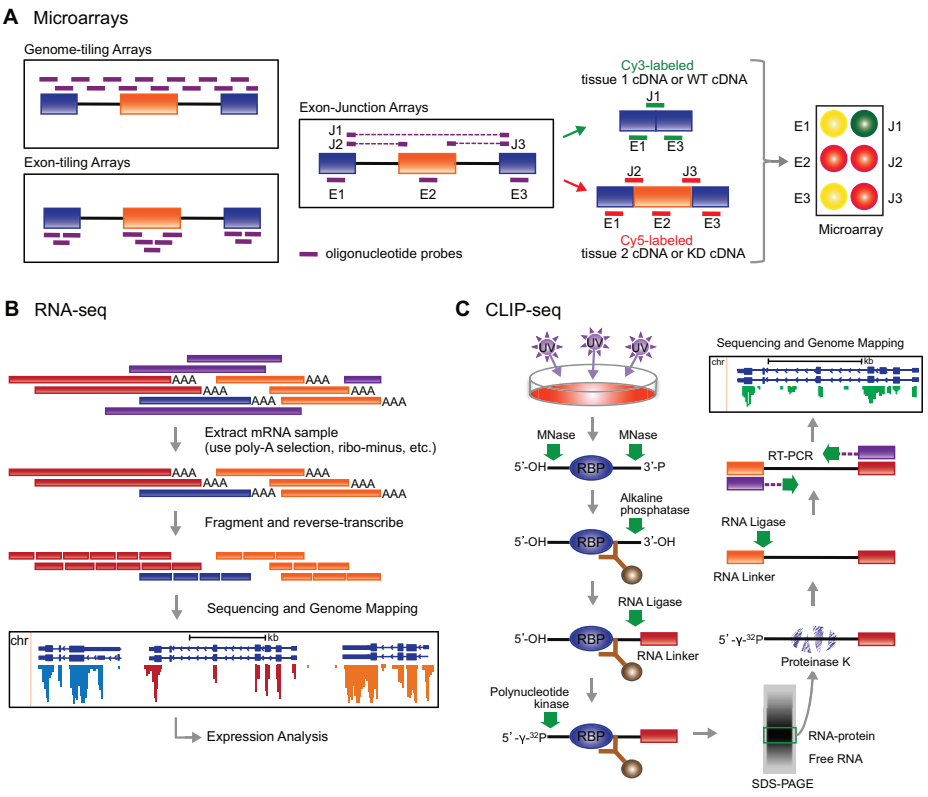


Figure 1.5: A summary of genome-wide approaches for alternative splicing analysis

A. Microarrays with various oligonucleotide probe types have been applied to study alternative splicing. Whole-genome tiling arrays contain overlapping or non-overlapping probes that scan entire genomic regions. Exon-tiling, or all-exon arrays, contain overlapping or non-overlapping probes within each known or predicted exon of the genome. These tiling arrays are often assayed with biotin-labeled cDNA and then incubated with fluorescently labeled streptavidin for detection. Exon-junction microarrays contain probes for exons (E1, E2, E3) and putative exon-junctions (J1, J2, J3) within a transcript. Cy3 (green) and Cy5 (red) labeling of cDNA from two different samples allows for comparison of expressed transcript variants. Green or red spots on the microarray represent expression exclusively in the Cy3-labeled or Cy5-labeled samples, respectively, while yellow spots on the microarray represent equivalent expression between samples. **B.** The RNA-seq protocol begins with total RNA extraction from a cellular condition or tissue type. The RNA sample is enriched for mRNA with poly-A selection using poly-T oligos and ribo-minus selection to remove contaminating ribosomal RNA, and other selection methods. The mRNA is then fragmented and reverse transcribed. Lastly, the library is amplified, sequenced and mapped to the genome. **C.** The CLIP-seq protocol begins with UV-crosslinking of RBPs to bound RNA targets and the RBP-RNA complex is enriched using a specific antibody. Trimming with MNase enables nucleotide level resolution as the RNA bound by the protein is protected from digestion. After size selection, the protein is removed from its bound RNA and the RNA is amplified by RT-PCR, sequenced and mapped to the genome to determine significant regions of protein binding for the RBP.

1.6 References

Blanchette, M., Green, R.E., Brenner, S.E., and Rio, D.C. (2005). Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila. Genes Dev *19*, 1306-1314.

Cavaloc, Y., Bourgeois, C.F., Kister, L., and Stevenin, J. (1999). The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. RNA *5*, 468-483.

Cline, M.S., Blume, J., Cawley, S., Clark, T.A., Hu, J.S., Lu, G., Salomonis, N., Wang, H., and Williams, A. (2005). ANOSVA: a statistical method for detecting splice variation from expression data. Bioinformatics *21 Suppl 1*, i107-115.

Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., Robertson, A.J., Perkins, A.C., Bruce, S.J., Lee, C.C., Ranade, S.S., Peckham, H.E., Manning, J.M., McKernan, K.J., and Grimmond, S.M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods *5*, 613-619.

Cuperlovic-Culf, M., Belacel, N., Culf, A.S., and Ouellette, R.J. (2006). Data analysis of alternative splicing microarrays. Drug Discov Today *11*, 983-990.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Dreyfuss, G., Choi, Y.D., and Adam, S.A. (1984). Characterization of heterogeneous nuclear RNA-protein complexes in vivo with monoclonal antibodies. Mol Cell Biol *4*, 1104-1114.

Dreyfuss, G., Kim, V.N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. Nat Rev Mol Cell Biol *3*, 195-205.

Dreyfuss, G., Matunis, M.J., Pinol-Roma, S., and Burd, C.G. (1993). hnRNP proteins and the biogenesis of mRNA. Annu Rev Biochem *62*, 289-321.

Fairbrother, W.G., Yeh, R.F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science *297*, 1007-1013.

Friedman, B.A., Stadler, M.B., Shomron, N., Ding, Y., and Burge, C.B. (2008). Ab initio identification of functionally interacting pairs of cis-regulatory elements. Genome Res *18*, 1643-1651.

Gardina, P.J., Clark, T.A., Shimada, B., Staples, M.K., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S., Davies, C., Williams, A., and Turpaz, Y. (2006). Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics *7*, 325.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M., Ulrich, A., Wardle, G.S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129-141.

Han, S.P., Tang, Y.H., and Smith, R. (2010). Functional diversity of the hnRNPs: past, present and perspectives. Biochem J *430*, 379-392.

Homer, N., Merriman, B., and Nelson, S.F. (2009). BFAST: an alignment tool for large scale genome resequencing. PLoS One *4*, e7767.

Hu, G.K., Madore, S.J., Moldover, B., Jatkoe, T., Balaban, D., Thomas, J., and Wang, Y. (2001). Predicting splice variant from DNA chip expression data. Genome Res *11*, 1237-1245.

Huelga, S.C., and Yeo, G.W. (2012). Alternative Splicing in Stem Cells. In Computational Biology of Embryonic Stem Cells, M. Zhan, ed. (Bentham Science Publishers), pp. 147-160.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302*, 2141-2144.

Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods *7*, 1009-1015.

Keene, J.D., Komisarow, J.M., and Friedersdorf, M.B. (2006). RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nat Protoc *1*, 302-307.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.

Kim, J.B., Porreca, G.J., Song, L., Greenway, S.C., Gorham, J.M., Church, G.M., Seidman, C.E., and Seidman, J.G. (2007). Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. Science *316*, 1481-1484.

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol *17*, 909-915.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Lapuk, A., Marr, H., Jakkula, L., Pedro, H., Bhattacharya, S., Purdom, E., Hu, Z., Simpson, K., Pachter, L., Durinck, S., Wang, N., Parvin, B., Fontenay, G., Speed, T., Garbe, J., Stampfer, M., Bayandorian, H., Dorton, S., Clark, T.A., Schweitzer, A., Wyrobek, A., Feiler, H., Spellman, P., Conboy, J., and Gray, J.W. (2010). Exon-level

microarray analyses identify alternative splicing programs in breast cancer. Mol Cancer Res *8*, 961-974.

Li, H., Lovci, M.T., Kwon, Y.S., Rosenfeld, M.G., Fu, X.D., and Yeo, G.W. (2008a). Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. Proc Natl Acad Sci U S A *105*, 20179-20184.

Li, H., Ruan, J., and Durbin, R. (2008b). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res *18*, 1851-1858.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., Darnell, J.C., and Darnell, R.B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464-469.

Lorenz, C., von Pelchrzim, F., and Schroeder, R. (2006). Genomic systematic evolution of ligands by exponential enrichment (Genomic SELEX) for the identification of protein-binding RNAs independent of their expression levels. Nat Protoc *1*, 2204-2212.

Lukong, K.E., Chang, K.W., Khandjian, E.W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. Trends Genet *24*, 416-425.

Martinez-Contreras, R., Cloutier, P., Shkreta, L., Fisette, J.F., Revil, T., and Chabot, B. (2007). hnRNP proteins and splicing control. Adv Exp Med Biol *623*, 123-147.

Mayeda, A., Helfman, D.M., and Krainer, A.R. (1993). Modulation of exon skipping and inclusion by heterogeneous nuclear ribonucleoprotein A1 and pre-mRNA splicing factor SF2/ASF. Mol Cell Biol *13*, 2993-3001.

McKnight, S.L., and Miller, O.L., Jr. (1976). Ultrastructural patterns of RNA synthesis during early embryogenesis of Drosophila melanogaster. Cell *8*, 305-319.

Misquitta-Ali, C.M., Cheng, E., O'Hanlon, D., Liu, N., McGlade, C.J., Tsao, M.S., and Blencowe, B.J. (2011). Global profiling and molecular characterization of alternative splicing events misregulated in lung cancer. Mol Cell Biol *31*, 138-150.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet *40*, 1413-1415.

Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A.L., Mohammad, N., Babak, T., Siu, H., Hughes, T.R., Morris, Q.D., Frey, B.J., and Blencowe, B.J. (2004). Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol Cell *16*, 929-941.

Sanford, J.R., Coutinho, P., Hackett, J.A., Wang, X., Ranahan, W., and Caceres, J.F. (2008). Identification of nuclear and cytoplasmic mRNA targets for the shuttling protein SF2/ASF. PLoS One *3*, e3369.

Sanford, J.R., Wang, X., Mort, M., Vanduyn, N., Cooper, D.N., Mooney, S.D., Edenberg, H.J., and Liu, Y. (2009). Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. Genome Res *19*, 381-394.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., and Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A *100*, 15776-15781.

Tacke, R., and Manley, J.L. (1995). The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. EMBO J *14*, 3540-3551.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212-1215.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial analysis of gene expression. Science *270*, 484-487.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

Wang, L., Xi, Y., Yu, J., Dong, L., Yen, L., and Li, W. (2010). A statistical method for the detection of alternative splicing using RNA-seq. PLoS One *5*, e8529.

Wang, Z., and Burge, C.B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA *14*, 802-813.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873-881.

Yeo, G.W., Coufal, N.G., Liang, T.Y., Peng, G.E., Fu, X.D., and Gage, F.H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat Struct Mol Biol *16*, 130-137.

Yeo, G.W., Xu, X., Liang, T.Y., Muotri, A.R., Carson, C.T., Coufal, N.G., and Gage, F.H. (2007). Alternative splicing events identified in human embryonic stem cells and neural progenitors. PLoS Comput Biol *3*, 1951-1967.

Zhang, Z., Hesselberth, J.R., and Fields, S. (2007). Genome-wide identification of spliced introns using a tiling microarray. Genome Res *17*, 503-509.

2 RNA binding proteins affecting splicing in human disease and development

2.1 Background

The precise control of RNA processing by RNA binding proteins (RBPs) is important for the fate of all RNAs in the cell, defects of which can lead to numerous accounts of human genetic diseases and cancer pathologies in many tissues (Carpenter et al., 2006; Lagier-Tourenne et al., 2010; Melko and Bardoni, 2010). Specifically, alternative splicing is under the coordinated control of many RBPs, and because of the high prevalence of alternative splicing in all multi-exon human genes (Pan et al., 2008; Wang et al., 2008), aberrant splicing is widespread in many diseases. Disruption of the fine-tuned balance of RBPs on RNA targets caused by sequence changes in the RNA targets (cis), or changes in the localization or expression of RBPs (*trans*) can produce improperly spliced RNAs.

2.1.1 Regulatory sequence mutations yield aberrant splicing

At the sequence level, it has been shown that more than half of human disease causing mutations yield aberrant splicing (Lopez-Bigas et al., 2005). Disruptions in key regulatory sequences near and in exons can effect which trans-acting factors bind to activate or repress the splicing of a particular exon. A classical disease example of misregulated alternative splicing in *cis* is spinal muscular atrophy (SMA). SMA is characterized by the loss of SMN1 gene, which produces an efficient SMN protien, and the retention of SMN2 gene, which produces a truncated and less effificient SMN protein (Lefebvre et al., 1995). The difference between SMN1 and SM2 genes is a single nucleotide difference in exon 7. This difference is enough to allow RNA binding protein

hnRNP A1 to bind to *SMN2* and cause exclusion of exon 7, yeilding an unstable SMNΔ7 protein with decreased activity (Kashima et al., 2007). The identification of this binding event, led Adrian Krainer and collegues to partner with ISIS pharmecueticals and explore molecular interventions, anti-sense oligonucleotides, that can block hnRNP A1 binding to SMN2 as a disease treatment (Passini et al., 2011). Successful candidates are currently in clinical trials for disease therapy.

2.1.2 Mislocalized RNA binding proteins yield aberrant splicing

Other genetic diseases with abberant splicing are caused not by sequnce changes in the RNA targets, but by the misregulation of the *trans*-acting RBPs important for regulation. Mutations in many RNA binding proteins have lead to the mislocalization of the protein, causing either a toxic gain-of-function or loss-of-function phenotype.

In neurodegenerative diseases amyotrophic lateral sclerosis (ALS) and frontotemporal lobar degeneration (FTLD), RNA and DNA binding proteins TDP-43 (*TARDBP*) and FUS/TLS are mutated and misslocalized to protein aggregates in the cytoplasm. It has been suggested that misslocalization of the TDP-43 protein leads to a loss-of-function phenotype, preventing its known and widespread regulation of alternative splicing patterns (Polymenidou et al., 2011b).

Similarly, in Myotonic Dystrophy (DM) expanded nucleotide repeats CUG occur in the 3'UTR of *DMPK* and CCUG occur in an intron of *ZNF9*. These extended repeats cause sequestering of two RBPs known to bind CUG-rich sequences, muscleblind protein (MBNL) and CUG-binding protien (CUGBP1). Sequestration of MBNL reduces the cellular concentration of MBNL causing a loss-of-funtion phenotype and negativly affects the alternative splicing events that are dependent on this protein (Kanadia et al., 2006). CUGBP, on the other hand, becomes more stable when bound to the repeat and

exibits a toxic-gain-of-function phenotype that also disrupts normal splicing patterns (Kuyumcu-Martinez et al., 2007).

2.1.3 Increased abundance of RNA binding proteins in cancer

Cancer pathologies, including colon, lung and breast cancers, have been shown to have major defects in splicing, with large proportions of mRNAs being aberrantly spliced (Gardina et al., 2006; Lapuk et al., 2010; Misquitta-Ali et al., 2011). The normal cellular balance of RBPs is highly disrupted in cancer cells, due to the overexpression of many RBPs important for splicing, including SR proteins and the most abundant class of RBPs, the hnRNP proteins. The SR protein SF2/ASF a known proto-oncogene amplified in human tumors (Karni et al., 2007). Elevated levels of SF2/ASF regulated splicing lead to inactivation of its specific target genes, including tumor suppressor BIN1, and transform cancerous cells (Karni et al., 2007). The hnRNP proteins have been shown to directly affect several alternative splicing events that affect cancer progression (David and Manley, 2010). As an example, recent reports have identified that hnRNP A1 binds to repress exon 9 of pyruvate kinase, resulting in the inclusion of the mutually exclusive exon 10, and conferring a proliferative advantage to cancer cells (Chen et al., 2010; Clower et al., 2010; David et al., 2010).

2.1.4 Alternative splicing in development and stem cell biology

Unexpectedly, in addition to genes important in human diseases, genes important for human development and the biology of stem cells have not escaped regulation by alternative splicing. As an example, *OCT4*, which when overexpressed in differentiated cells aids in reprogramming cells to a pluripotent state (Takahashi et al., 2007), is regulated by alternative splicing. Differential expression of various *OCT4* isoforms has been observed before and after differentiation, each having unique

functions, suggesting regulatory roles for stem cell pluripotency (Atlasi et al., 2008; Cauffman et al., 2006; Lee et al., 2006). Other genes implicated in stem cell biology with functionally distinct isoforms produced by alternative splicing include: insulin-like growth factor 1 (*IGF-1*) (Ates et al., 2007), DNA methyltransferase 3 Beta (*DNMT3B*) (Gopalakrishnan et al., 2009), protein kinase C delta (*PKCδ*) (Patel et al., 2006), and vascular endothelial growth factor A (*VEGFA*) (Lin et al., 2008). Identifying alternative splicing events in stem cells and understanding how they are regulated has far-reaching implications to our understanding of human development and disease.

Here we present two detailed studies of RNA binding proteins directly related to development and human disease, and involved in splicing. In the first study, for a RBP not previously implicated in splicing, LIN28, we examine the indirect effect on splicing. In the second study, for an RBP previously implicated in splicing, TDP-43, we examine the effect of specific disease mutations on its splicing.

2.2 Indirect regulation of alternative splicing by Lin28 in stem cell development

Conserved across bilaterian animals, LIN28A (herein referred to as LIN28) is highly expressed early in development and is selectively downregulated during differentiation (Moss et al., 1997; Yang and Moss, 2003). Consistent with this pattern of expression, LIN28 has been shown to be important in the maintenance of embryonic stem (ES) cell pluripotency and efficacy of induced pluripotent stem cell (iPSC) derivation (Newman and Hammond, 2010; Yu et al., 2007). Of the factors used in reprogramming, LIN28 is unique in its classification as an RBP, rather than as a transcription factor. Altered expression levels of RBPs often results in human disease. Notably, aberrant upregulation of LIN28 has been found in a range of different cancer

cells and primary tumor tissues (Cao et al., 2011); (Viswanathan et al., 2009; West et al., 2009).

While alternative splicing has only recently been implicated in stem cell biology, it is clear that alternative splicing plays significant roles in fine-tuning the proteome during lineage specification. Many RNA binding proteins are known to *directly* affect alternative splicing, mainly through the binding and occlusion of other splicing factors. Interestingly, we found that LIN28 binds to mRNA regions within transcripts that code for splicing factors, including *TDP-43, FUS/TLS, TIA-1,* and *hnRNP F* and controls their protein abundance (Wilbert et al., 2012). We therefore expected that LIN28 misregulation could result in changes in alternative splicing (AS). We show that LIN28 can strongly affect alternative splicing levels *indirectly*, through the regulation of splicing factor abundance.

2.2.1 Increased Levels of LIN28 in Somatic Cells Causes Widespread Changes in Alternative Splicing

To explore splicing changes dependent on LIN28, we subjected total RNA from a stable Flp-In HEK293 cell line that constitutively expresses a C-terminal V5-tagged human LIN28 protein at physiological levels (LIN28-V5 293 cells) and control Flp-In-293 cells to splicing-sensitive microarray (HJAY) analysis. We identified 1,985 differentially regulated AS events in the presence of LIN28 expression, out of 14,643 events detected on the array (Figure 2.1A). These events are comprised of isoform changes in approximately 1,965 genes. This number of AS events is comparable to the numbers regulated by well-studied splicing factors such as hnRNP proteins, RBFOX2 and HuR (Huelga et al., 2012; Mukherjee et al., 2011; Venables et al., 2009). Since we've seen little evidence of LIN28 binding to intronic regions (Wilbert et al., 2012), we reasoned that LIN28 likely interacts with cytoplasmic, mature mRNA transcripts, which suggests that the observed AS events are most likely the downstream result of LIN28 regulation of

splicing factors. We successfully validated a number of these AS changes by semiquantitative RT-PCR with an 85% validation rate (Figure 2.1B and Figure 2.3A). As an interesting example, we validated the alternative splicing of a 63 nucleotide (nt) cassette exon 23a in the neurofibrimin 1 (NF1) gene, which is skipped upon expression of LIN28-V5 (Figure 2.1B). As a known negative regulator of the *Ras* signaling pathway, accurate control of NF1 isoforms are important in cancer and neuronal differentiation (Patrakitkomjorn et al., 2008), thereby providing a glimpse into signaling pathways that LIN28 may affect through regulation of AS.

To analyze the extent of alternative splicing events affected due to the regulation of a single splicing factor by LIN28, we overexpressed a plasmid harboring the open reading frame of TDP-43 fused to a C-terminal GFP in Flp-In-293 cells, reproducing the upregulation of TDP-43 upon LIN28 expression observed in LIN28-V5 293 cells. We subjected total RNA to splicing-sensitive microarray analysis (Figure 2.3B), identifying a total of 865 AS events that changed, including 526 differentially spliced cassette exons (Figure 2.1A). Of the cassette exons affected by stable LIN28-V5 expression in our cell line, we identified a significantly overlapping subset of 113 cassettes (13%) that were also affected upon upregulation of TDP-43 ($p < 10^{-5}$, hypergeometric test), with 70% of the cassette events changing in the same direction (Figure 2.1C). Of the hundreds of splicing factors that LIN28 is predicted to regulate, LIN28 affects a statistically significant overlapping set of alternative splicing events with at least one splicing factor, TDP-43.

2.2.2 Decreased levels of LIN28 and LIN28B in embryonic stem cells modulates translation of RNA binding proteins

We were surprised to find that depletion of LIN28 in hES cells resulted in less than half of the number of AS events as in LIN28-V5 293 cells, and that few of these events were reciprocal (Figure 2.3C and Figure 2.2A). In addition, despite the high

concordance between hES and LIN28-V5 293 cells of the location of LIN28 binding sites on target mRNAs, its splicing factor targets did not display a decrease in protein levels expected upon knockdown of LIN28 (Figure 7C). Given that LIN28B, the paralog of LIN28, was significantly enhanced when LIN28 was depleted in hES cells (Figure 2.2B) and that LIN28 and LIN28B interact with a common set of mRNAs encoding splicing factors (Figure 2.3D), we hypothesized the LIN28B may compensate for loss of LIN28. To address this relation between LIN28 and LIN28B, we electroporated hES cells with siRNAs that individually depleted LIN28 and LIN28B, as well as both proteins simultaneously (Figure 2.2C). Interestingly, we observed that hnRNP F increases at the protein level with depletion of LIN28B, TDP-43 is downregulated when either LIN28 or LIN28B was depleted but not further downregulated by depletion of both, and FUS/TLS was reduced only when both LIN28 and LIN28B were concurrently depleted. Therefore, LIN28 and LIN28B may exhibit synergistic (FUS/TLS), and both repressive (hnRNP F) and enhancing (TDP-43, FUS/TLS) effects on translation of their mRNA targets in stem cells. Our observations that LIN28 and LIN28B have differing effects on their targets, and that LIN28 levels affect LIN28B expression (Figure 2.2B), reveal an alternative mechanism through which LIN28 and LIN28B expression can shape cell fate and homeostasis.

2.2.3 LIN28 Summary

In summary, upregulation of splicing factor protein levels in response to an increase in LIN28 in somatic cells leads to widespread changes in alternative splicing patterns. Surprisingly, downregulation of LIN28 in hES cells does not always result in reciprocal changes for these RBPs. Furthermore, LIN28B does not in general compensate for lack of LIN28 function, despite also interacting with mRNAs encoding these RBPs, and has different, or sometimes synergistic, effects on these targets. This

cell type specific control of gene regulatory targets by LIN28 presents an alternative mechanism through which LIN28 and LIN28B expression can shape cell fate and homeostasis. This genome-wide study reveals avenues by which LIN28 impacts gene regulatory networks through direct regulation of its mRNA targets, and provides a valuable framework for future characterization of the molecular roles of LIN28 and LIN28B in biological pathways.

2.3 Disease mutations in TDP-43 lead to splicing changes in neuronal genes

Amyotrophic lateral sclerosis (ALS) and frontotemporal lobar degeneration with ubiquitinated inclusions (FTLD-U) are progressive, adult-onset neurodegenerative diseases with overlapping clinical and pathological features (Da Cruz and Cleveland, 2011; Liscic et al., 2008; Lomen-Hoerth et al., 2002). ALS is characterized by the selective loss of upper and lower motor neurons leading to progressive fatal paralysis and muscle atrophy. A large majority (~90%) of ALS and FTLD-U cases are without a known genetic cause. Importantly, in these sporadic cases, the appearance of ubiquitinated inclusions within the affected neurons of the nervous system characterizes both ALS and FTLD-U patients, suggesting an overlapping mechanism underlying both diseases. Biochemical characterization of brains and spinal cords from ALS and FTLD-U patients identified TAR DNA Binding Protein (TDP-43) as the major protein component of these ubiquitinated inclusions (Arai et al., 2006; Neumann et al., 2006). The discovery of ALS-linked mutations in the glycine-rich C-terminal domain of TDP-43 (Gitcho et al., 2008; Kabashi et al., 2008; Sreedharan et al., 2008) demonstrated a pathological role of TDP-43 in both diseases. The subsequent identification of mutations in a structurally and functionally related nucleic acid binding protein, FUS/TLS (fused in sarcoma/translocated in liposarcoma) (Kwiatkowski et al., 2009; Vance et al., 2009), further implicated defects in RNA processing in ALS pathogenesis.

TDP-43 is a multi-functional nucleic acid binding protein. Within the nervous system, TDP-43 binds to >6,000 pre-mRNAs and affects the levels of ~600 mRNAs and the splicing patterns of another 950 (Polymenidou et al., 2011a). Structurally, the 414 amino acid protein consists of two RNA recognition motifs (RRM1 and RRM2) (Buratti et al., 2001; Ou et al., 1995), nuclear import and export signal (Winton et al., 2008), and a glycine-rich region implicated in protein-protein interactions (Ayala et al., 2005; Wang et al., 2004) that include components of the RNA splicing machinery (Buratti et al., 2005; D'Ambrogio et al., 2009).

It remains unresolved whether toxicity to motor neurons from mutations in TDP-43 is mediated through a gain of toxic property, loss of function, or a combination of both. By generation of transgenic mice encoding levels of wild-type or mutant human TDP-43 comparable to endogenous TDP-43, we demonstrate mutant-dependent, age-dependent motor neuron disease from ALS-linked TDP-43 mutants in the absence of overexpression, cytoplasmic accumulation of a 35 kD TDP-43 fragment, or insoluble TDP-43 aggregates. Accompanying autoregulation-mediated reduction of endogenous wild type TDP-43 are splicing alterations previously identified to be TDP-43 dependent (Polymenidou et al., 2011a). Additional splicing alterations are identified by systematic genome-wide analyses of alternative splicing that are indicative of both enhancement and loss of function by the TDP-43 mutants for individual RNA substrates, from which we conclude that ALS-linked mutations confer both loss and gain of function properties to TDP-43 and that these act intranuclearly to induce splicing alterations that may underlie age-dependent motor neuron disease.

2.3.1 Widespread Splicing Changes from Altered TDP-43 Levels Within the Central Nervous System

To evaluate if normal patterns of alternative splicing were disrupted upon reduction of mouse TDP-43 and its replacement with wild type or mutant human TDP-43 at or above the level of endogenous TDP-43 in normal mice, RNA extracted from cortices of two month old mice were examined using Affymetrix splicing sensitive microarrays (Figure 2.4A). Analysis of splicing changes in cortices of TDP-43$^{Wild\text{-}Type}$, TDP-43^{Q331K}, and TDP-43$^{Q331K\text{-}Low}$ revealed that 824, 1195, and 208 exons, respectively, were differentially spliced relative to mRNA from non-transgenic mice, using a statistical threshold that captured as many significant changes as possible (absolute separation score of < 0.3) (Du et al., 2010; Sugnet et al., 2006) (Figure 2.4A).

We focused on the largest represented class of alternative splicing events, namely cassette exons (i.e., exons that can be included or excluded in the final mRNA [white slice of pie charts, Figure 2.4A]). For comparison, we reanalyzed our previous splicing array data (Polymenidou et al., 2011a) using a consistent threshold (absolute separation score of < 0.3), identifying 737 cassette exons that were alternatively spliced upon depletion of endogenous mouse TDP-43 in the central nervous system (Figure 2.6). These exons - whose abundance is normally regulated by TDP-43 levels - were then compared with the 314, 533, and 97 cassette exons that were misregulated in human TDP-43$^{Wild\text{-}Type}$, TDP-43^{Q331K}, and TDP-43$^{Q331K\text{-}Low}$ transgenic animals.

To determine if misregulated exons were associated with endogenous TDP-43 binding (as measured by cross-linking and immunoprecipitation followed by deep sequencing (Polymenidou et al., 2011a)), we separated the 737 exons into “direct” (175, that contained TDP-43 binding sites within 2 kb of the exon) and “indirect” (562 without TDP-43 binding sites near the exon) targets (Figure 2.4B). Notably, when comparing the

percentages of direct and indirect targets that overlapped with TDP-43^{Q331K}-dependent misregulated exons, we observed a statistically significant ($p<0.0005$ by chi-square analysis) 1.8 fold-enrichment of direct targets (26% (45/175) compared to indirect ones (14% (81/562)). A similar (2.4 fold) enrichment was seen for altered splicing of direct targets in TDP-43$^{Q331K\text{-}Low}$ (6% of direct targets (11/175) compared to 2.6% of indirect targets (15/562)). No significant enrichment was seen in a similar comparison for TDP-43$^{Wild\text{-}Type}$ (9% (16/175) to 7% (39/562)) (Figure 2.4C). Thus, the human mutant transgenes preferentially affect direct, endogenous TDP-43-dependent alternative splicing events.

2.3.2 Mutant TDP-43 Retention or Enhancement of TDP-43 Function in Splicing of Most pre-mRNAs

To test whether the splicing changes for cassette exons reflected a loss or gain of normal TDP-43 function, we compared the directionality of exon inclusion when TDP-43 was depleted or in response to expression in mice of the transgene-encoded human TDP-43 (Figure 2.4D, i-vi). Of the cassette exon splicing changes observed in the TDP-43 transgenic mice, we found that 55 of 314 in TDP-43$^{Wild\text{-}Type}$ (Figure 2.4D, i-ii), 126 of 533 in TDP-43^{Q331K} (Figure 2.4D, iii-iv), and 26 of 97 TDP-43$^{Q331K\text{-}Low}$ (Figure 2.4D, v-vi) overlapped with cassette exon changes identified following TDP-43 depletion. For directly bound exon targets whose splicing was altered by expression of human TDP-43, the majority of the affected cassette exons overlapping with direct targets changed in a direction opposite to loss of TDP-43 (TDP-43$^{Wild\text{-}Type}$ [13 of 16], TDP-43^{Q331K} [35 of 45], TDP-43$^{Q331K\text{-}Low}$ [10 of 11]) (in gray, Figure 2.4D, i, iii, and v).

Furthermore, at higher TDP-43^{Q331K} levels (and the correspondingly stronger reduction in endogenous mouse TDP-43), splicing of 35 of 45 directly bound cassette exons (Figure 2.4D, iii) changed in a manner consistent with an elevated normal function

of endogenous mouse TDP-43 (the opposite of TDP-43 depletion). In other words, these cassette exons whose inclusion (or exclusion) was previously shown to be dependent on mouse TDP-43 were more included (or excluded) when human TDP-43^{Q331K} was expressed, consistent with this mutant retaining normal (or enhanced) TDP-43 activity for most pre-mRNAs.

2.3.3 Mutant TDP-43–Dependent Loss-of-Function for Splicing of Specific pre-mRNAs

Ten direct exon targets in TDP-43^{Q331K} mice were altered in a manner indicative of a reduction in endogenous TDP-43 function in facilitating the splicing of these TDP-43 target exons (Figure 2.4D, iii). Exons whose splicing were reported altered by microarray analysis, as well as ones of biological interest but were not represented on the array, were then analyzed by semi-quantitative RT-PCR (Figure 2.4E-F, Table 2.1, Figure 2.7). To identify and/or validate splicing changes that are specifically affected by ALS-linked mutant TDP-43, we focused on changes occurring only in both TDP-43^{Q331K} and TDP-43$^{Q331K\text{-}Low}$-transgenic mice, but not in TDP-43$^{Wild\text{-}Type}$ mice (whose accumulated transgene level is similar to that of TDP-43$^{Q331K\text{-}Low}$ transgenic mice). Within these, the presence of TDP-43 mutant produced splicing changes in *Kncip2, Abhd14a, Ctnnd1* and *Atp2b1* that mimicked reduction of endogenous TDP-43 activity (Table 2.1, Figure 2.4E and Figure 2.7A), demonstrating a “loss of normal function” of human TDP-43^{Q331K}, which even at a high level of expression did not replace the endogenous mouse TDP-43 for the splicing of these exons.

2.3.4 Dose-Dependent Enhanced Splicing of Some pre-mRNAs by Human TDP-43

Within the exons that overlapped with those that changed upon TDP-43 depletion (Figure 2.4D, i, iii, v), we found some that primarily displayed a dose-dependent TDP-43^{Q331K} or TDP-$43^{Wild\text{-}Type}$ pattern of splicing in the opposite direction to what occurs after TDP-43 depletion (Figure 2.4F; *Sort1* and Figure 2.7B; *Ppm2c, AU040829,* and *Caly,* whose exons were more excluded, in contrast to more included in TDP-43 depletion). Added to this, for other pre-mRNA targets, enhancement of normal TDP-43 function in splicing was seen for TDP-43^{Q331K}, as demonstrated by dose-dependent splicing changes in a direction opposite what is seen with TDP-43 depletion (e.g., *Eif4h and Taf1b,* whose exons were more excluded, in contrast to increased inclusion in TDP-43 depletion) (Figure 2.4E, Table 2.1).

These findings for individual RNAs were supported globally by a genome-wide analysis of direct TDP-43 targets whose exons were included or excluded upon human TDP-43 expression (Figure 2.4G). On average, ~20% of unaffected exons represented on the array were direct targets (green line in Figure 2.4G). However, when focusing on alternatively changing cassette exons, we found an overall increase in the percentages of direct TDP-43 exon targets (~43%, TDP-43^{Q331K} and TDP-$43^{Wild\text{-}Type}$; up to 53% in TDP-$43^{Q331K\text{-}low}$) that were excluded in the presence of the human TDP-43 protein, compared with only ~20% of the included exons (Figure 2.4G, ii). Surprisingly, 39% of excluded exons in the TDP-43^{Q331K} animals not previously demonstrated as regulated by endogenous TDP-43 depletion were also direct targets (Figure 2.4G, iii). That is, some exons that were unchanged following TDP-43 depletion displayed a dose-dependent pattern of splicing in the direction of enhanced exclusion in the presence of the human TDP-43 (Table 2.1, Figure 2.4F; *Atxn2,* and Figure 2.7C; *Zfp414,* and *Kcnj3,* whose

exons were more excluded, in contrast to no change upon TDP-43 depletion). This supports a hypothesis that certain exons, although bound by TDP-43, are only misregulated upon elevation of wild-type or mutant TDP-43, but are not altered upon its loss. In other words, elevation or enhancement of TDP-43 activity through dose or mutation produces aberrant splicing events separate from those resulting from loss of TDP-43.

2.3.5 Mutant-Specific Splicing Alterations Accompany Motor Neuron Disease Development

Having identified that the human transgenes cause both loss and enhancement of normal function, we focused on alternative splicing changes within the spinal cord, the tissue that develops the degenerative phenotype. RNA extracted from the spinal cords of 2 month old non-transgenic, TDP-43$^{Wild\text{-}Type}$ and TDP-43^{Q331K} animals prior to the onset of significant motor dysfunction were analyzed on the same splicing sensitive microarrays (Figure 2.5A). We identified 4462 and 4399 alternative splicing events in the TDP-43$^{Wild\text{-}Type}$ and TDP-43^{Q331K} spinal cord conditions, respectively (Figure 2.5A). Approximately 38% of changes in TDP-43$^{Wild\text{-}Type}$ cortex (118) and 42% of changes observed in TDP-43^{Q331K} cortex (226) were also present in the spinal cord (Figure 2.5B). Notably, with the exception of a cassette exon within the *Kctd9* gene, altered splicing of all 1,060 cassette exons that overlapped between spinal cord in TDP-43$^{Wild\text{-}Type}$ and TDP-43^{Q331K} mice occurred in the same direction, demonstrating that at least for this subset of RNA targets the mutant remains functionally similar to wild-type TDP-43 (Figure 2.5C).

Use of semi-quantitative RT-PCR validated a set of splicing changes that were consistent with those identified in the cortex (Figure 2.5D and Table 2.2), again demonstrating that TDP-43^{Q331K} mutant confers both loss and retention of normal TDP-43 function. Furthermore, among the 419 splicing events unique in TDP-43^{Q331K} spinal

cord were several RNAs whose encoded proteins are involved in neurological function and transmission (Table 2.3), including the synaptic cell-adhesion molecules neurexins 1 and 3 (*Nrxn1* and *Nrxn3*), protein phosphatase 3 (also known as calcineurin (*Ppp3ca*)), and glutamate receptor 2 (*Gria2,* encoding GluR2), which has been proposed to modify motor neuron vulnerability to excitotoxicity (Pizzi et al., 2000; Van Damme et al., 2007; Van Damme et al., 2005). In contrast, the 394 events unique to the TDP-43$^{\text{Wild-Type}}$ spinal cord condition did not include a similar representation from genes involved in neurological function.

2.3.6 Mutant TDP-43 Summary

Systematic genome-wide splicing arrays and computational analyses have revealed that nearly complete replacement of endogenous TDP-43 with human TDP-43 is accompanied by widespread changes in alternative pre-mRNA splicing, with some changes reflecting an “enhancement of normal function” exacerbated by increased accumulation of TDP-43. Furthermore, we have found that a small subset of splicing changes is uniquely dependent on the TDP-43$^{\text{Q331K}}$ mutant protein. Within these mutant-dependent changes, splicing alterations include an enhancement of normal function, while others are characteristic of TDP-43 loss of function. Hence, the TDP-43$^{\text{Q331K}}$ mutant confers both gain and loss of function properties, both of which are associated with mutant-dependent motor neuron disease. Additionally, the high overlap in splicing alterations in both TDP-43$^{\text{Wild-Type}}$ and TDP-43$^{\text{Q331K}}$ spinal cord further emphasizes that the TDP-43$^{\text{Q331K}}$ mutant remains functionally similar to its wild-type counterpart. Nevertheless, the TDP-43$^{\text{Q331K}}$ mutation also produces a subset of unique splicing alterations within the spinal cord, specifically in genes involved in neurological function.

The dependency of some splicing events solely on the level of human TDP-43 - wild-type or mutant (e.g., *Sort1*) - highlights a sensitivity in TDP-43-dependent splicing to

levels of accumulated TDP-43 protein. Depletion of endogenous mouse TDP-43 has been demonstrated to have widespread effects on both splicing and levels of mRNAs in the central nervous system (Polymenidou et al., 2011a). Our work here demonstrates that an increase in accumulated TDP-43 protein affects splicing just as broadly, results consistent with TDP-43's previously observed role in exon repression in the nervous system (Polymenidou et al., 2011a). Together, we propose that maintenance of a homeostatic level of TDP-43 protein is critical for its splicing function in the central nervous system, and that disruption of this level produces widespread aberrant alternative splicing events.

Finally, our systematic splicing analysis enables us to propose a molecular model for human TDP-43 function in which it drives: (1) dosage-enhanced increases in exon exclusion of normal endogenous mouse TDP-43 targets, (2) mutant-enhanced increases in exon exclusion of normal TDP-43 targets, and (3) mutant-specific decreases in exon exclusion, due to a loss of function on normal TDP-43 targets (Figure 2.5E). Taken together, the unique splicing changes in TDP-43^{Q331K} transgenic mice within genes in spinal cords could provide a molecular basis for selective vulnerability of motor neurons to such mutants.

2.4 Methods

2.4.1 Human cell culture and stable cell line generation

The LIN28 open reading frame (*Homo sapiens*, GenBank: DQ896719) was cloned from a Gateway pENTR221 vector (Open Biosystems) into the Gateway pEF5/FRT/V5 destination vector (Life Technologies) to generate the V5-tagged LIN28. To generate LIN28-V5 293 stable cell lines, pEF5/FRT/LIN28-V5 plasmid was co-transfected along with the FLP Recombinase expressing plasmid pOG44 into Flp-In-293 cells. Stably

transected clones were propagated in media supplemented with 75-100 μg/ml hygromycin B (Life Technologies) and several independent clonal cell lines were established. Human ES cell lines H9 and HUES6 were grown in feeder-free conditions with mTeSR media (STEMCELL Technologies) and on Matrigel (BD Biosciences).

2.4.2 Lentiviral shRNA-mediated and siRNA-mediated knockdown of LIN28 and LIN28B

To achieve knockdown of LIN28, we utilized an shRNA construct targeting human LIN28 in the pLKO.1 vector (TRCN0000102579; Open Biosystems). As a control, a pLKO.1 vector containing an shRNA toward GFP was used (Open Biosystems). To deplete LIN28 and LIN28B, we utilized On-TARGET*plus* SMARTpool siRNAs from Dharmacon (LIN28A: L-018411-01-0005, LIN28B: L-028584-01-0005, On-TARGET*plus* Non-Targeting Pool: D-001810-10-05).

2.4.3 TDP-43 overexpression

Flp-In-293 cells were grown to ~70% confluency and transfected with TDP-43-GFP (Liu-Yesucevitz et al., 2010) or control pEGFP-C2 (Clontech) plasmid using Lipofectamine-2000 (Life Technologies) according to manufacturer's instructions.

2.4.4 Microarray analysis and validation for splicing changes

Microarray data analysis was performed as previously described (Polymenidou et al., 2011a). For each microarray condition, the log2 ratio of skipping intensities to inclusion intensities was estimated using least-squares analysis. Significantly changing splicing events between LIN28-V5 293 cells and untreated Flp-In-293 cells, TDP-43 overexpression and control overexpression Flp-In-293 cells, and LIN28 knockdown and control knockdown H9 hES cells were identified using a q-value < 0.05 and an absolute separation score > 0.5. Significantly changing splicing events between TDP-43^{Q331K},

TDP-43$^{Q331K\text{-}low}$, TDP-43$^{Wild\text{-}Type}$ and control non-transgenic mice were identified using a q-value < 0.05 and an absolute separation score > 0.3.

To test candidate splicing targets as identified by microarray, cDNA was generated by reverse transcribing 1ug of total RNA extracted from the cortex and spinal cord of 2 month old transgenic animals using oligo (dT) primer and Superscript III reverse transcriptase (Invitrogen) according to the manufacturer's instructions. PCR amplification was performed for 28 cycles using cDNA template with primers falling in exons flanking the alternate cassette exons. Products were separated on 10% acrylamide-TBE gels and then stained with SYBR Gold (Invitrogen). Gel imaging and quantification of the isoforms was performed using the Biorad Chemidoc software. The intensity ratios between products including the cassette exon and skipping the cassette exon were averaged for a minimum of 3 biological replicates per genotype. Splicing alterations significantly different from the non-transgenic RNAs were determined with Student's T-test (*, $p<0.05$, **, $p<0.01$ and ***, $p<0.001$).

2.4.5 Western blot analysis

Membrane incubations with anti-GAPDH (Abcam ab8245), anti-LIN28 (Abcam ab46020), anti-LIN28B (Cell Signaling 4196), anti-FUS/TLS (Santa Cruz Biotechnologies SC-47711), anti-TDP43 (Aviva ARP35837_P050), anti-hnRNP F (Santa Cruz Biotechnologies SC-10045), and IMP2 (MBL RN008P) were performed overnight. Secondary antibodies anti-rabbit (Calbiochem 401393 or Cell Signaling 7074), anti-mouse (Cell Signaling 7076), and anti-Goat (Promega V-4771), and chemiluminescence reagents (Thermo Pierce) were used according to manufacturers' recommendations.

2.5 Lin28 and TDP-43 publication acknowledgement

The Lin28 work described here has been published as part of a manuscript entitled “LIN28 Binds Messenger RNAs at GGAGA Motifs and Regulates Splicing Factor Abundance” in *Molecular Cell* (Wilbert et al., 2012). This dissertation author was the secondary author who helped to write the manuscript, analyze the splicing array data, and direct experiments.

The TDP-43 work described here has been published as part of a manuscript entitled “ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43” in *PNAS* (Arnold et al., 2013). This dissertation author was a secondary author who helped to write the manuscript, analyze the splicing array data, and direct experiments.

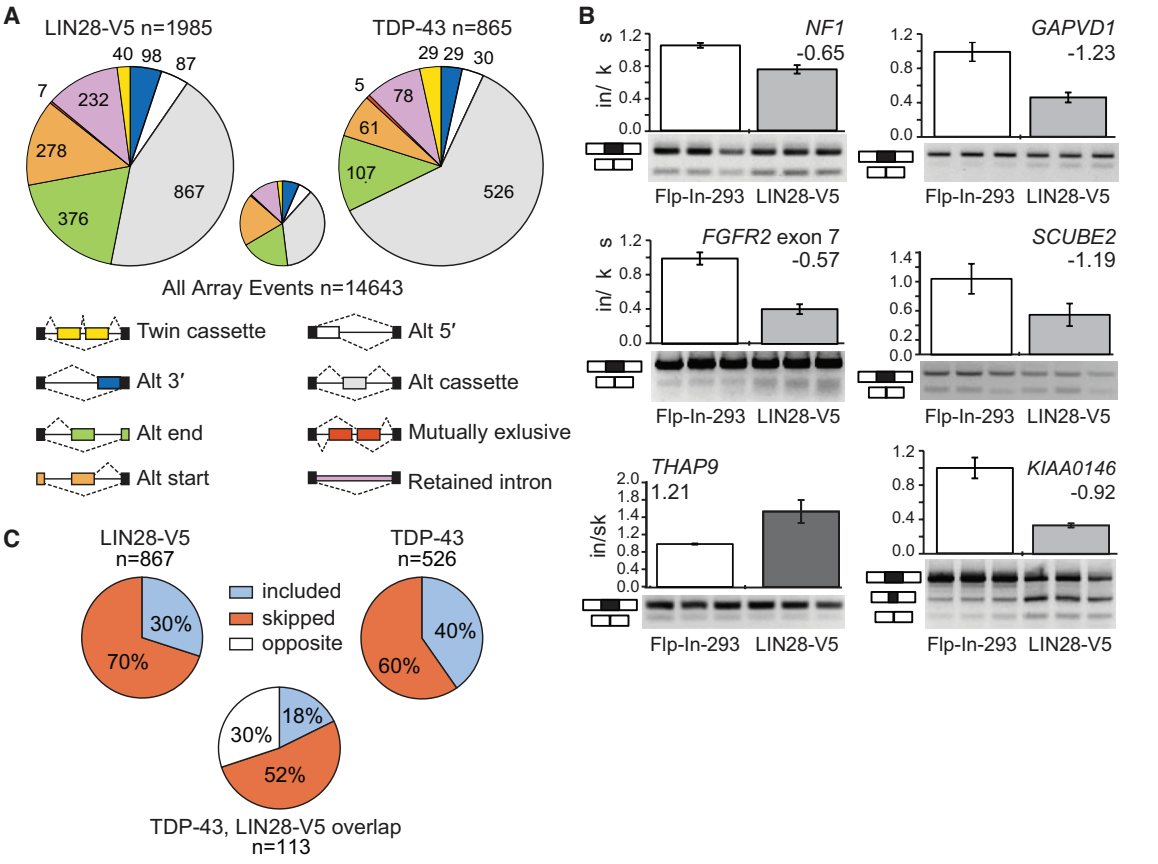


Figure 2.1: LIN28 expression in somatic cells results in thousands of alternative splicing events, in part through regulation of TDP-43 levels.

A. The pie charts display the number of each type of alternative splicing event changed upon overexpression of LIN28-V5 (left) or TDP-43 (right) in Flp-In-293 cells, as detected by splicing-sensitive microarray analyses. The small pie chart (center) represents the distribution of alternative splicing event types detected on the microarray.
B. RT-PCR validations of alternative cassette events detected by microarray analysis. The KIAA0146 event has two cassette exons, thus it generates three isoforms (one where both cassettes are included, one where one cassette is included, and one where both cassettes are skipped). All plots show significant differences between control Flp-In-293 and LIN28-V5 293 cells ($p < 0.05$, Students t-test). Bars represent an average and error bars represent the standard deviation across biological triplicates.
C. The percent of included versus skipped alternative cassette exons upon overexpression of LIN28-V5 (left) or TDP-43 (right) in Flp-In-293 cells. For the alternative cassette exons that changed in both conditions (n = 113), the percent of exons affected in the same (where the exon is included or skipped in both conditions) or opposite direction are shown below.

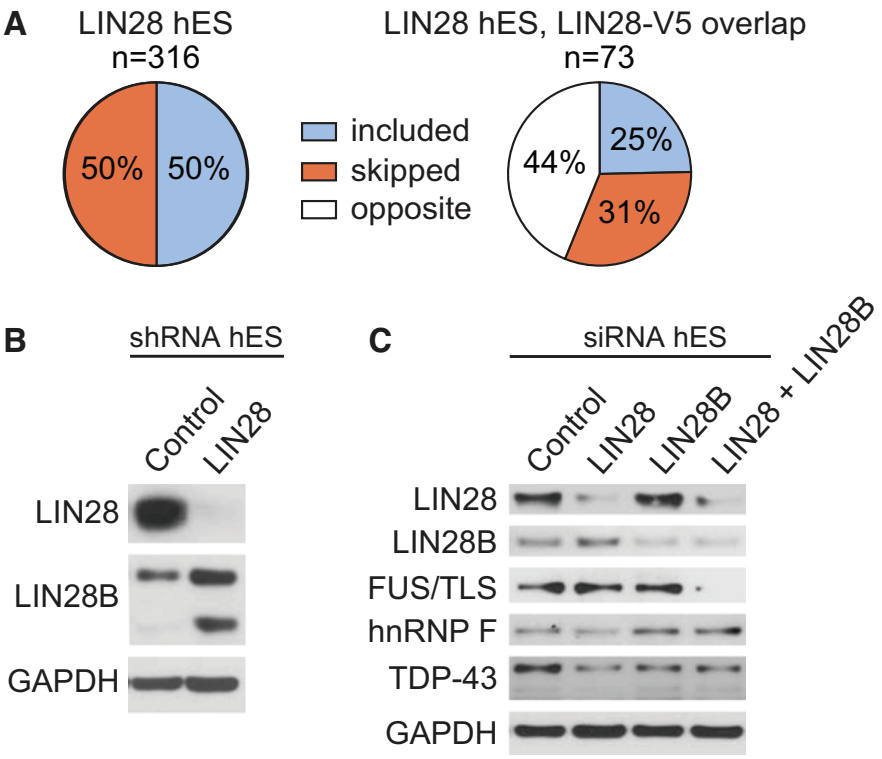


Figure 2.2: LIN28 and LIN28B affect splicing factors differently in human ES cells.

A. The percent of included versus skipped alternative cassette exons upon depletion of LIN28 in hES cells is shown. Of the alternative cassette events changed in both the LIN28 hES and LIN28-V5 293 experiments (n = 73), the percent of exons affected in the same (where the exon is included or skipped in both conditions) or opposite direction are shown below. The direction of cassette exon splicing changes due to LIN28 depletion in hES cells is flipped to correspond to LIN28 overexpression.
B. Western blot analysis of LIN28B levels upon shRNA-mediated depletion of LIN28 in hES cells. GAPDH serves as a loading control.
C. Western blot analysis of LIN28, LIN28B, and splicing factors upon siRNA-mediated depletion of LIN28, LIN28B, or both in hES cells.

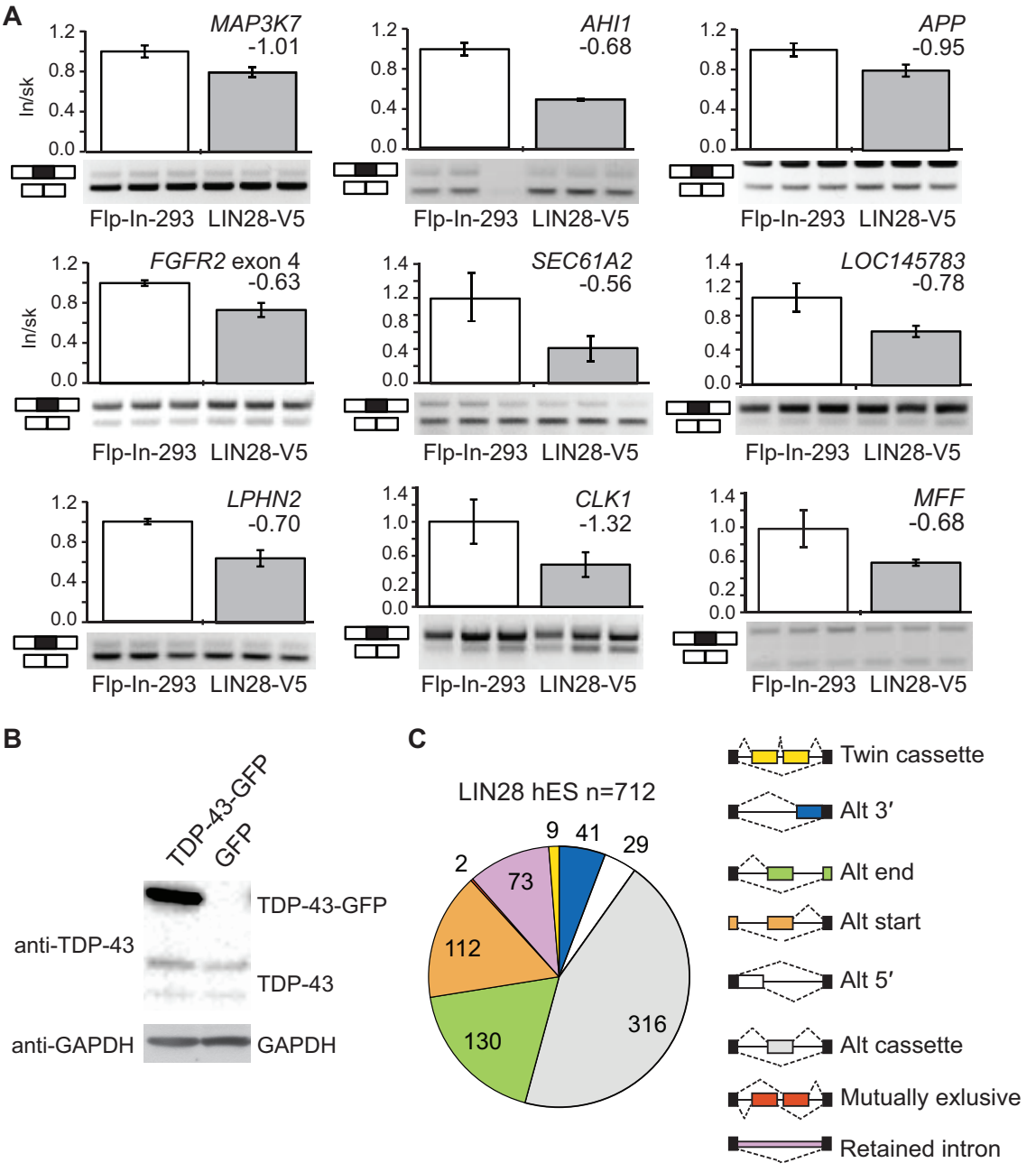
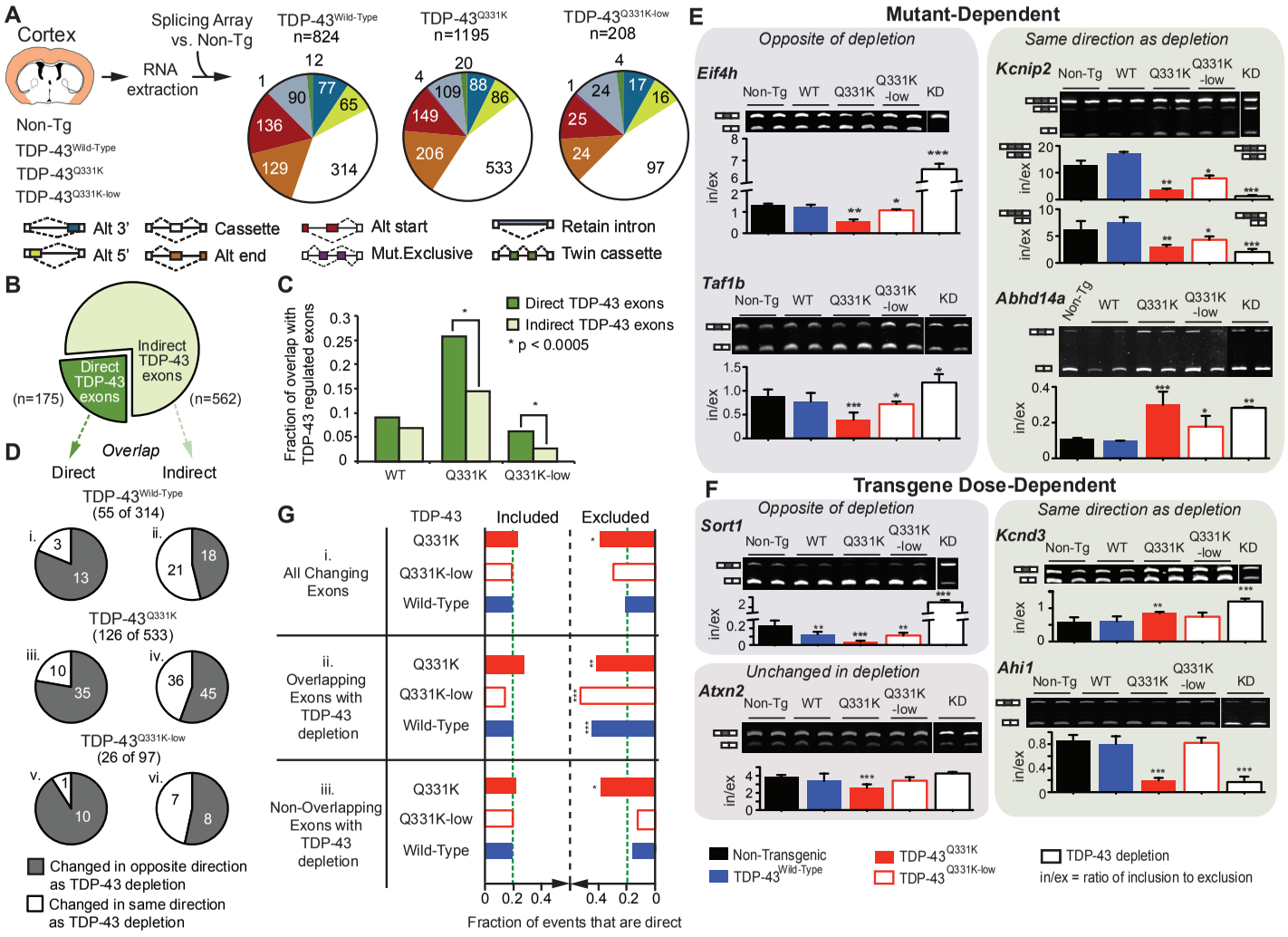


Figure 2.3: LIN28 regulates alternative splicing events

A. Additional RT-PCR validations of alternative cassette events detected by microarray analysis. All plots show significant differences between control Flp-In 293 and LIN28-V5 293 cells ($p < 0.05$, Students t-test). Bar graphs represent the average value and error bars display the standard deviation across biological triplicates. **B.** Western blot analysis using antibodies against TDP-43 and GAPDH (loading control) in Flp-In-293 cells transfected with a TDP-43-GFP or control GFP expression vector. **C.** The pie charts display the number of each type of alternative splicing event changed upon knockdown of LIN28 in H9 hES cells (different event types depicted to the right).

Figure 2.4: Both enhanced activity and loss-of-function for specific RNA splicing targets directly bound by TDP-43^{Q331K}

A. Experimental strategy to identify differentially regulated splicing events using Affymetrix splicing sensitive microarrays to analyze RNAs extracted from cortices of 2-month old non-transgenic, TDP-43$^{Wild\text{-}Type}$, TDP-43^{Q331K}, or TDP-43$^{Q331K\text{-}Low}$ transgenic mice. **B**. TDP-43 regulated cassette exons identified from TDP-43 depletion experiments in mouse (Polymenidou et al., 2011). Cassette exons are divided into direct targets, bound by TDP-43 within 2 kilobases, and indirect targets, not bound by TDP-43. **C**. Bar plot displaying the fraction of overlap of TDP-43-regulated exons, direct (dark green) and indirect (light green), with exons that changed upon expression of TDP-43$^{Wild\text{-}Type}$, TDP-43^{Q331K} and TDP-43$^{Q331K\text{-}low}$. The TDP-43^{Q331K} and TDP-43$^{Q331K\text{-}low}$ overlapping exons show a significant enrichment (by chi-square analysis: $p<0.0005$) in the fraction that are direct TDP-43 regulated exons compared to the fraction that are indirect TDP-43 regulated exons. No enrichment was found in TDP-43$^{Wild\text{-}Type}$. **D**. Overlap of significantly changed cassette exons observed in the cortices from TDP-43 transgenic mice with TDP-43 regulated cassette exons. Direct and indirect exons that are also regulated in TDP-43 transgenic mice are represented as exons that are changed in the opposite (colored gray) or same (included or excluded in both, colored white) direction of TDP-43 depletion. **E-F**. RT-PCR validation for a subset of alternate cassette exons identified in cortex by splicing sensitive microarrays and that are (**E**) mutant-dependent or (**F**) transgene dose-dependent and that change in the same or opposite direction following knockdown (KD) in TDP-43 following anti-sense oligonucleotide infusion within striatum. Bar plots show the ratio of inclusion to exclusion calculated from the mean intensities ($n\geq3$ biological replicates, ± S.D). Representative gel analyses are images shown with duplicate biological replicates. *$p<0.05$, **$p<0.01$ and ***$p<0.001$ with Student's T-test; "in/ex" = ratio of inclusion to exclusion. **G**. Bar graph depicting the percentages of differentially included or excluded exons upon human TDP-43 expression (green dashed line). The fraction of unchanged exons in any cortex experiment (TDP-43$^{Wild\text{-}Type}$, TDP-43^{Q331K}, TDP-43$^{Q331K\text{-}low}$) that have TDP-43 binding. (i) "All Changing Exons" includes all exons that were either included or excluded (25% of these were direct TDP-43 targets). (ii) "Overlapping Exons with TDP-43 KD" includes only exons that changed following TDP-43 depletion (i.e., exons regulated by endogenous TDP-43). A significant increase was found in the percentage of direct targets (~43% for TDP-43^{Q331K} and TDP-43$^{Wild\text{-}Type}$, and 53% for TDP-43$^{Q331K\text{-}low}$) for cassette exons that are excluded upon human TDP-43 expression, with only a modest ~20% of included exons that are direct targets. (iii) "Non-overlapping exons with TDP-43 KD" includes only exons not previously shown to be regulated by endogenous TDP-43 depletion, in which 39% of excluded exons in the TDP-43^{Q331K} animals were direct targets.


A
Cortex
Non-Tg
TDP-43Wild-Type
TDP-43Q331K
TDP-43Q331K-low
RNA extraction
Splicing Array vs. Non-Tg
TDP-43Wild-Type n=824
TDP-43Q331K n=1195
TDP-43Q331K-low n=208
Alt 3'
Alt 5'
Cassette
Alt end
Alt start
Mut.Exclusive
Retain intron
Twin cassette
B
Direct TDP-43 exons
Indirect TDP-43 exons
(n=175)
(n=562)
C
Direct TDP-43 exons
Indirect TDP-43 exons
* p < 0.0005
Fraction of overlap with TDP-43 regulated exons
WT
Q331K
Q331K-low
D
Overlap
Direct
Indirect
TDP-43Wild-Type (55 of 314)
TDP-43Q331K (126 of 533)
TDP-43Q331K-low (26 of 97)
Changed in opposite direction as TDP-43 depletion
Changed in same direction as TDP-43 depletion
G
TDP-43
Included
Excluded
i. All Changing Exons
ii. Overlapping Exons with TDP-43 depletion
iii. Non-Overlapping Exons with TDP-43 depletion
Q331K
Q331K-low
Wild-Type
Fraction of events that are direct
E
Mutant-Dependent
Opposite of depletion
Same direction as depletion
Eif4h
Taf1b
Kcnip2
Abhd14a
Non-Tg
WT
Q331K
Q331K -low
KD
in/ex
F
Transgene Dose-Dependent
Opposite of depletion
Same direction as depletion
Unchanged in depletion
Sort1
Kcnd3
Atxn2
Ahi1
Non-Transgenic
TDP-43Wild-Type
TDP-43Q331K
TDP-43Q331K-low
TDP-43 depletion
in/ex = ratio of inclusion to exclusion

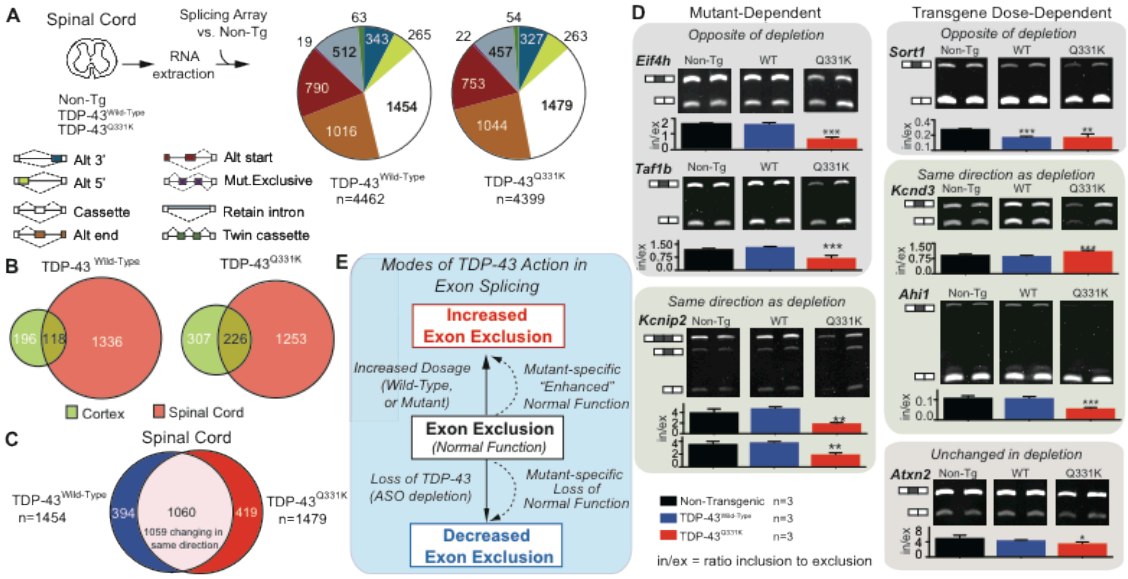


Figure 2.5: Unique splicing alterations in the spinal cord of TDP-43^{Q331K} mice include changes in genes involved in neurological function and transmission

A. Experimental strategy to identify differentially regulated splicing events in spinal cords of 2 month old non-transgenic, TDP-43$^{Wild\text{-}Type}$, or TDP-43^{Q331K} transgenic mice. Pie charts display total alternative splicing events significantly altered in the spinal cords of transgenic mice, relative to non-transgenic animals (colors representing the types of events on the array – defined at the right).
B. Diagram showing overlap between events changing in cortex and spinal cords of TDP-43$^{Wild\text{-}Type}$ and TDP-43^{Q331K} animals.
C. Overlap of spinal cord TDP-43$^{Wild\text{-}Type}$ and TDP-43^{Q331K} alternative cassette exons reveals a set of 1060 common splicing events, 1059 of which change in the same direction.
D. RT-PCR validation for a subset of spinal cord alternate cassette exons identified in (**A**). Bar plots show mean intensities of n≥3 biological replicates (± S.D). Representative gel images shown, with duplicate biological replicates.
E. Modes of TDP-43 Action in Exon Splicing.

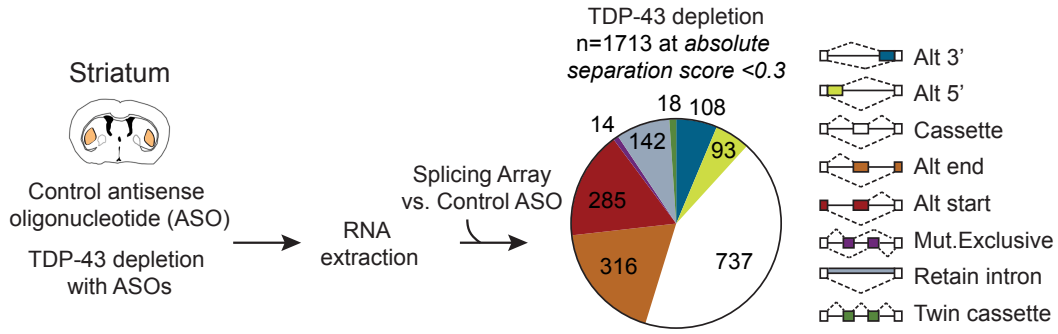


Figure 2.6: Experimental strategy for the re-analysis of alternative splicing events following TDP-43 depletion.

Experimental strategy to identify differentially regulated splicing events following depletion of mouse TDP-43 using anti-sense oligonucleotides targeted against mouse TDP-43, as described in a previous study from our group (Polymenidou et al., 2011a). RNA extracted from the striatum of mice treated with TDP-43 or Control ASOs was subjected to Affymetrix splicing sensitive microarray analysis. After comparison to control ASO splicing arrays, pie charts display the total alternative splicing events significantly altered in TDP-43 depleted mice, utilizing a commonly used statistical threshold of absolute separation score of <0.3. Pie chart colors represent the types of events represented on the array and are displayed schematically below.

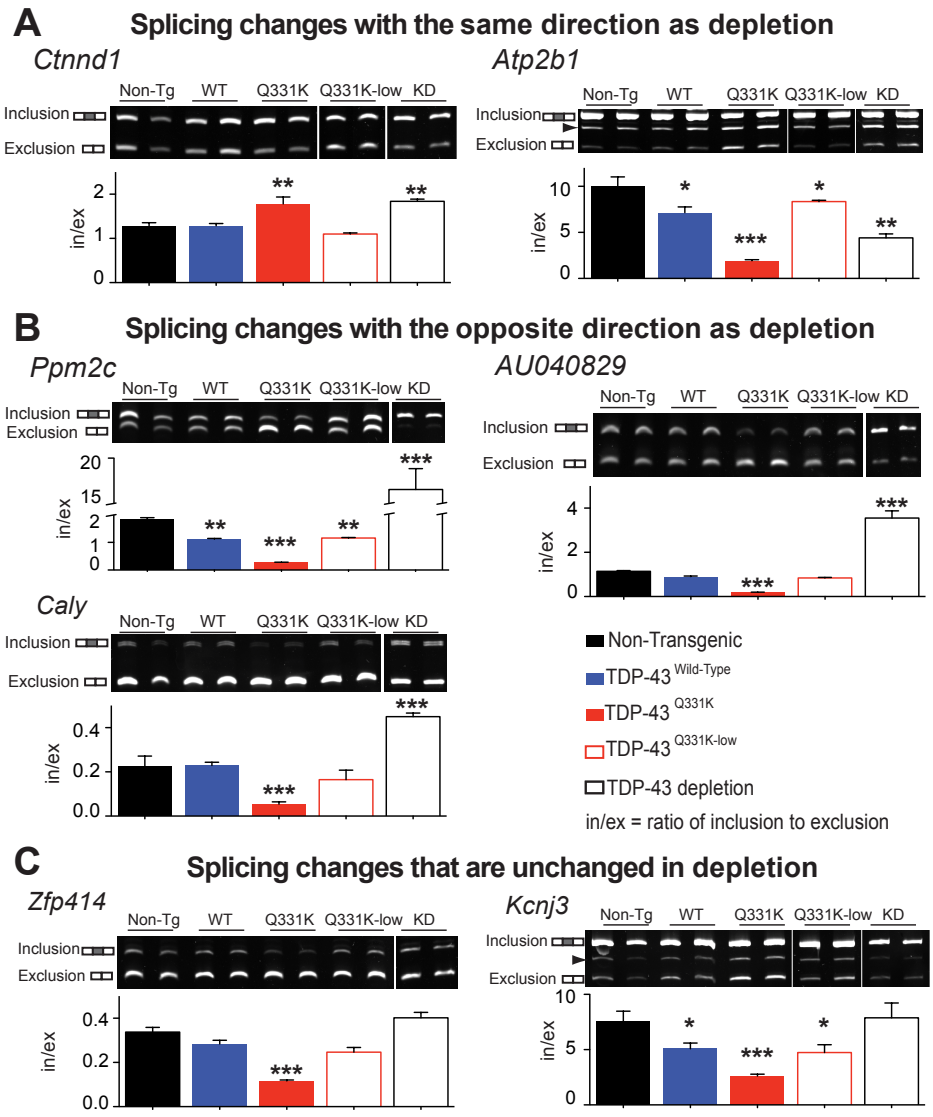


Figure 2.7: Validation of additional splicing changes detected by splicing-sensitive array.

Additional RT-PCR validations for alternately-spliced exons identified by splicing sensitive microarrays in the presence of the human TDP-43 transgene or following TDP-43 depletion in the brain. The directionality of splicing changes is compared with RNA from knockdown (KD) of TDP-43 in mouse striatum by anti-sense oligonucleotides.
A. Alternative exons of *Ctnnd1* and *Atp2b1* showed transgene dose-dependent changes in the same direction as depletion. For *Atp2b1*, the arrowhead indicates a nonspecific band or heterodimer of the two amplified splicing isoforms.
B. Alternative exons of *Ppm2c*, *AU040829*, and *Caly* shows targets that changed in the presence of the human TDP-43 transgene in a direction opposite to that following TDP-43 depletion in the brain. (C) Alternative exons of *Zfp414* and *Kcmj3* shows targets that changed in the presence of the TDP-43 human transgene but remained unchanged following TDP-43 depletion in the brain.

Table 2.1: Summary of Exon Changes in Cortices from TDP-43 Transgenic Mice from RT-PCR Confirmation Gels and Microarrays

	RT-PCR Confirmation Gels				Microarray Separation Score (compared to non-transgenic; absolute threshold <0.3)		
Target	Wild-Type	Q331K	Q331K-low	TDP-43 Depletion (Polymenidou et al., 2011b)	Wild-Type	Q331K	Q331K-low
Mutant Dependent							
Eif4h	NS	More Excluded*	More Excluded*	More Included	0	-1.521	-0.367
Taf1b	NS	More Excluded*	More Excluded	More Included	0	-2.02	-0.564
Kcnip2	Not Significant (NS)	More Excluded*	More Excluded*	More excluded	ND	ND	ND
Dose Dependent							
Sort1	More Excluded*	More Excluded*	More Excluded*	More Included	-0.946	-0.401	-0.802
Kcnd3	NS	More included*	NS	More included	0	0.357	0
Ahi1	NS	More Excluded	NS	More excluded	0	-1.982	0
Atxn2	NS	More Excluded	NS	not changed	0	-0.75	0
Ctnnd	NS	More included*	NS	More included	0	0.575	0

* = statistically significant by Student's T-test with $p<0.05$, compared to non-tg
ND= not determined

Key

	opposite direction as knockdown, suggestive of enhancement-of-function
	same direction as knockdown, suggestive of loss-of-function
	changed by transgene but unchanged in knockdown, , suggestive of enhancement-of-function

Table 2.2: Summary of Exon Changes in Spinal Cords from TDP-43 Transgenic Mice from RT-PCR Confirmation Gels and Microarrays

	RT-PCR Confirmation Gels			Microarray Separation Score (compared to non-transgenic; absolute threshold <0.3)	
Target	Wild-Type	Q331K	TDP-43 Depletion	Wild-Type	Q331K
Mutant Dependent					
Eif4h	NS	More Excluded*	More included	0	-1.259
Taf1b	NS	More Excluded*	More included	0	-1.341
Kcnip2	NS	More Excluded*	More excluded	ND	ND
Dose Dependent					
Sort1	More Excluded*	More Excluded*	More included	-0.752	-0.882
Kcnd3	NS	More included*	More included	0	0.34
Ahi1	NS	More excluded*	More excluded	0	0.34
Atxn2	NS	More Excluded*	not changed	0	-0.836

* = statistically significant by Student's T-test with $p<0.05$, compared to non-tg
ND= not determined

Key

	opposite direction as knockdown, suggestive of enhancement-of-function
	same direction as knockdown, suggestive of loss-of-function
	changed by transgene but unchanged in knockdown, , suggestive of enhancement-of-function

Table 2.3: Genes containing splicing changes in TDP-43^{Q331K} Spinal Cord involved in neurological processing and transmission

Gene Symbol	Gene Name	**Microarray Separation Score** (compared to non-transgenic; absolute threshold <0.3)
APBA2	amyloid beta A4 precursor protein-binding family	-0.526
CHRM1	muscarinic acetylcholine receptor M1	0.36
COCH	cochlin precursor	0.384
GAL3ST1	galactosylceramide sulfotransferase	0.381
GRIA2	glutamate receptor 2 isoform 3 precursor	-0.306
GTF3C2	general transcription factor 3C polypeptide 2	0.326
NR2F6	nuclear receptor subfamily 2 group F member 6	-0.303
NRXN1	neurixin-1-alpha	0.37
NRXN3	neurexin-3-alpha	-0.775
OLFR1033	olfactory receptor 1033	-0.408
OPA1	dynamin-like 120 kDa protein, mitochondrial	-0.313
PDE6D	retinal rod rhodopsin-sensitive cGMP	0.576
PICK1	PRKCA-binding protein	0.33
PPP3CA	calcineurin or serine/threonine-protein phosphatase 2B	-0.553
RIMS1	regulating synaptic membrane exocytosis protein	-0.574
SCRIB	protein scribble homolog	-0.584
STX3	syntaxin-3	0.316
UNC13B	protein unc-13 homolog B	0.332

2.6 References

Arai, T., Hasegawa, M., Akiyama, H., Ikeda, K., Nonaka, T., Mori, H., Mann, D., Tsuchiya, K., Yoshida, M., Hashizume, Y., and Oda, T. (2006). TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Biochem Biophys Res Commun *351*, 602-611.

Arnold, E.S., Ling, S.C., Huelga, S.C., Lagier-Tourenne, C., Polymenidou, M., Ditsworth, D., Kordasiewicz, H.B., McAlonis-Downes, M., Platoshyn, O., Parone, P.A., Da Cruz, S., Clutario, K.M., Swing, D., Tessarollo, L., Marsala, M., Shaw, C.E., Yeo, G.W., and Cleveland, D.W. (2013). ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. Proceedings of the National Academy of Sciences of the United States of America *110*, E736-745.

Ates, K., Yang, S.Y., Orrell, R.W., Sinanan, A.C., Simons, P., Solomon, A., Beech, S., Goldspink, G., and Lewis, M.P. (2007). The IGF-I splice variant MGF increases progenitor cells in ALS, dystrophic, and normal muscle. FEBS Lett *581*, 2727-2732.

Atlasi, Y., Mowla, S.J., Ziaee, S.A., Gokhale, P.J., and Andrews, P.W. (2008). OCT4 spliced variants are differentially expressed in human pluripotent and nonpluripotent cells. Stem Cells *26*, 3068-3074.

Ayala, Y.M., Pantano, S., D'Ambrogio, A., Buratti, E., Brindisi, A., Marchetti, C., Romano, M., and Baralle, F.E. (2005). Human, Drosophila, and C.elegans TDP43: nucleic acid binding properties and splicing regulatory function. J Mol Biol *348*, 575-588.

Buratti, E., Brindisi, A., Giombi, M., Tisminetzky, S., Ayala, Y.M., and Baralle, F.E. (2005). TDP-43 binds heterogeneous nuclear ribonucleoprotein A/B through its C-terminal tail: an important region for the inhibition of cystic fibrosis transmembrane conductance regulator exon 9 splicing. J Biol Chem *280*, 37572-37584.

Buratti, E., Dork, T., Zuccato, E., Pagani, F., Romano, M., and Baralle, F.E. (2001). Nuclear factor TDP-43 and SR proteins promote in vitro and in vivo CFTR exon 9 skipping. EMBO J *20*, 1774-1784.

Cao, D., Allan, R.W., Cheng, L., Peng, Y., Guo, C.C., Dahiya, N., Akhi, S., and Li, J. (2011). RNA-binding protein LIN28 is a marker for testicular germ cell tumors. Hum Pathol *42*, 710-718.

Carpenter, B., MacKay, C., Alnabulsi, A., MacKay, M., Telfer, C., Melvin, W.T., and Murray, G.I. (2006). The roles of heterogeneous nuclear ribonucleoproteins in tumour development and progression. Biochim Biophys Acta *1765*, 85-100.

Cauffman, G., Liebaers, I., Van Steirteghem, A., and Van de Velde, H. (2006). POU5F1 isoforms show different expression patterns in human embryonic stem cells and preimplantation embryos. Stem Cells *24*, 2685-2691.

Chen, M., Zhang, J., and Manley, J.L. (2010). Turning on a fuel switch of cancer: hnRNP proteins regulate alternative splicing of pyruvate kinase mRNA. Cancer Res *70*, 8977-8980.

Clower, C.V., Chatterjee, D., Wang, Z., Cantley, L.C., Vander Heiden, M.G., and Krainer, A.R. (2010). The alternative splicing repressors hnRNP A1/A2 and PTB influence pyruvate kinase isoform expression and cell metabolism. Proc Natl Acad Sci U S A *107*, 1894-1899.

D'Ambrogio, A., Buratti, E., Stuani, C., Guarnaccia, C., Romano, M., Ayala, Y.M., and Baralle, F.E. (2009). Functional mapping of the interaction between TDP-43 and hnRNP A2 in vivo. Nucleic Acids Res *37*, 4116-4126.

Da Cruz, S., and Cleveland, D.W. (2011). Understanding the role of TDP-43 and FUS/TLS in ALS and beyond. Curr Opin Neurobiol *21*, 904-919.

David, C.J., Chen, M., Assanah, M., Canoll, P., and Manley, J.L. (2010). HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. Nature *463*, 364-368.

David, C.J., and Manley, J.L. (2010). Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes Dev *24*, 2343-2364.

Du, H., Cline, M.S., Osborne, R.J., Tuttle, D.L., Clark, T.A., Donohue, J.P., Hall, M.P., Shiue, L., Swanson, M.S., Thornton, C.A., and Ares, M., Jr. (2010). Aberrant alternative splicing and extracellular matrix gene expression in mouse models of myotonic dystrophy. Nat Struct Mol Biol *17*, 187-193.

Gardina, P.J., Clark, T.A., Shimada, B., Staples, M.K., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S., Davies, C., Williams, A., and Turpaz, Y. (2006). Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics *7*, 325.

Gitcho, M.A., Baloh, R.H., Chakraverty, S., Mayo, K., Norton, J.B., Levitch, D., Hatanpaa, K.J., White, C.L., 3rd, Bigio, E.H., Caselli, R., Baker, M., Al-Lozi, M.T., Morris, J.C., Pestronk, A., Rademakers, R., Goate, A.M., and Cairns, N.J. (2008). TDP-43 A315T mutation in familial motor neuron disease. Ann Neurol *63*, 535-538.

Gopalakrishnan, S., Van Emburgh, B.O., Shan, J., Su, Z., Fields, C.R., Vieweg, J., Hamazaki, T., Schwartz, P.H., Terada, N., and Robertson, K.D. (2009). A novel DNMT3B splice variant expressed in tumor and pluripotent cells modulates genomic DNA methylation patterns and displays altered DNA binding. Mol Cancer Res *7*, 1622-1634.

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., Jr., and Yeo, G.W. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Rep *1*, 167-178.

Kabashi, E., Valdmanis, P.N., Dion, P., Spiegelman, D., McConkey, B.J., Vande Velde, C., Bouchard, J.P., Lacomblez, L., Pochigaeva, K., Salachas, F., Pradat, P.F., Camu, W., Meininger, V., Dupre, N., and Rouleau, G.A. (2008). TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. Nat Genet *40*, 572-574.

Kanadia, R.N., Shin, J., Yuan, Y., Beattie, S.G., Wheeler, T.M., Thornton, C.A., and Swanson, M.S. (2006). Reversal of RNA missplicing and myotonia after muscleblind overexpression in a mouse poly(CUG) model for myotonic dystrophy. Proc Natl Acad Sci U S A *103*, 11748-11753.

Karni, R., de Stanchina, E., Lowe, S.W., Sinha, R., Mu, D., and Krainer, A.R. (2007). The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nat Struct Mol Biol *14*, 185-193.

Kashima, T., Rao, N., David, C.J., and Manley, J.L. (2007). hnRNP A1 functions with specificity in repression of SMN2 exon 7 splicing. Hum Mol Genet *16*, 3149-3159.

Kuyumcu-Martinez, N.M., Wang, G.S., and Cooper, T.A. (2007). Increased steady-state levels of CUGBP1 in myotonic dystrophy 1 are due to PKC-mediated hyperphosphorylation. Mol Cell *28*, 68-78.

Kwiatkowski, T.J., Jr., Bosco, D.A., Leclerc, A.L., Tamrazian, E., Vanderburg, C.R., Russ, C., Davis, A., Gilchrist, J., Kasarskis, E.J., Munsat, T., Valdmanis, P., Rouleau, G.A., Hosler, B.A., Cortelli, P., de Jong, P.J., Yoshinaga, Y., Haines, J.L., Pericak-Vance, M.A., Yan, J., Ticozzi, N., Siddique, T., McKenna-Yasek, D., Sapp, P.C., Horvitz, H.R., Landers, J.E., and Brown, R.H., Jr. (2009). Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. Science *323*, 1205-1208.

Lagier-Tourenne, C., Polymenidou, M., and Cleveland, D.W. (2010). TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. Hum Mol Genet *19*, R46-64.

Lapuk, A., Marr, H., Jakkula, L., Pedro, H., Bhattacharya, S., Purdom, E., Hu, Z., Simpson, K., Pachter, L., Durinck, S., Wang, N., Parvin, B., Fontenay, G., Speed, T., Garbe, J., Stampfer, M., Bayandorian, H., Dorton, S., Clark, T.A., Schweitzer, A., Wyrobek, A., Feiler, H., Spellman, P., Conboy, J., and Gray, J.W. (2010). Exon-level microarray analyses identify alternative splicing programs in breast cancer. Mol Cancer Res *8*, 961-974.

Lee, J., Kim, H.K., Rho, J.Y., Han, Y.M., and Kim, J. (2006). The human OCT-4 isoforms differ in their ability to confer self-renewal. J Biol Chem *281*, 33554-33565.

Lefebvre, S., Burglen, L., Reboullet, S., Clermont, O., Burlet, P., Viollet, L., Benichou, B., Cruaud, C., Millasseau, P., Zeviani, M., and et al. (1995). Identification and characterization of a spinal muscular atrophy-determining gene. Cell *80*, 155-165.

Lin, H., Shabbir, A., Molnar, M., Yang, J., Marion, S., Canty, J.M., Jr., and Lee, T. (2008). Adenoviral expression of vascular endothelial growth factor splice variants

differentially regulate bone marrow-derived mesenchymal stem cells. J Cell Physiol *216*, 458-468.

Liscic, R.M., Grinberg, L.T., Zidar, J., Gitcho, M.A., and Cairns, N.J. (2008). ALS and FTLD: two faces of TDP-43 proteinopathy. Eur J Neurol *15*, 772-780.

Liu-Yesucevitz, L., Bilgutay, A., Zhang, Y.J., Vanderweyde, T., Citro, A., Mehta, T., Zaarur, N., McKee, A., Bowser, R., Sherman, M., Petrucelli, L., and Wolozin, B. (2010). Tar DNA binding protein-43 (TDP-43) associates with stress granules: analysis of cultured cells and pathological brain tissue. PLoS One *5*, e13250.

Lomen-Hoerth, C., Anderson, T., and Miller, B. (2002). The overlap of amyotrophic lateral sclerosis and frontotemporal dementia. Neurology *59*, 1077-1079.

Lopez-Bigas, N., Audit, B., Ouzounis, C., Parra, G., and Guigo, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett *579*, 1900-1903.

Melko, M., and Bardoni, B. (2010). The role of G-quadruplex in RNA metabolism: involvement of FMRP and FMR2P. Biochimie *92*, 919-926.

Misquitta-Ali, C.M., Cheng, E., O'Hanlon, D., Liu, N., McGlade, C.J., Tsao, M.S., and Blencowe, B.J. (2011). Global profiling and molecular characterization of alternative splicing events misregulated in lung cancer. Mol Cell Biol *31*, 138-150.

Moss, E.G., Lee, R.C., and Ambros, V. (1997). The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. Cell *88*, 637-646.

Mukherjee, N., Corcoran, D.L., Nusbaum, J.D., Reid, D.W., Georgiev, S., Hafner, M., Ascano, M., Jr., Tuschl, T., Ohler, U., and Keene, J.D. (2011). Integrative Regulatory Mapping Indicates that the RNA-Binding Protein HuR Couples Pre-mRNA Processing and mRNA Stability. Mol Cell.

Neumann, M., Sampathu, D.M., Kwong, L.K., Truax, A.C., Micsenyi, M.C., Chou, T.T., Bruce, J., Schuck, T., Grossman, M., Clark, C.M., McCluskey, L.F., Miller, B.L., Masliah, E., Mackenzie, I.R., Feldman, H., Feiden, W., Kretzschmar, H.A., Trojanowski, J.Q., and Lee, V.M. (2006). Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Science *314*, 130-133.

Newman, M.A., and Hammond, S.M. (2010). Lin-28: an early embryonic sentinel that blocks Let-7 biogenesis. Int J Biochem Cell Biol *42*, 1330-1333.

Ou, S.H., Wu, F., Harrich, D., Garcia-Martinez, L.F., and Gaynor, R.B. (1995). Cloning and characterization of a novel cellular protein, TDP-43, that binds to human immunodeficiency virus type 1 TAR DNA sequence motifs. J Virol *69*, 3584-3596.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet *40*, 1413-1415.

Passini, M.A., Bu, J., Richards, A.M., Kinnecom, C., Sardi, S.P., Stanek, L.M., Hua, Y., Rigo, F., Matson, J., Hung, G., Kaye, E.M., Shihabuddin, L.S., Krainer, A.R., Bennett, C.F., and Cheng, S.H. (2011). Antisense oligonucleotides delivered to the mouse CNS ameliorate symptoms of severe spinal muscular atrophy. Science translational medicine *3*, 72ra18.

Patel, N.A., Song, S.S., and Cooper, D.R. (2006). PKCdelta alternatively spliced isoforms modulate cellular apoptosis in retinoic acid-induced differentiation of human NT2 cells and mouse embryonic stem cells. Gene Expr *13*, 73-84.

Patrakitkomjorn, S., Kobayashi, D., Morikawa, T., Wilson, M.M., Tsubota, N., Irie, A., Ozawa, T., Aoki, M., Arimura, N., Kaibuchi, K., Saya, H., and Araki, N. (2008). Neurofibromatosis type 1 (NF1) tumor suppressor, neurofibromin, regulates the neuronal differentiation of PC12 cells via its associating protein, CRMP-2. J Biol Chem *283*, 9399-9413.

Pizzi, M., Benarese, M., Boroni, F., Goffi, F., Valerio, A., and Spano, P.F. (2000). Neuroprotection by metabotropic glutamate receptor agonists on kainate-induced degeneration of motor neurons in spinal cord slices from adult rat. Neuropharmacology *39*, 903-910.

Polymenidou, M., Lagier-Tourenne, C., Hutt, K.R., Huelga, S.C., Moran, J., Liang, T.Y., Ling, S.C., Sun, E., Wancewicz, E., Mazur, C., Kordasiewicz, H., Sedaghat, Y., Donohue, J.P., Shiue, L., Bennett, C.F., Yeo, G.W., and Cleveland, D.W. (2011a). Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nature neuroscience *14*, 459-468.

Polymenidou, M., Lagier-Tourenne, C., Hutt, K.R., Huelga, S.C., Moran, J., Liang, T.Y., Ling, S.C., Sun, E., Wancewicz, E., Mazur, C., Kordasiewicz, H., Sedaghat, Y., Donohue, J.P., Shiue, L., Bennett, C.F., Yeo, G.W., and Cleveland, D.W. (2011b). Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nat Neurosci *14*, 459-468.

Sreedharan, J., Blair, I.P., Tripathi, V.B., Hu, X., Vance, C., Rogelj, B., Ackerley, S., Durnall, J.C., Williams, K.L., Buratti, E., Baralle, F., de Belleroche, J., Mitchell, J.D., Leigh, P.N., Al-Chalabi, A., Miller, C.C., Nicholson, G., and Shaw, C.E. (2008). TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. Science *319*, 1668-1672.

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., and Ares, M., Jr. (2006). Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLoS Comput Biol *2*, e4.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861-872.

Van Damme, P., Bogaert, E., Dewil, M., Hersmus, N., Kiraly, D., Scheveneels, W., Bockx, I., Braeken, D., Verpoorten, N., Verhoeven, K., Timmerman, V., Herijgers, P.,

Callewaert, G., Carmeliet, P., Van Den Bosch, L., and Robberecht, W. (2007). Astrocytes regulate GluR2 expression in motor neurons and their vulnerability to excitotoxicity. Proc Natl Acad Sci U S A *104*, 14825-14830.

Van Damme, P., Braeken, D., Callewaert, G., Robberecht, W., and Van Den Bosch, L. (2005). GluR2 deficiency accelerates motor neuron degeneration in a mouse model of amyotrophic lateral sclerosis. J Neuropathol Exp Neurol *64*, 605-612.

Vance, C., Rogelj, B., Hortobagyi, T., De Vos, K.J., Nishimura, A.L., Sreedharan, J., Hu, X., Smith, B., Ruddy, D., Wright, P., Ganesalingam, J., Williams, K.L., Tripathi, V., Al-Saraj, S., Al-Chalabi, A., Leigh, P.N., Blair, I.P., Nicholson, G., de Belleroche, J., Gallo, J.M., Miller, C.C., and Shaw, C.E. (2009). Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. Science *323*, 1208-1211.

Venables, J.P., Klinck, R., Koh, C., Gervais-Bird, J., Bramard, A., Inkel, L., Durand, M., Couture, S., Froehlich, U., Lapointe, E., Lucier, J.F., Thibault, P., Rancourt, C., Tremblay, K., Prinos, P., Chabot, B., and Elela, S.A. (2009). Cancer-associated regulation of alternative splicing. Nature structural & molecular biology *16*, 670-676.

Viswanathan, S.R., Powers, J.T., Einhorn, W., Hoshida, Y., Ng, T.L., Toffanin, S., O'Sullivan, M., Lu, J., Phillips, L.A., Lockhart, V.L., Shah, S.P., Tanwar, P.S., Mermel, C.H., Beroukhim, R., Azam, M., Teixeira, J., Meyerson, M., Hughes, T.P., Llovet, J.M., Radich, J., Mullighan, C.G., Golub, T.R., Sorensen, P.H., and Daley, G.Q. (2009). Lin28 promotes transformation and is associated with advanced human malignancies. Nature genetics *41*, 843-848.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

Wang, H.Y., Wang, I.F., Bose, J., and Shen, C.K. (2004). Structural diversity and functional implications of the eukaryotic TDP gene family. Genomics *83*, 130-139.

West, J.A., Viswanathan, S.R., Yabuuchi, A., Cunniff, K., Takeuchi, A., Park, I.H., Sero, J.E., Zhu, H., Perez-Atayde, A., Frazier, A.L., Surani, M.A., and Daley, G.Q. (2009). A role for Lin28 in primordial germ-cell development and germ-cell malignancy. Nature *460*, 909-913.

Wilbert, M.L., Huelga, S.C., Kapeli, K., Stark, T.J., Liang, T.Y., Chen, S.X., Yan, B.Y., Nathanson, J.L., Hutt, K.R., Lovci, M.T., Kazan, H., Vu, A.Q., Massirer, K.B., Morris, Q., Hoon, S., and Yeo, G.W. (2012). LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. Molecular cell *48*, 195-206.

Winton, M.J., Igaz, L.M., Wong, M.M., Kwong, L.K., Trojanowski, J.Q., and Lee, V.M. (2008). Disturbance of nuclear and cytoplasmic TAR DNA-binding protein (TDP-43) induces disease-like redistribution, sequestration, and aggregate formation. J Biol Chem *283*, 13302-13309.

Yang, D.H., and Moss, E.G. (2003). Temporally regulated expression of Lin-28 in diverse tissues of the developing mouse. Gene Expr Patterns *3*, 719-726.

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R., Slukvin, II, and Thomson, J.A. (2007). Induced pluripotent stem cell lines derived from human somatic cells. Science *318*, 1917-1920.

3 Global splicing patterns regulated by 6 hnRNP proteins

3.1 Background

Nascent transcripts produced by RNA polymerase II in eukaryotic cells are subject to extensive processing prior to the generation of a mature messenger RNA (mRNA). In human cells, a sizable fraction of these precursor mRNAs (pre-mRNAs) is associated with heterogeneous nuclear ribonucleoparticle (hnRNP) proteins, forming large hnRNP-RNA complexes (Dreyfuss et al., 1993). These complexes can contain any combination of at least 20 major hnRNP proteins named hnRNP A1 through hnRNP U. Rivaling histones in their abundance, hnRNP proteins assemble on pre-mRNAs in a combinatorial arrangement dictated by their relative abundances and recognition of specific RNA sequences (Dreyfuss et al., 2002).

HnRNP proteins are multifunctional, participating in all crucial aspects of RNA processing, including pre-mRNA splicing, mRNA export, localization, translation and stability (Han et al., 2010). Of particular interest is the increasingly evident impact of hnRNP proteins on the regulation of alternative splicing (AS), where splice sites within a pre-mRNA are preferentially selected or repressed to produce different mRNA isoforms. The process of AS contributes drastically to proteome diversity, with recent estimates that greater than 90% of human genes can undergo AS (Wang et al., 2008). The proper control of AS is essential for cellular development and homeostasis. Consequently, misregulation of specific AS events has been shown to result in several diseases, such as spinal muscular atrophy (Kashima and Manley, 2003), and global misregulation of AS has been shown to occur in multiple forms of cancer (Gardina et al., 2006; Lapuk et al., 2010; Misquitta-Ali et al., 2011). In addition, hnRNP proteins are misregulated in

particular cancer cell types and have been shown to directly affect AS events that promote cancer (David et al., 2010; Golan-Gerstl et al., 2011).

Most hnRNP proteins have been implicated in AS, however, a majority of these reports were based on single-gene studies and *in vitro* assays. The first systematic AS analysis for the hnRNP protein family used siRNAs against 14 of the major hnRNP proteins in three different human cell lines, followed with an RT-PCR screen restricted to 56 AS events (Venables et al., 2008). Recent advances in genome-wide technologies have allowed for more global assessments of AS. Progress has been made using splicing-sensitive microarrays or RNA sequencing to identify genome-wide AS events (Katz et al., 2010) and using biochemical approaches such as iCLIP (Konig et al., 2010) and CLIP-seq (Katz et al., 2010) to identify binding sites for hnRNP C and for hnRNP H1, respectively. In Drosophila, genome-wide AS changes were identified for the orthologous hnRNP A/B family and, combined with immunoprecipitation strategies, revealed that the 4 hnRNP family members tested bound and regulated a set of overlapping, but distinct target RNAs (Blanchette et al., 2009). Nevertheless, it is unknown whether this mode of co-regulation is conserved in humans.

In this study we used genome-wide approaches to identify hnRNP-regulated AS events in human cells and examined the impact of their regulation. Using siRNAs to deplete the protein levels of 6 major hnRNP proteins, A1, A2/B1, F, H1, M, and U, and splicing-sensitive human exon-junction microarrays (Affymetrix HJAY) we have identified thousands of hnRNP-dependent exons. A battery of analyses demonstrated that the majority of hnRNP-dependent AS events are regulated by multiple hnRNP proteins and depletion of individual members of a set of hnRNP proteins acts to cause similar sets of AS changes. Using antibodies specific for hnRNP proteins A1, A2/B1, F, M and U, we performed cross-linking and immunoprecipitation followed by high throughput

sequencing (CLIP-seq or HITS-CLIP) to identify direct binding sites that contribute to the regulation of AS events. Combination of our data revealed six new interrelated RNA splicing maps for the major hnRNP proteins. Finally, RNA sequencing (RNA-seq) was used to show that hnRNP proteins regulate a significant subset of RNA binding proteins (RBPs) and cancer-associated genes, implicating hnRNP proteins in the regulation of other aspects of RNA processing and general pathways in cancer.

3.2 Results

3.2.1 Identification of thousands of hnRNP-dependent human AS events

To identify genome-wide alternative splicing (AS) events regulated by hnRNP proteins, we reduced the protein levels of hnRNP proteins A1, A2/B1, F, H1, M and U to 8-28% of their endogenous levels in human 293T cells (Figure 3.1A). Splicing-sensitive microarrays were used to discover AS events that change significantly upon depletion of specific hnRNP proteins relative to a non-targeting control siRNA (Figure 3.1B). Hierarchical clustering of normalized intensities for all microarrays showed that the triplicate experiments for specific hnRNP depletions group together appropriately (Figure 3.7A). Using a published analysis procedure (Sugnet et al., 2006), we identified 6,555 altered AS events upon hnRNP depletion, representing 4,638 human genes. This widespread regulation of AS events is consistent with the reported splicing regulatory roles of hnRNP proteins A1, A2/B1, F, H1 and M (Figure 3.1C). Depletion of the well-studied splicing regulator hnRNP A1 affects the largest number of AS events (2,576). Surprisingly, depletion of hnRNP U, which previously showed little evidence for splicing regulation (Martinez-Contreras et al., 2007), affects almost as many AS events (2,418) as hnRNP A1. As hnRNP proteins affected different types of AS in similar ratios (Figure 3.1C), we focused our analysis on the most represented category, cassette exons,

where a single exon was either included or skipped from a mature mRNA transcript. In total, the inclusion of 2,586 cassette exons were affected by hnRNP protein depletion, ranging from 546 exons for hnRNP F to 1,116 exons for hnRNP U. Thus far, this dataset contains thousands more human AS events whose splicing regulation is dependent on hnRNP proteins than previously published hnRNP studies (Katz et al., 2010; Konig et al., 2010; Llorian et al., 2010; Venables et al., 2008; Xue et al., 2009).

3.2.2 HnRNP proteins activate and repress cassette exons

To provide insight into whether hnRNP proteins activate or repress exon recognition, we determined what fraction of exons showed increased inclusion (or exclusion) when a particular hnRNP was depleted, a response expected for exons repressed (or activated) by the regulator. By this criterion, hnRNP proteins A1 and A2/B1 repress the largest fractions of cassette exons (Figure 3.1D). This is consistent with splicing repression by hnRNP A1 and A2/B1 (Martinez-Contreras et al., 2007); however, their presence also appears necessary for exon inclusion. In fact, hnRNP proteins F, H1, M, and U appear to activate exon inclusion for a majority of the cassette exons they affect (Figure 3.1D). The same trend is also observed for other types of AS events, where after depletion, more than half of the events dependent on hnRNP F, H1, M and U show increased skipping, and more than half of the events dependent on hnRNP A1 and A2/B1 show increased inclusion. Semi-quantitative RT-PCR experiments successfully validated 89% of the 84 cassette exon events tested, a rate consistent with previous reports using this platform (Du et al., 2010; Polymenidou et al., 2011; Sugnet et al., 2006) (See Figure 3.1E for select hnRNP A1-dependent events and Figure 3.7B-G for additional events tested). Splicing changes in 293T cells were reproduced when we depleted hnRNP A2/B1 with an antisense oligonucleotide in primary human fibroblasts (Figure 3.7H-I), indicating that the regulation of many of these AS events is preserved in

primary human cells. Thus, we conclude that these 6 hnRNP proteins, act both as repressors and activators of exon inclusion, and are important for maintaining correct levels of inclusion of hundreds to thousands of differentially spliced AS events in the human transcriptome.

3.2.3 Multiple hnRNP proteins regulate a majority of AS events in a hierarchy of collaboration

The high abundance of hnRNP proteins in human cells suggests that a considerable level of multiplicity or redundancy is likely to exist to regulate RNA processing. Indeed, more than half (54%) of the 2,586 differentially spliced cassette exons were affected by more than 1 hnRNP protein (Figure 3.2A). Moreover, 47% of the 1,407 shared cassette exons were affected by 3 or more hnRNP proteins (Figure 3.2B), recapitulated also in the other types of AS events (Figure 3.2A). These global observations in human cells are confirmed by studies of individual AS events, such as the *c-src* cassette exon N1 which is regulated by hnRNP F, hnRNP H1, PTB (hnRNP I), and hnRNP A1 (Black, 2003), as well as a study of the hnRNP A/B subfamily in a Drosophila cell-line (Blanchette et al., 2009). Using thousands of AS events, our findings demonstrate that hnRNP proteins act in a concerted manner at many events in human cells.

We next set out to identify relationships between hnRNP proteins and the sets of cassette exons they regulate. Hierarchical clustering across regulated cassette exon sets revealed that activities of hnRNP A1 and M are the most distinct among the 6 hnRNP proteins (Figure 3.2C, dendrogram). To examine each relationship, we counted the percent of cassette exons regulated independently by pairs of hnRNP proteins, termed “overlapping events”. To illustrate, 29% of cassette exons affected by depletion of hnRNP M are also affected by depletion of hnRNP A2/B1 (Figure 3.2C row 6, column

1, sum of percentages), and correspondingly 19% of cassettes affected by depletion of hnRNP A2/B1 are also affected by M (Figure 3.2C row 1, column 6, sum of percentages). HnRNP F and H1 have very similar protein domain structures and bind to the same motif (Caputi and Zahler, 2001), which correlates with their significant fraction of overlapping events (206 total; $P < 0.05$, hypergeometric test). HnRNP A1 and A2/B1 also have similar protein domain structures; however, we only observed a modest fraction of overlapping events (330 total; Figure 3.2C). Instead, we observed significant fractions of overlapping events regulated between hnRNP proteins A2/B1 and H1, A2/B1 and F, and F and U (Figure 3.2C solid lined circles). These results identify novel relationships between hnRNP proteins in AS regulation.

To determine whether pairs of hnRNP proteins might act in concert or antagonistically, we noted the direction of change generated by each pair for each of their overlapping events. We found that most changed in the same direction (76% of targets, on average, across all pair-wise comparisons). In fact, a significant fraction of the changes in 11 of the 15 pair-wise comparisons of hnRNP-regulated cassette exons were in the same direction ($P < 0.002$, hypergeometric test; Figure 3.2C, highlighted in yellow), implying that multiple hnRNP proteins generally act in concert to modulate exon inclusion. Notably, 62% of the overlapping events between hnRNP F and H1 and 73% of the overlapping events between hnRNP F and U were activated upon depletion of either protein. RT-PCR confirmed 14 examples of cooperative AS regulation by hnRNP F and U, many with direct evidence for binding for the hnRNP proteins in the proximity of the exon (Figure 3.2D; Figure 3.8A). Of all pair-wise comparisons, hnRNP A1 and M regulated a more distinct set of cassette exons in comparison with the other hnRNP proteins, and they showed fewer overlapping events in the same direction (Figure 3.2C). We also validated 3 events where hnRNP A1 acted antagonistically to hnRNP U (Figure

3.8B), 2 events where hnRNP A1 and M acted similarly (Figure 3.8C), and 1 event where hnRNP A1 and M acted similarly, but antagonistically to hnRNP U (Figure 3.2E). We conclude that indirectly or directly, hnRNP proteins A2/B1, H1, F and U act cooperatively to regulate AS in similar ways, while hnRNP proteins A1 and M influence changes in opposition from the other hnRNP proteins.

3.2.4 Mapping sites of physical interaction between 6 hnRNP proteins and ~10,000 pre-mRNAs

The complex regulation observed above could be achieved by a number of indirect mechanisms. To determine whether joint regulation by hnRNP proteins involves direct binding by multiple proteins, we identified RNA binding sites of the hnRNP proteins using CLIP-seq. HnRNP proteins have been shown to bind degenerate RNA sequences that are relatively short (~4-5 nucleotides in length), rendering identification of binding sites by solely bioinformatic means unreliable. Instead, CLIP employs immunoprecipitation of cross-linked protein-RNA complexes to purify and sequence RNA regions bound to the protein of interest, and despite inherent limitations of specificity common to immunoprecipitation based methods, CLIP has been successfully used to identify direct binding sites across many cell-types, tissues and organisms (Chi et al., 2009; Hafner et al., 2010; Konig et al., 2010; Licatalosi et al., 2008; Mukherjee et al., 2011; Polymenidou et al., 2011; Sanford et al., 2009; Yeo et al., 2009). We performed CLIP-seq for hnRNP A1, A2/B1, F, M, and U (Figure 3.9A, Table 3.1), yielding comparable results to data generated for hnRNP H1 (Katz et al., 2010), which we incorporated with our own to generate a complete set of CLIP-seq data for the hnRNP proteins of interest.

A total of 85,921 CLIP-derived clusters (CDCs) were identified, representing binding sites for 6 hnRNP proteins in 10,557 protein-coding genes (Table 3.1). These

binding sites are distributed uniformly across pre-mRNA sequences, with the exception of hnRNP M, which shows a 5′ bias (Figure 3.9B). HnRNP binding sites are predominantly intronic, occurring within 500 nt of an exon, as expected for splicing regulators, but also in distal intronic regions (Figure 3.3A, pie charts). Accounting for the total length of the region reveals enrichment for binding within exonic regions and 3′ untranslated regions (3′ UTRs) for hnRNP A1, A2/B1, F, and U (Figure 3.3A, bar plots). The enrichment is consistent with the continuous shuttling of most hnRNP proteins from the nucleus to the cytoplasm and can mediate gene regulation via other RNA processing events such as transport and stability (Dreyfuss et al., 2002). HnRNP M demonstrates a stronger preference for binding in distal intronic regions, resembling that of another member of the hnRNP family, *TDP-43,* in mouse brain (Polymenidou et al., 2011). The distributions of hnRNP binding sites in different genic regions indicate that hnRNP proteins, in particular hnRNP M, are likely to have additional RNA processing functions other than splicing regulation.

3.2.5 *In vivo* hnRNP sequence motifs and enrichment near regulated exons

HnRNP proteins frequently recognize degenerate RNA sequences, as determined from affinity purification and SELEX (systematic evolution of ligands by exponential enrichment) experiments to identify the *in vitro* binding motif of the proteins (Burd and Dreyfuss, 1994; Swanson and Dreyfuss, 1988). The HOMER algorithm (Heinz et al., 2010) was applied to each set of hnRNP CDCs to discover the *in vivo* binding motifs of each hnRNP protein (two representatives of the four most significantly enriched motifs in Figure 3.9C are depicted in Figure 3.3A). For hnRNP A1, the predicted UAG(G) motif resembles the well-characterized UAGG hnRNP A1 binding site (Hutchison et al., 2002). For hnRNP A2/B1, not only did we recover the UAG sequence among the most significantly enriched motifs, but we also identified a G/A-rich motif similar to the

previously published UAGRGA motif (Hutchison et al., 2002). HnRNP F binds fairly G-rich sequences similar to the known *in vitro* GGGA motif (Caputi and Zahler, 2001). Additional G-rich motifs with interspersed uridines and adenosines are also enriched for hnRNP F, which is consistent with the motifs identified in SELEX experiments coupled with high-throughput sequencing (Zefeng Wang, personal communication). HnRNP H1 binds G-rich sequences interspersed with adenosines, which is very similar to its known GGGA motif (Caputi and Zahler, 2001; Katz et al., 2010). HnRNP M binds to GU-rich sequences that often contain a UU sequence within the motif, collaborating with *in vitro* observations that hnRNP M interacts with poly-U and poly-G sequences (Datar et al., 1993). HnRNP U CDCs are enriched for GU-rich sequences, also in agreement with a previous report showing that poly-U and poly-G sequences were the preferred *in vitro* binding sequences for hnRNP U (Kiledjian and Dreyfuss, 1992).

Next, we studied hnRNP binding sites near all annotated exons categorized as either constitutive or AS based on available expressed sequence tag (EST) or mRNA transcript evidence. A distance of 2 kilobases (kb) up and downstream of the exon was chosen to capture the hnRNP interactions with the most potential to affect splicing. While individual hnRNP CDCs are associated with 5% of constitutive exons on average, a small but significantly higher 7% on average are observed for AS exons for each of the 6 hnRNP proteins ($P < 10^{-28}$, Fisher's exact test) (Figure 3.3B). Consistent with our finding that hnRNP proteins activate and also repress exon recognition, these results reveal that hnRNP proteins directly associate with constitutive, but more preferentially AS exons in human cells. Focusing on cassette exons that were hnRNP-activated (skipped upon hnRNP depletion) or hnRNP-repressed (included upon hnRNP depletion), we observed that an average of 11% of cassettes harbor a CDC of the regulating hnRNP within 2kb (Figure 3.3C; Figure 3.9D). Although it is likely that not all binding sites are retrieved by

the CLIP procedure, we considered this subset of cassette exons direct targets of hnRNP proteins. Repeating our analysis with the quartile of cassette exons that change most strongly, corresponding to exons that are more uniquely regulated by a single hnRNP protein (Figure 3.9E), did not alter the fraction of direct targets (Figure 3.9F). In agreement with our pair-wise comparisons of microarray-detected AS cassette exons, we observe that hnRNP A1 and M more frequently bind near exons that are included upon depletion ($P < 0.05$, Fisher's exact test), implicating these proteins in direct repression of exon inclusion (Figure 3.3C). In contrast, hnRNP proteins A2/B1, F, H1 and U appear to preferentially bind near exons skipped upon their depletion, implying that these proteins directly activate exon inclusion.

We also considered what fraction of a particular set of hnRNP-regulated cassettes contained binding sites for the other hnRNP proteins within 2kb of the event. Significant CDCs for any of the 6 hnRNP proteins fall within 2kb of 41% of regulated exons on average (Figure 3.9G), and CLIP-seq reads for any of the 6 hnRNP proteins fall within 2kb of 95% of regulated exons on average (Figure 3.9H). Furthermore, we observe that binding of a particular hnRNP to the exons that change upon its depletion is not significantly enriched over binding to the same exons by the other hnRNP proteins (Table 3.2). One notable exception to this observation is hnRNP M, which shows a 4-fold enrichment of binding to its regulated events over binding of other hnRNP proteins to hnRNP M regulated events. Interestingly, we also see a 2-fold enrichment of binding by hnRNP M and hnRNP A2/B1 to hnRNP A1-regulated events. These observations demonstrate that hnRNP proteins are likely in a dynamic but overlapping complex that cooperate highly in the regulation of AS.

3.2.6 RNA splicing maps revealed for hnRNP proteins

To determine whether a pattern of position-dependent binding of the hnRNP proteins is associated with some aspect of AS, we computed the fraction of hnRNP-activated or hnRNP-repressed cassette exons that had an overlapping CLIP-seq read at each flanking nucleotide position. A similar calculation was used to determine the fraction of binding surrounding exons that remained unchanged during hnRNP depletion. A non-parametric statistical test was used to identify positions that were significantly enriched for hnRNP binding (Figure 3.4, black dots denote positions $P < 0.05$, chi-square test). This analysis revealed that each hnRNP had a unique pattern of binding around exons they directly regulate. For example, hnRNP A1 significantly bound the 3′ end of exons that were included upon depletion of hnRNP A1 (i.e., repressed). This observation is consistent with previous reports that hnRNP A1 binds exonic regions to repress usage of a nearby 5′ splice site (Martinez-Contreras et al., 2007). In contrast, we have shown that hnRNP A2/B1 preferentially binds activated more than repressed exons (Figure 3.3C), but is found at precise locations within ~150 nucleotides up and downstream of its repressed exons (Figure 3.4), suggestive of a “looping out” model of repression as proposed for the case of the hnRNP A1 pre-mRNA (Hutchison et al., 2002; Martinez-Contreras et al., 2006). HnRNP M binds proximal to and within exons to repress splicing (Figure 3.3C, Figure 3.4). HnRNP H1 binding is enriched within the flanking intronic regions of exons it activates, similar to a recently published study for hnRNP H1 (Katz et al., 2010). In contrast, we also observe hnRNP H1 binding to the exons it activates, consistent with previous findings that hnRNP H1 binds exons and activates exon inclusion (Caputi and Zahler, 2002; Chen et al., 1999). Unlike splicing factors that are more tissue-specific, such as the *NOVA* and *RBFOX* proteins, the RNA-maps for the hnRNP proteins appear less constrained proximal to the regulated exons

(Licatalosi et al., 2008; Yeo et al., 2009). We also find that hnRNP proteins are neither enriched near exons that are alternatively spliced in a tissue specific manner (Figure 3.10), nor are they enriched near exons that are predicted to be alternatively spliced in both human and mouse (Figure 3.10, last column). This shows that hnRNP proteins are not preferentially selected to specify very tissue-specific or evolutionarily conserved alternative splicing events. This likely reflects a more general control of RNA metabolism by ubiquitously expressed hnRNP proteins that are dynamically in complex with overlapping activities, including but not limited to AS.

3.2.7 HnRNP proteins regulate RNA binding proteins

Gene ontology analysis of 715 hnRNP RNA targets bound by all 6 hnRNP proteins reveals that RNA processing, in particular the molecular process of “RNA splicing” ($P = 2.63\text{x}10^{-8}$) and “mRNA metabolic process” ($P = 9.11\text{x}10^{-9}$) are highly enriched categories. Interestingly, inspection of the list of genes that have CLIP-seq clusters reveals that hnRNP genes are targets of hnRNP proteins. Specifically, hnRNP A1, A2/B1, C, F, H, I, K, L, M, Q, R, and U were bound by at least one of the 6 hnRNP proteins assayed, and all of the hnRNP proteins assayed directly associated with their own transcript (Figure 3.5A). The hnRNP proteins often bound to the 3′UTRs of hnRNP transcripts, which suggests auto-regulation of their own and regulation of other hnRNP transcripts.

To measure the transcript levels of hnRNP genes and other targets which were not all represented on the splicing microarrays, we subjected polyA-selected RNA from the cells depleted of hnRNP expression to strand-specific RNA sequencing (RNA-seq) (Parkhomchuk et al., 2009). For 25,921 annotated human genes, we computed the number of mapped reads per kilobase of exon, per million mapped reads (RPKM) as a normalized measure of gene expression in the hnRNP-depleted and the control siRNA

treated 293T cells (Mortazavi et al., 2008). The RPKM values reflected reduced levels of the hnRNP transcripts, consistent with significant down-regulation at their protein level upon siRNA treatment (Figure 3.5B). Interestingly, in several cases we observed significant upregulation of other hnRNP proteins upon depletion of a specific hnRNP protein, such as depletion of hnRNP F leading to an increase in hnRNP U protein levels. This increase in protein levels, which is unlikely a result of off-target siRNA effects that would downregulate mRNA levels, appears to be post-transcriptional as relatively little changes in mRNA levels of the hnRNP genes were observed (Figure 3.5B). Our findings strongly suggest that compensatory relationships exist, supporting the cooperativity of binding and regulation among the hnRNP proteins. The observation that these compensatory reactions are asymmetric (i.e. hnRNP F does not increase upon hnRNP U depletion) also reveals further complexity in their relationships. Furthermore, our results here imply that the alternative splicing changes we discovered using the microarrays are likely an underestimate of the AS events that are regulated by an individual hnRNP protein due to compensation between hnRNP proteins.

We next explored specific examples of auto- and cross-regulation of hnRNP proteins. HnRNP A1 has been shown to bind *in vitro* and regulate the skipping of its own cassette exon 7B (Blanchette and Chabot, 1999). The hnRNP A1 isoform containing exon 7B yields a protein with less splicing activity (Mayeda et al., 1994). Close to this exon in the hnRNP A1 transcript, we find an hnRNP U binding site (Figure 3.5A), suggesting that hnRNP U may also contribute to the splicing of exon 7B. Indeed, using RT-PCR to detect the splicing of exon 7B upon hnRNP U depletion confirms that hnRNP U represses the recognition of exon 7B (Figure 3.5C, $P < 0.002$, *t*-test). Interestingly, we also find that hnRNP A2/B1, F, and H1 repress this exon, perhaps indirectly, as there

was no significant binding evidence for these hnRNP proteins, nor hnRNP A1, near the event *in vivo* in these cells.

HnRNP A2/B1 has been shown to autoregulate its own expression levels by splicing of an intron within its 3′UTR that subsequently results in nonsense mediated decay (NMD) of the RNA transcript (McGlincy et al., 2010). Using a splicing-sensitive quantitative RT-PCR (qRT-PCR) strategy, we demonstrate that hnRNP A1 also contributes to the splicing of the hnRNP A2/B1 3′UTR, implicating hnRNP A1 in the regulation of hnRNP A2/B1 (Figure 3.5D, $P < 0.001$, *t*-test). This splicing event in the hnRNP A2/B1 3′UTR is directly bound only by hnRNP A2/B1 and hnRNP A1, indicating that the regulation is specific and direct (Figure 3.5A). Previous work has proposed that RBPs often facilitate splicing changes within other RBPs to yield NMD candidate transcripts as a mechanism of homeostatic control (Ni et al., 2007). These events can also occur through the inclusion of an intron or exon containing a premature termination codon (PTC). We also tested another NMD candidate event where an intron is retained in the hnRNP H1 transcript resulting in an in-frame PTC. This region is of particular interest because it is strongly bound by hnRNP A1, A2/B1, F, M, and U (Figure 3.5A). Using qRT-PCR we measured the ratio of the hnRNP H1 isoforms, one with the intron removed and the other with the intron retained. Our results reveal that hnRNP A1, A2/B1, M and U affect the inclusion of this intron in hnRNP H1 (Figure 3.5E, $P < 0.04$, *t*-test). Notably, the regulation of hnRNP H1 by hnRNP M correlates with an increase in H1 protein levels upon depletion of hnRNP M (Figure 3.5B). We conclude that in addition to previous literature that hnRNP proteins are subject to auto-regulation (Blanchette and Chabot, 1999; McGlincy et al., 2010; Rossbach et al., 2009), hnRNP proteins cross-regulate each other, adding to the complexity of the cooperative regulation of AS by the hnRNP proteins.

Aside from hnRNP genes, other splicing factors from the SR protein family such as SRSF1-7 were also direct targets, containing a CDC of at least 1 hnRNP protein assayed. To identify if other pre- or mRNA binding proteins were direct targets, we compiled a list of 443 genes containing RNA binding domains predicted to bind pre-mRNA and mRNA. Interestingly, 70% of the identified RBPs (311 of 443) are bound by at least one hnRNP protein ($P < 1.5x10^{-36}$, hypergeometric test). Therefore, hnRNP proteins interact with the RNAs of a significant proportion of all RBPs. Furthermore, this number of bound RBPs is likely an underestimate, as hnRNP binding for cell-specific or lowly abundant RBP transcripts may have been missed. HnRNP proteins also regulate a significant subset of the RNAs encoding RNA binding proteins (82 of the 311 to which they are bound), at either the expression or AS level as detected by splicing array ($P < 1.5x10^{-5}$, hypergeometric test) (Figure 3.6A). Intriguingly, we detect regulation of RBPs important for the microRNA pathway including *Drosha*, *Dicer1*, *EIF2C1 (Ago1)*, *EIF2C2 (Ago2)* and also other splicing factors including, PTBP1 and 2, SRSF5, NOVA1, and other hnRNP proteins D, DL, K, Q (Syncrip) and R. For RBP transcripts that are bound and regulated by multiple hnRNP proteins at the expression level, we consistently see that all hnRNP proteins tested affect the transcripts in the same direction, all causing down-regulation or all causing up-regulation. These findings demonstrated that hnRNP proteins have regulatory responsibilities, and likely play roles in managing mRNA levels of other RNA binding proteins important for AS and microRNA-mediated silencing.

3.2.8 Cancer-associated genes are direct targets of hnRNP proteins

In our analysis of hnRNP-bound transcripts, we also observed strong enrichment for the gene ontology terms “cell cycle” ($P < 1.38x10^{-22}$) and “response to DNA damage stimulus” ($P < 1.19x10^{-18}$). Widespread misregulation of AS has been observed in several types of cancer including lung (Misquitta-Ali et al., 2011), breast (Lapuk et al.,

2010), and colon cancer (Gardina et al., 2006). Underlying this effect on AS is likely the misregulation of the levels of hnRNP proteins in cancer (Carpenter et al., 2006). To identify hnRNP-regulated AS events within cancer-associated genes, we identified all hnRNP-regulated cassette exons within a set of 368 genes where specific mutations in these genes are known to contribute to oncogenesis (Futreal et al., 2004). We found 77 hnRNP-regulated cassette exons that occurred specifically in these genes, as determined by the splicing array. Using the Fluidigm microfluidic multiplex qRT-PCR platform, we independently validated 41 of the cancer-associated splicing events (Figure 3.12).

As hnRNP proteins may bind to targets to regulate other aspects of RNA processing or other AS events that the arrays were not designed to interrogate, we searched the 368 cancer-associated genes for direct hnRNP binding to their pre-mRNAs. This experiment shows that 71% of the Sanger cancer-associated genes (261 of 368) have pre-mRNAs that are bound by at least one hnRNP protein ($P < 5.3 \times 10^{-32}$, hypergeometric test) and 57 of the 261 are significantly regulated by the bound hnRNP at the expression level and/or the AS level (causing a change in inclusion of a cassette exon upon depletion of an hnRNP; $P < 0.04$, hypergeometric test; Figure 3.6B). Of the 46 cancer-associated AS events that are regulated by multiple hnRNP proteins, 39 (85%) are regulated by all hnRNP proteins in the same splicing direction. This trend again highlights our finding that many hnRNP proteins regulate RNA processing in a concerted manner, and suggests that hnRNP proteins may cooperate and have redundant roles to maintain normal cellular homeostasis, disruption of which can lead to cancer.

3.3 Discussion

3.3.1 Cooperative regulation by hnRNP proteins

The precise control of RNA processing by RBPs is essential for maintaining correct gene expression, defects of which can lead to genetic diseases such as ALS, Fragile X syndrome, as well as cancer (Carpenter et al., 2006; Lagier-Tourenne et al., 2010; Melko and Bardoni, 2010). With the emergence of genome-wide methods for detecting direct binding of RBPs on their RNA substrates, coupled with technologies to measure changes in RNA processing genome-wide, a new understanding of the global function of individual RBPs is emerging. Here we present a genome-wide analysis that integrates regulation and binding information for six highly abundant and ubiquitously expressed RNA binding proteins in human cells. We have identified 6,555 AS events regulated by six major hnRNP proteins, A1, A2/B1, H1, F, M, and U, representing the largest set of RBP-regulated human exons demonstrated thus far. With this data we have established that these members of the hnRNP family seem to cooperate with each other in a finely connected network of regulation, a discovery that would likely be missed in analysis of a single hnRNP protein.

The hnRNP protein family was first grouped together based on their abundance and ability to crosslink to RNA, however recent studies have debated that perhaps the hnRNP proteins acted too dissimilarly to be grouped into the same family (Han et al., 2010). Here we have shown that the hnRNP proteins are similar and cooperative in their regulatory roles, with significant overlap of their bound RNA targets and also in regulation of sets of AS exons. This is in contrast to the study in Drosophila of the hnRNP A/B sub-family, which identified only partial overlapping hnRNP-bound targets (Blanchette et al., 2009). Interestingly, of common cassette exon targets, we observed a concordance of effects by hnRNP proteins A2/B1, F, H1 and U, and opposing effects by

hnRNP proteins A1 and M. Thus our genome-wide datasets reveal new insights governing the cooperativity of hnRNP proteins. Also, resonating with previous studies (Licatalosi et al., 2008; Mukherjee et al., 2011; Polymenidou et al., 2011; Sanford et al., 2009; Yeo et al., 2009), we identified that the hnRNP proteins targeted their own, and in several cases shown here, other hnRNP pre-mRNAs, as well as other RBP transcripts. This regulation of many RBPs, including many splicing factors (Figure 3.6B), has the potential to yield a cascade of direct and also indirect splicing changes in human cells when a single hnRNP is depleted. Consistent with this idea, we have found that on average less than 50% of the cassette exons have direct evidence for hnRNP binding (Figure 3.9G), suggesting that a portion of the molecular changes we have observed may be due to downstream regulation of other regulatory factors. For example, hnRNP F regulates the elongation factor ELL (Figure 3.6B), which in turn could impact a broader network of RNA processing factors via Polymerase II elongation-coupled effects on AS (Ip et al., 2011). Further saturation of our CLIP-seq libraries in the future may provide more evidence for bound targets. The absence of detectable hnRNP binding sites may also be due to the variability of UV cross-linking efficiency. A smaller scale study in human cells previously suggested potential cross-regulation between the hnRNP proteins and other RNA binding proteins, albeit without direct protein-RNA evidence to support their conclusions (Venables et al., 2008). In contrast, this wide-spread cross-regulation was not observed in Drosophila cell-lines (Blanchette et al., 2009), nor in another study of a splicing factor, *RBFOX1* in mouse brains (Gehman et al., 2011). RNA processing, and AS in particular is likely far more complex in human, especially for the abundant, ubiquitously expressed RBPs like the hnRNP protein family. Integrated with AS information, our binding data has also enabled us to build splicing RNA maps that, as a whole, support that combinatorial control by hnRNP proteins could be achieved

through direct binding of certain hnRNP proteins, followed by the coordination and recruitment of other hnRNP proteins to nearby, weakly bound sites (Blanchette et al., 2009).

3.3.2 Targets of hnRNP proteins enriched in cancer pathways

Consistent with the increasing evidence that hnRNP proteins contribute to cancer pathogenesis, we have discovered that cancer-associated genes are direct and regulated targets. While hnRNP proteins have been directly linked to cancer, their specific mechanisms for affecting tumorigenesis largely remain to be identified. Here, we have provided a glimpse into what cancer-related pathways may be specifically targeted in cancers. The direct mechanism in which the hnRNP proteins interact with these genes and affect cancer progression remains to be further examined. We have provided a web resource (http://rnabind.ucsd.edu/) that enables users to upload sequences within protein-coding genes, which will be automatically inspected for hnRNP binding and regulation of AS within the gene. Taken together, the thousands of hnRNP regulated splicing changes that we have identified have the potential to provide insight into hnRNP-regulated splicing events in cancer and other disease systems.

3.4 Methods

HnRNP A1, A2/B1, F, H1, M and U were individually depleted in human 293T cells by transfecting siRNAs at a final concentration of 100nM (Table 3.3 for all siRNA sequences) using Lipofectamine-2000 (Invitrogen). Transfections were repeated 48 hrs later and cells were harvested in TRIzol (Invitrogen) 72 hrs after the first transfection. Extracted RNA was subjected to splice-sensitive microarrays and strand-specific RNA-seq, prepared based on a previously described method (Parkhomchuk et al., 2009) with minor modifications (Detailed in Section 3.2.12). To acquire RNA targets, untreated

human 293T cells were subjected to UV cross-linking and the frozen lysates were subjected to the CLIP-seq protocol for hnRNP A1, A2/B1, F, M, or U as previously described (Yeo et al., 2009), using anti-hnRNP A1 (Novus Biologicals), anti-hnRNP A2/B1 (Santa Cruz Biotechnologies), anti-hnRNP F (Santa Cruz Biotechnologies), anti-hnRNP M (Aviva Systems Biology), or anti-hnRNP U (Bethyl Laboratories), respectively. RNA libraries were subjected to standard Illumina sequencing. Details of computational analyses are described in the sections below. All data (splicing arrays, CLIP-seq and RNA-seq) are deposited at the Gene Expression Omnibus under accession number GSE34992.

3.4.1 Cell culture and transfections

Human 293T cells were obtained from ATCC and cultured in DMEM supplemented with 10% FBS at 37˚C and 5% CO_2. Cells were transfected at 40% confluency using 5µl/ml Lipofectamine-2000 (Invitrogen) and siRNAs at a final concentration of 100nM. For hnRNP A1, A2/B1, H1, M and U, equal mixtures of multiple siRNA sequences were used and for hnRNP F only one siRNA sequence was used (See **Table 3.3** for all siRNA sequences). Media was changed 4-6 hrs after the initial transfection, and the transfection was repeated 48 hrs later. Cells intended for RNA analysis were suspended and lysed in TRIzol (Invitrogen) 72 hrs after the first transfection. Human primary fibroblasts were grown at 37˚C and 5% CO_2 in DMEM supplemented with 10% FBS, NEAA, and L-glutamine. At 90% confluency, the cells were washed with PBS and transfected with antisense oligonucleotides (ISIS Pharmaceuticals) at a final concentration of 12.5nM using 5µl/ml Lipofectamine-2000 diluted in OptiMEM (See **Table 3.3** for antisense oligonucleotide (ASO) sequences). Growth media was added 4 hrs after, followed by a full media change the next day. Cells were harvested 48 hours after transfection.

3.4.2 Western blot analysis

Cells intended for protein analysis were suspended in NP-40 lysis buffer (50mM Tris-HCl pH 7.5, 150mM NaCl, 0.01% NP-40, Roche Complete Protease Inhibitors EDTA Free), sonicated for 30 seconds in 4˚C water bath sonicator, and then centrifuged at 14,000 RPM for 20 minutes at 4˚C. The lysate supernatant was used for western blotting. Western blots were performed using 25 µg of the obtained protein lysate separated on 10% Bis-Tris gels (Invitrogen) in 1X MOPS Buffer (Invitrogen) and transferred to Hybond-P membrane (Amersham Biosciences). Membranes were incubated overnight with anti-GAPDH (Abcam), anti-hnRNP A1 (Novus Biologicals), anti-hnRNP A2/B1 (Santa Cruz Biotechnologies), anti-hnRNP F (Santa Cruz Biotechnologies), anti-hnRNP M (Aviva Systems Biology), anti-hnRNP H1 (Bethyl Laboratories), anti-hnRNP U (Bethyl Laboratories), diluted to 1:5000, 1:1000, 1:200, 1:200, 1:2000, 1:2000 and 1:2000, respectively. Secondary antibodies (GE Healthcare) were diluted to 1:7000 and chemiluminescence reagents (Pierce) were used according manufacturers recommendations.

3.4.3 Microarray analysis for splicing changes

Microarray data analysis was performed as previously described (Polymenidou et al., 2011). For each microarray condition, the $\log_2$ ratio of skipping intensities to inclusion intensities was estimated using least-squares analysis. Significantly changing splicing events between hnRNP depletion and control were identified using a q-value < 0.05 and an absolute separation score > 0.5.

3.4.4 RT-PCR and qRT-PCR validations

cDNA was generated by reverse transcribing 1 µg of the obtained total RNA extracted from the control and hnRNP-depleted cells using oligo (dT) primer and

Superscript III reverse transcriptase (Invitrogen) according to the manufacturer's instructions. To test candidate splicing targets, PCR amplification was preformed for 32-36 cycles using ~30ng cDNA template with primers falling in exons flanking the alternate cassette exons (all primer sequences are listed in Table 3.4). Products were separated on 2% agarose gels followed by staining with SYBR safe (Invitrogen). Quantification of the isoforms was performed with ImageJ software. Intensity ratios between products including the cassette exon and skipping the cassette exon were averaged for 3 biological replicates for the hnRNP depletion and control experiments. Quantitative RT-PCR (qRT-PCR) was carried out on 7900HT Fast Real-time PCR System (Applied Biosystems) using approximately 25ng of cDNA with Power SYBR-Green PCR Master mix (Applied Biosystems) in triplicate. Gene expression values were normalized to GAPDH levels and are shown as fold change relative to sample control. Primer pairs used for qRT-PCR are listed in Table 3.4 and verified for a single product by a melting curve analysis.

3.4.5 Crosslinking immunoprecipitation followed by high-throughput sequencing (CLIP-seq)

Three 10cm plates of confluent, untreated human 293T cells were subjected to UV cross-linking on ice. CLIP-seq libraries for hnRNP A1, A2/B1, F, M, and U were constructed as previously described (Yeo et al., 2009), using anti-hnRNP A1 (Novus Biologicals), anti-hnRNP A2/B1 (Santa Cruz Biotechnologies), anti-hnRNP F (Santa Cruz Biotechnologies), anti-hnRNP M (Aviva Systems Biology), and anti-hnRNP U (Bethyl Laboratories). Libraries were subjected to standard Illumina sequencing protocols.

3.4.6 Human (hg18) gene structure annotation

The human genome sequence (hg18) and annotations for protein-coding genes were obtained from the University of California, Santa Cruz (UCSC) Genome Browser. Known human genes (knownGene containing 66,803 entries) and 26,570 known isoform clusters (knownIsoforms) with annotated exon alignments to the human genomic sequence were processed as follows. All mRNAs that were aligned to hg18 and were greater than 300 nt were clustered together with the known isoform clusters. A total of 4,162,937 spliced ESTs were mapped onto the high-quality gene clusters to annotate AS events. Final annotated gene regions were clustered together so that any overlapping portion of these databases was defined by a single genomic position. To identify 5′ and 3′ untranslated regions we relied on the coding annotation in UCSC known genes that we extended 1.5kb downstream or upstream the start and stop codons, respectively.

3.4.7 CLIP-seq data processing and cluster generation

CLIP-seq reads were processed as previously described (Polymenidou et al., 2011). Briefly, reads were trimmed to remove sequencing adaptors and homopolymeric runs >10nt, and mapped to the human genome (hg18) using Bowtie (version 0.12.2 with parameters –q –l 20 –m 5 –k 5 —best). Multiple sequencing runs for a single CLIP experiment were combined, collapsing reads from the same library preparation with identical sequence into one unique read to eliminate redundancies caused by PCR amplification. For the purposes of defining significant clusters, all reads falling at the exact position in the genome were collapsed and considered a single read. Significant clusters were calculated as previously described (Polymenidou et al., 2011).

3.4.8 Motif Analysis

De novo motif finding was implemented using hnRNP clip derived clusters (CDCs), which were extended equally from the center such that all CDCs were 150nt in length. For an appropriate background sequence set, all repeat-masked pre-mRNA sequences were extracted from the human hg18 genome. The Homer algorithm was used for this analysis (findMotifs.pl with parameters –len 5 –homer1 –chopify –norevopp –rna –fasta) (Heinz et al., 2010).

3.4.9 RNA splicing maps

CLIP-seq reads were mapped to the set of all 4,829 splicing array-detected cassette exons at each nucleotide position in four regions, –50 nt to +400 nt near the upstream exon, –400 nt to +50 nt at the 3′ end of the cassette exon, –50 nt to +400 nt at the 5′ end of the cassette exon, and –400 nt to +50 nt near the downstream exon. If at least one CLIP-seq read overlapped a given nucleotide position, the cassette exon was counted as having binding at that position. For each hnRNP, cassette exons were separated into three groups, repressed, activated and unchanged. Using R (version 2.11.1), we plotted at each nucleotide position the fraction of activated cassette exons with binding, the fraction of repressed cassette exons with binding, and the fraction of unchanged cassette exons with binding. Significant peaks around regulated exons compared to unchanged exons were computed using the chisq.test() function in R and a Bonferonni corrected p-value cutoff of 0.05.

3.4.10 Tissue specific exon analysis

The unprocessed single-end RNA-seq Illumina human BodyMap data was downloaded from http://sra.dnanexus.com. Using an in-house analysis pipeline, we mapped the reads to the human genome (hg18) using the GSNAP short-read alignment

algorithm (Wu and Nacu, 2010). Next, we ran the MISO algorithm (Katz et al., 2010) on all mapped reads for each dataset in order to compute percent spliced in (PSI) values for each exon. We next applied a Z-score statistic to identify exons that were specifically enriched in the 16 tissues (|Z-score| > 2). We defined sets of exons that were specific to each of the 16 tissues, taking all exons that passed the Z-score cutoff in only one tissue (the exon is differentially spliced only in one of the 16 tissues). For each tissue specific exon, we counted hnRNP CDCs within 2kb. We also generated a randomly distributed set of clusters (RDCs), taking the same size and number of CDCs and randomly shuffling their location in a gene, and counted the number of RDCs within 2kb of all tissue specific exons as a background count. Comparisons of the number of exons with nearby CDCs with the number of exons with nearby RDCs determined whether the set of exons were enriched for hnRNP binding. Using a Fisher exact test and a p-value cutoff of 0.001, no set of exons were identified as significantly enriched for hnRNP binding. We repeated this same enrichment analysis for all ACEscan exons, downloaded from the UCSC genome browser.

3.4.11 Gene ontology analysis

We used the **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (DAVID Bioinformatic Resources 6.7; http://david.abcc.ncifcrf.gov/) for all genes with hnRNP binding. The target gene set for each hnRNP was analyzed individually, as well as the intersection of all hnRNP target genes using RefSeq identifiers for each gene.

3.4.12 RNA sequencing preparation (RNA-seq)

Strand-specific RNA-seq libraries were prepared based on the method described in Parkhomchuk et al., 2009, with a few modifications. 10 µg of total RNA was DNase treated (Turbo DNase, Ambion) and then subjected to 2 rounds of polyadenylation

selection using Oligo $(dT)_{25}$ Dynabeads (Invitrogen) according to manufacturer's recommendations. Eluted mRNA was then reverse transcribed with a mixture of random hexamers and oligo-dT primers. Unincorporated nucleotides were then removed using G-25 spin columns (GE Healthcare) and the second strand of the cDNA was then synthesized in the presence of dUTP instead of dTTP. This reaction was then brought to 350 µl and sonicated on high for 1 hour of 30 seconds on, 30 seconds off sonication bursts. Fragments ranging from 150-225 bp were then selected from a 6% polyacrylamide gel, before the ends of each fragment were repaired using T4 DNA Polymerase, Klenow DNA Polymerase, and T4 PNK (NEB) at 20°C for 1 hour. Enzymes, buffers, and nucleotides in all reactions were removed using Qiagen PCR Purification MinElute columns. Following the end repair, a single dATP was added to each 3′ end using Klenow Exo minus (NEB) at 37°C for 1 hr. The adaptors were then ligated using 2000 units per µl T4 DNA ligase (NEB) at room temperature for 30 minutes. The dUTPs were then removed using Uracil N-Glycosylase (Fermentas) at 37°C for 30 minutes. One half of the library was then subjected to 15 cycles of PCR amplification using Phusion High Fidelity DNA Polymerase (NEB). The library was then separated on a 6% polyacrylamide gel and fragments between 250-325 nucleotides were selected. 8 pM of the resulting amplified libraries were subjected to standard Illumina sequencing protocols.

3.4.13 RNA-seq data processing and gene expression analysis

Strand-specific RNA-seq reads from control siRNA treatment, and each hnRNP depletion experiment were processed as previously described (Polymenidou et al., 2011). An average of 70% of reads mapped uniquely to our gene structure database, using Bowtie (version 0.12.2, with parameters –l 20 –m 5 –k 5 —best —un —max –q). Multiple sequencing runs of the same condition were combined, collapsing reads from

the same library preparation with identical sequence into one unique read to eliminate redundancies caused by PCR amplification. For each condition, the mRNA RPKM (reads per kilobase of million mapped reads) was computed for each gene as a metric for expression. Significantly up- and down-regulated genes upon hnRNP depletion were identified as previously described (Polymenidou et al., 2011), using a local Z-score analysis with cutoffs of $Z > 2$ for upregulated and $Z < -2$ for downregulated.

3.4.14 RBP and cancer target analysis

RNA binding proteins were computationally filtered from PFAM database (http://pfam.sanger.ac.uk/), selecting all proteins in the genome that contained any of the 81 known RNA binding domains. This list was further filtered to remove proteins containing the RNA binding domains specific for tRNAs, snoRNAs and rRNAs, with the remaining 443 genes containing RNA binding domains predicted to bind pre-mRNA and mRNA sequences. This list of 443 RNA binding proteins were filtered for those bound by at least one of the 6 hnRNP proteins assessed in this study (as indicated by CLIP-seq). Genes with binding were filtered further for those significantly regulated at the expression level (from RNA-seq) or those that contained a significantly regulated alternative cassette exon (from splicing arrays) in at least one hnRNP depletion experiment. We also repeated this analysis for a set of 457 cancer-associated genes that were obtained from the Sanger Cancer Gene Census (http://www.sanger.ac.uk/genetics/CGP/Census/), of which 368 could be cross-referenced to our gene annotations. Genes that were bound and regulated according to the above criteria were loaded into Cytoscape (http://www.cytoscape.org/) to visualize the hnRNP regulation network.

3.5 HnRNP publication acknowledgement

Chapter 3 is in whole published as a manuscript entitled “Integrative Genome-wide Analysis Reveals Cooperative Regulation of Alternative Splicing by hnRNP Proteins” in Cell Reports (Huelga et al., 2012). This dissertation author was the primary author who wrote the manuscript, directed experiments, and analyzed the data.

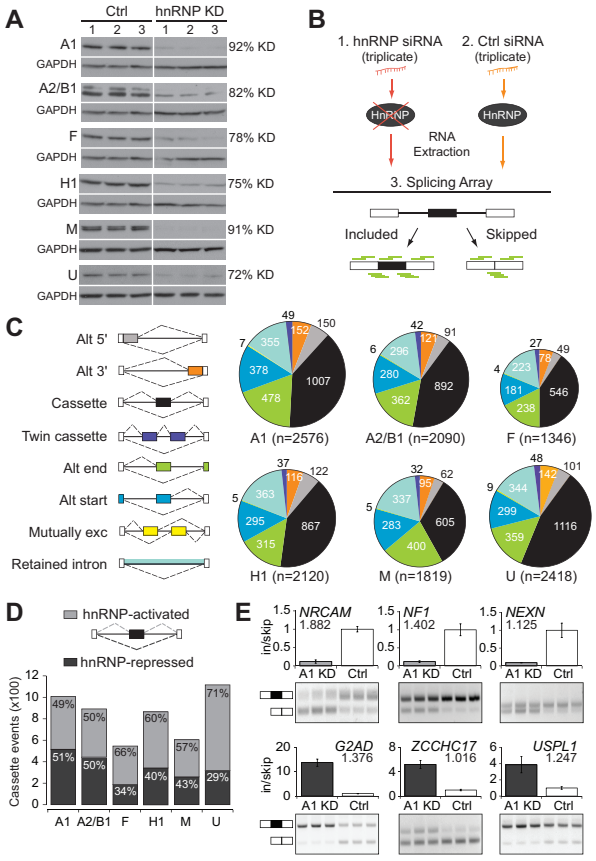


Figure 3.1: Thousands of hnRNP-dependent alternative splicing events identified in human cells

A. Western analysis demonstrating successful depletion of hnRNP proteins A1, A2/B1, F, H1, M, and U in human 293T cells in triplicate experiments. GAPDH was used as a loading control. The extent of protein depletion was quantified by densitometric analysis. **B.** Experimental strategy for identification of alternative splicing (AS) events. Total RNA was extracted from human 293T cells treated with (1) hnRNP-targeting siRNA(s), or (2) a control non-targeting siRNA, and (3) was subjected to splicing-sensitive microarrays (probes depicted in green). **C.** Pie charts of differentially regulated AS events as detected by splicing arrays. Colors represent the types of AS events that were detected and the size of each pie chart reflects the number of events detected. **D.** Cassette exons differentially regulated upon depletion of individual hnRNP proteins, relative to control. Bars are divided to display the proportion of events that are less included upon depletion (hnRNP-activated, light gray), or more included upon depletion (hnRNP-repressed, dark gray). **E.** RT-PCR validation of hnRNP A1-regulated cassette exons ($P < 0.01$, *t*-test). Light gray indicates activation of exon inclusion and dark gray indicated repression of exon inclusion. Absolute separation score predicted by the splicing array analysis is listed below the affected gene name. Bar plots represent densitometric analysis of bands and error bars are standard deviations within triplicate experiments (See also Figure 3.7 as Supplement).

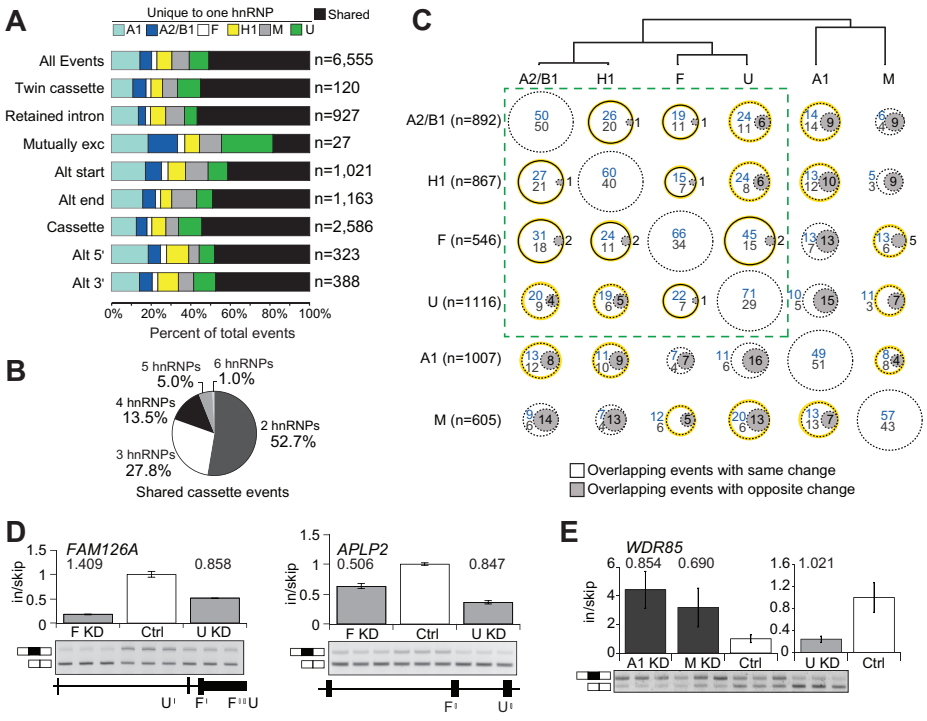


Figure 3.2: Multiple hnRNP proteins co-regulate AS events

A. For each type of AS event, the percentage of events that was affected in more than one hnRNP depletion is in black, and the percentage of events that was affected uniquely in a single hnRNP depletion. **B.** Distribution of shared cassette exons affected by 2 to 6 hnRNP proteins. **C.** Comparisons of the overlap and directionality of cassette exons affected by pairs of hnRNP proteins. The dendrogram displays the hierarchical clustering of all cassette exons regulated by each hnRNP protein. Circles represent the percent of overlapping events for the hnRNP in the row, with the hnRNP listed in the column. Circles with gray shading indicate overlapping events that were affected in opposing directions (depletion of one hnRNP protein leads to inclusion and depletion of the other hnRNP leads to skipping). For cassette exons that were affected in the same direction, blue values indicates the percentage of events that were hnRNP-activated while values in gray listed below represent percentage that were hnRNP-repressed. The solid black lines represent significant overlap of targets ($P < 0.05$, hypergeometric test) and the yellow highlighting denotes overlapping cassettes that are significantly changing in the same direction ($P < 0.002$, hypergeometric test). The dashed-line box contains the hnRNP proteins that are most similar. **D.** RT-PCR validations of cassette exons affected in the same direction upon depletion of hnRNP F or hnRNP U ($P < 0.01$, *t*-test). Binding of hnRNP F and U near the cassette exon is shown below. **E.** RT-PCR validation of a cassette exon similarly affected by depletion of hnRNPA1 or hnRNP M, but opposite from the change caused by depletion of hnRNP U ($P < 0.05$, *t*-test). Absolute separation score as predicted by the splicing array analysis is listed above the bars. Bar plots represent densitometric analysis of bands and error bars are standard deviations in triplicate experiments (See also Figure 3.8 as Supplement).

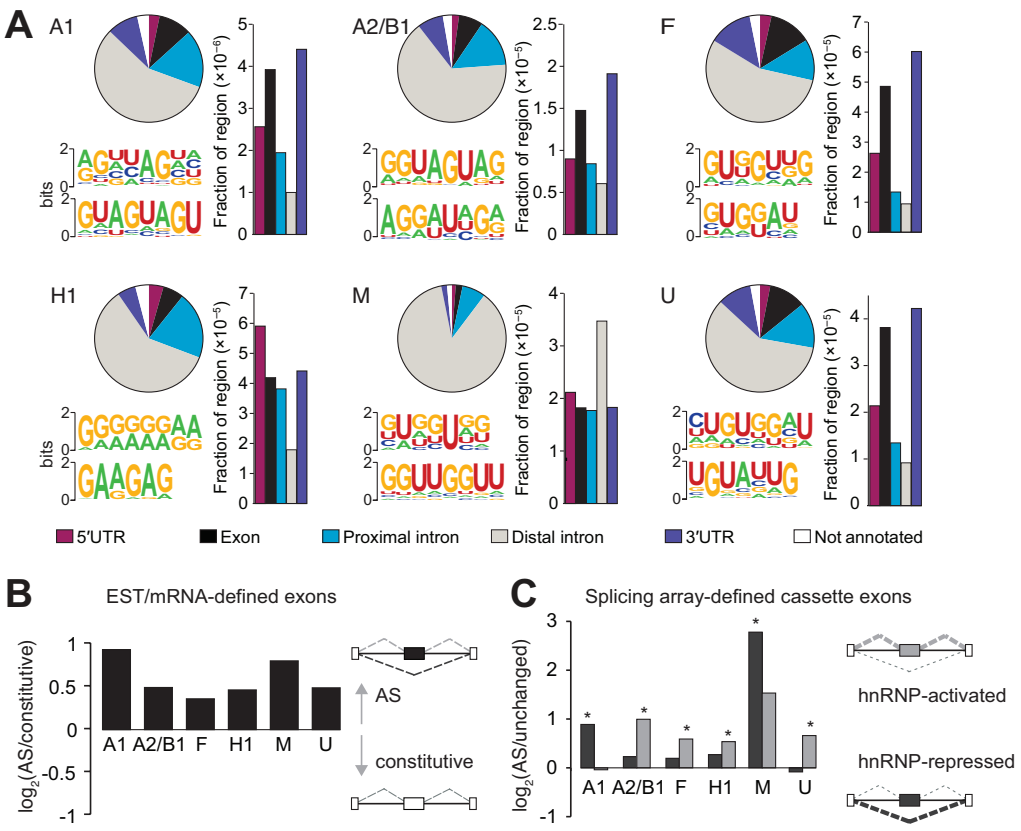


Figure 3.3: HnRNP proteins bind to sequence motifs *in vivo* and are enriched near regulated exons

A. Distribution of hnRNP binding sites within different categories of genic regions (pie charts). Bar plots represent the fraction of a particular region bound, relative to the total genomic size of the region. The most representative top motifs identified by the HOMER algorithm are also shown. **B.** Preference for hnRNP protein binding near expressed sequence tag (EST) or mRNA-defined cassette exons. The y-axis is the $\log_2$ ratio of the fraction of 41,754 cassette exons to the fraction of 222,125 constitutive exons harboring hnRNP binding clusters within 2 kilobases (kb) of the exon. All bars are significant ($P < 10^{-28}$, Fisher exact test). **C.** Preference for hnRNP proteins to bind proximal to splicing array-defined cassette exons that are activated or repressed by a specific hnRNP protein. The y-axis is the $\log_2$ ratio of the fraction of changed cassette exons to the fraction of unchanged cassette exons having hnRNP binding clusters within 2 kb of the exon. Dark gray bars represent the preference for hnRNP-repressed compared to unaffected exons and light gray bars represent the preference for hnRNP-activated compared to unaffected exons. A single asterisk denotes the most significant binding preference ($P < 0.05$, Fisher exact test). (See also Figure 3.9 as Supplement).

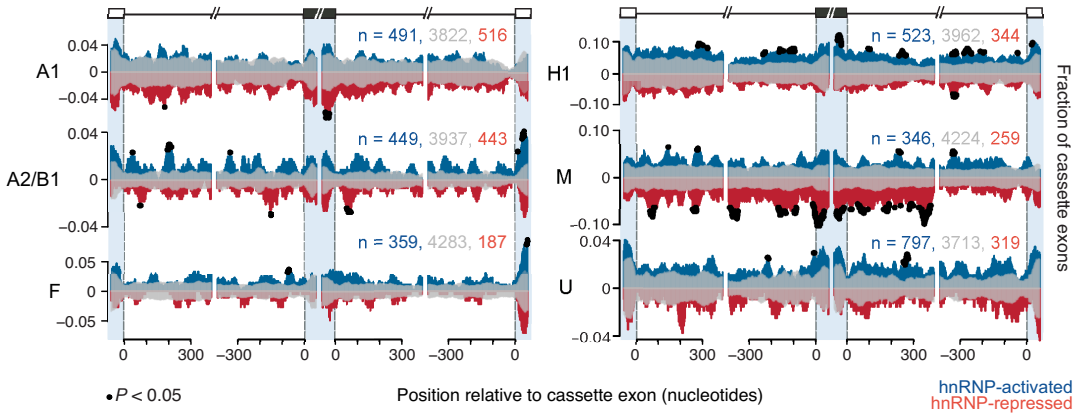


Figure 3.4: HnRNP proteins bind in a position-specific manner around cassette exons

RNA splicing maps showing the fraction of microarray-detected cassette exons with CLIP-seq reads located proximal to activated (positive y-axis, blue), repressed (negative y-axis, red), and unaffected cassette exons (reflected on positive and negative y-axis, transparent gray). The x-axis represents nucleotide position relative to upstream, cassette and downstream exons. Blue shading separates positions that fall in the upstream exon (left, –50-0), in the cassette exon (middle left, 0-50, middle right, –50-0) and in the downstream exon (right, 0-50). Black dots represent nucleotide positions where significantly more hnRNP-activated cassette exons or hnRNP-repressed cassette exons showed binding, compared to unchanged cassette exons ($P < 0.05$, chi-square test). The total counts for activated, unaffected, and repressed exons that were used to generate the maps are listed. (See also Figure 3.10 as Supplement).

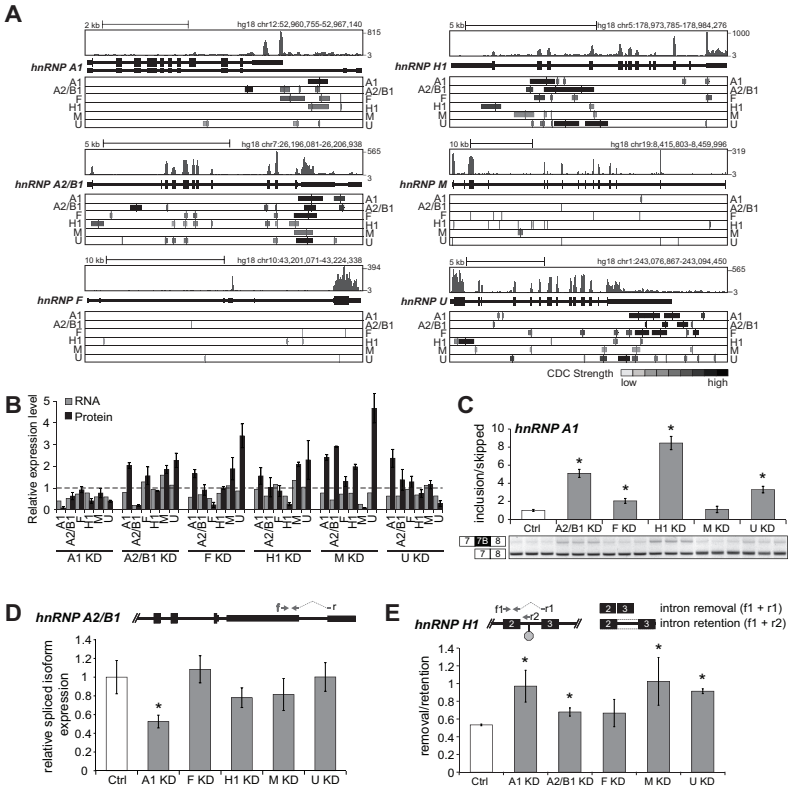


Figure 3.5: Cross-regulation of hnRNP proteins adds complexity to splicing regulation

A. CLIP-derived clusters (CDCs) identified for each hnRNP within each of the 6 hnRNP gene structures. CDCs are shaded according to strength of binding, where darker shading indicates more significant read coverage. RNA-seq read distributions in control siRNA treated 293T cells are shown above each hnRNP gene structure.
B. Quantification of hnRNP protein abundance (based on densitometric measurements of Figure S5 bands) and RNA abundance (based on RNA-seq RPKMs) within the context of the different hnRNP depletion conditions. Y-values are fold-changes over the control protein and RNA levels. **C.** RT-PCR splicing validation of hnRNP A1 exon 7B in each of the hnRNP KD experiments. Bars represent densitometric measurements of bands relative to control. Significant changes relative to control are starred ($P < 0.002$, *t*-test). **D.** Relative expression of an hnRNP A2/B1 spliced 3'UTR NMD candidate isoform in each of the hnRNP KD experiments. Primers 'f' and 'r' are depicted. Bars represent an average abundance by qRT-PCR across triplicate KD experiments, relative to control. All measurements were normalized to GAPDH expression. Significant changes relative to control are starred ($P < 0.001$, *t*-test). **E.** Ratio of relative hnRNP H1 isoform expression with intron removed (measured with primers f1 and r1 depicted) compared to relative hnRNP H1 isoform expression with intron retained (measured with primers f1 and r2 depicted) in each of the hnRNP KD conditions. Bars represent an average abundance by qRT-PCR across triplicate KD experiments relative to control. All measurements are normalized to GAPDH expression. Asterisks denote significant changes relative to control ($P < 0.04$, *t*-test).

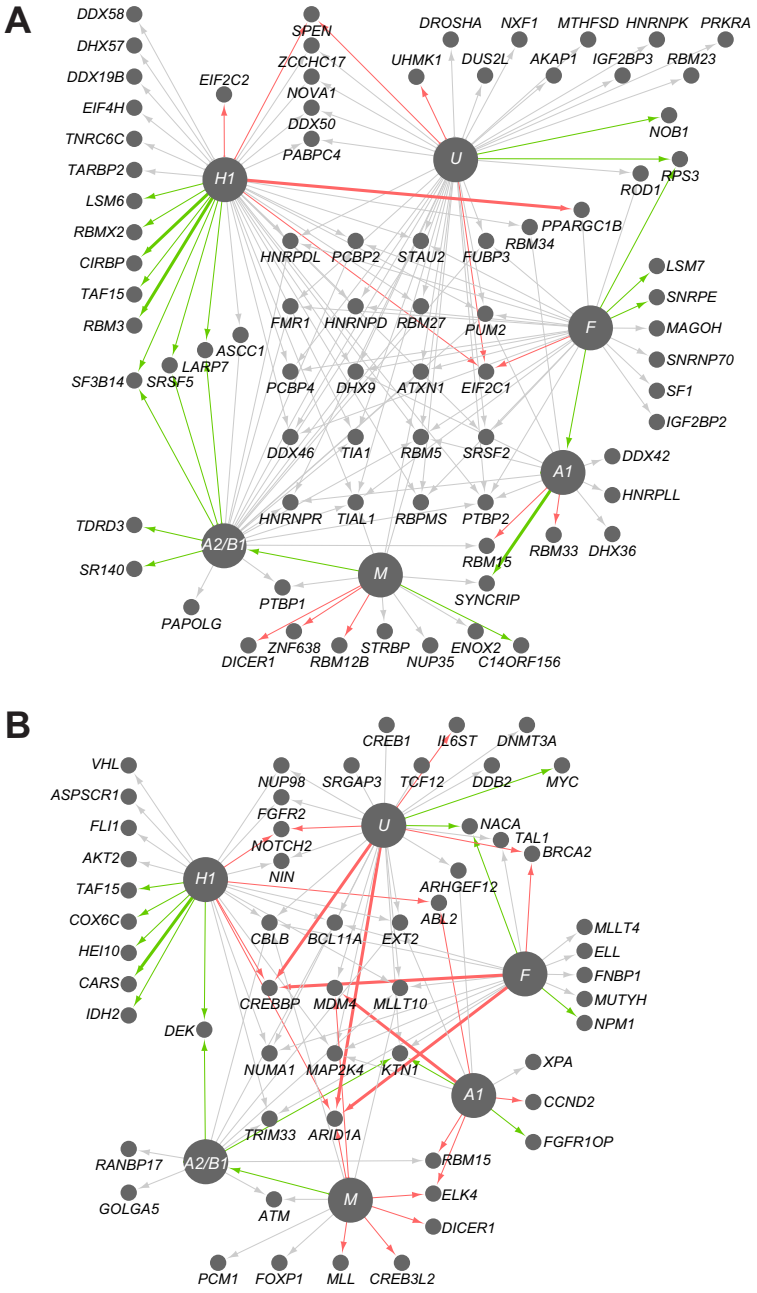


Figure 3.6: HnRNP proteins regulate other RBPs and cancer-associated transcripts

Connectivity diagrams of hnRNP-regulated and hnRNP-bound (**A**) RNA binding protein transcripts or (**B**) cancer-associated transcripts. Nodes represent hnRNP RNA targets and the hnRNP nodes are enlarged for emphasis. Each edge represents an hnRNP target that is both bound directly and regulated, colored according to the type of regulation. Gray represents that the gene contains a cassette exon that changes upon depletion of the hnRNP. Red and green edges reflect up and down-regulation of mRNA levels upon depletion of the hnRNP. Thicker edges represent genes with both an alternative cassette and an expression change. (See also Figure 3.12 as Supplemental)

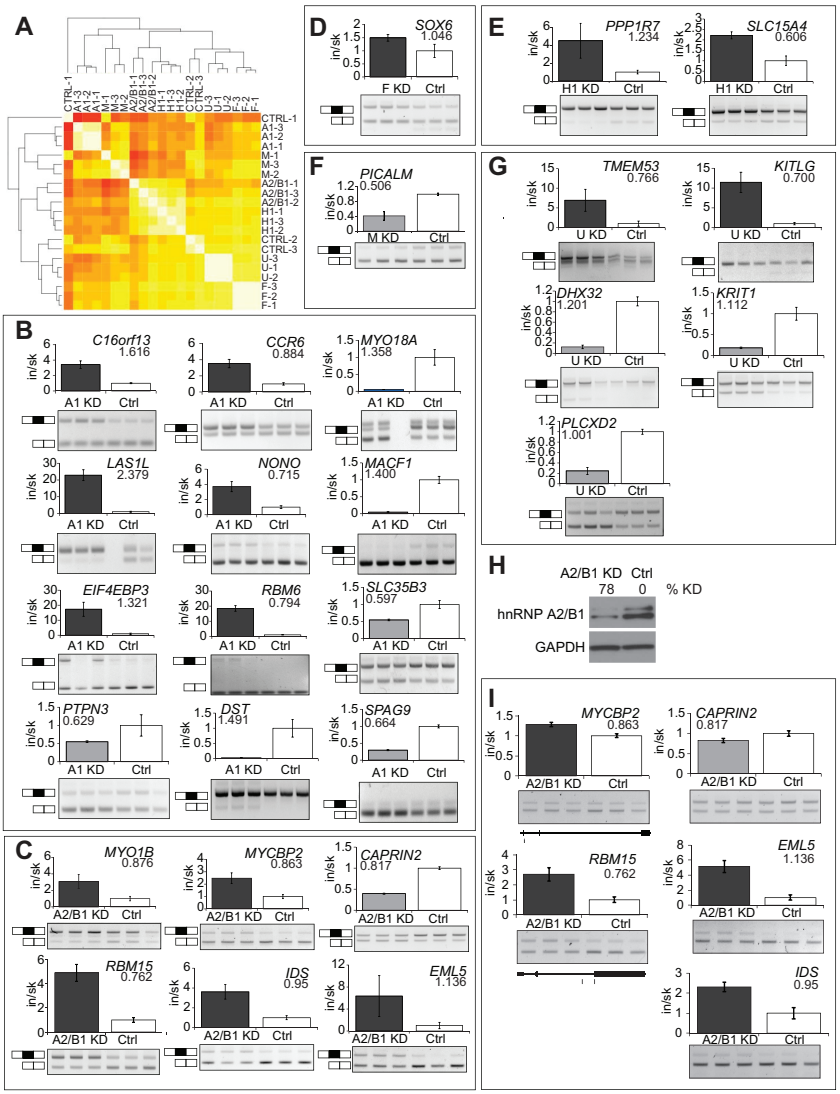


Figure 3.7: Correlation and validations of splicing array data

A. Correlation heatmap for normalized triplicate splicing array intensities. Analyses excluding the unclustered control experiment (CTRL-1) confirm that our results are unaffected due to this control experiment. **B-G.** Additional RT-PCR validations for individual hnRNP-regulated events. Bar plots represent densitometric analysis of bands and error bars correspond to standard deviation in triplicate experiments. Absolute separation score as predicted by the splicing array analysis is listed below the gene name. All events shown here are significantly different as measured by a *t*-test ($P < 0.05$). **H.** Western blot showing successful depletion of hnRNP A2/B1 in primary human fibroblasts using an anti-sense oligonucleotide (ISIS Pharmaceuticals). Quantification of percent knockdown was performed by densitometric analysis of the bands. GAPDH was used as a loading control. (I) RT-PCR validation of hnRNP A2/B1 regulated events in primary human fibroblasts ($P < 0.05$, *t*-test). For comparison, see Figure 3.7C.

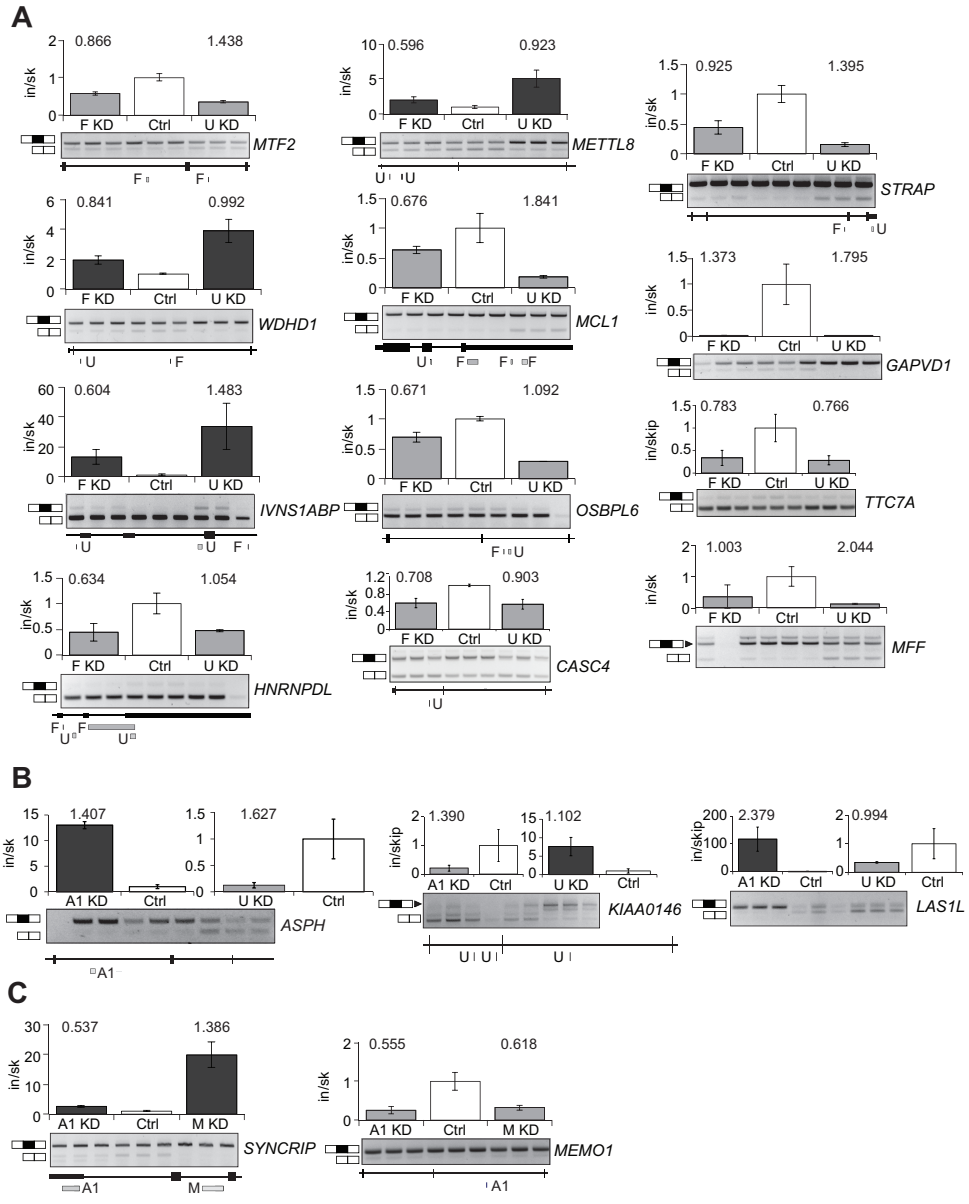


Figure 3.8: Validation of hnRNP co-regulated events

A. Additional RT-PCR validations of antagonistic splicing events co-regulated by hnRNP F and U. **B.** RT-PCR validations of antagonistic splicing events co-regulated by hnRNP A1 and U. **C.** RT-PCR validations of synergystic splicing events co-regulated by hnRNP A1 and M ($P < 0.05$, *t*-test). Bar plots represent densitometric analysis of bands. Absolute separation score as predicted by the splicing array analysis is listed above the bars. Binding of the regulating hnRNP proteins near the cassette exon is shown below if CLIP derived clusters are detected.

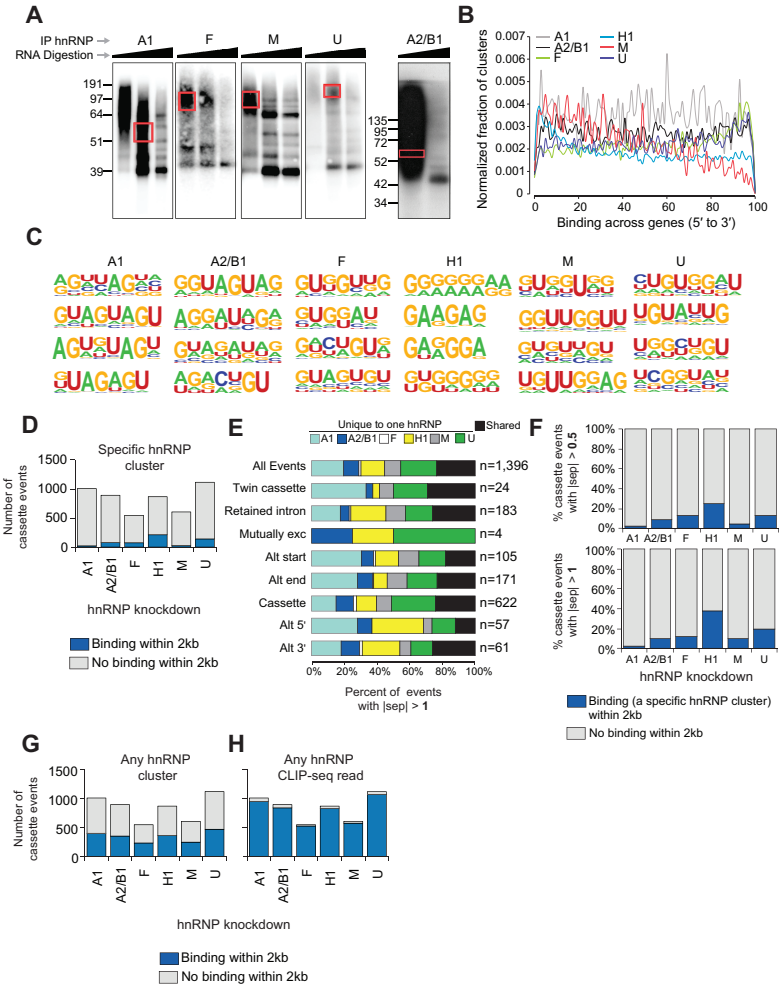


Figure 3.9: HnRNP CLIP-seq data analysis

A. Autoradiograph of hnRNP-RNA complexes digested with increasing amounts of micrococcal nuclease (MNase). The red box is the RNA that was excised and used for CLIP-seq library preparation. **B.** CLIP derived clusters plotted across composite gene structure displays weak preferences for binding at 3′ or 5′ gene ends. **C.** Top 4 enriched motifs as determined by the HOMER motif discovery algorithm. **D.** Fraction of changing AS events with binding (blue), or no binding (gray) in the exon and within 2kb. HnRNP binding is defined by a CLIP-seq cluster for the specific hnRNP whose depletion affected the event. **E.** For each type of AS event from the top quartile of events (|sep_score| > 1), the percentage of events that was affected in more than one hnRNP depletion (black) and the percentage of events that was affected uniquely in a single hnRNP depletion. For comparison, see **Figure 3.2A**. **F.** Comparison of the percent of all splicing events (|sep_score| > 0.05) that are direct targets, and the percent of only the top quartile of events (|sep_score| > 1) that are direct targets. No significant difference is observed. **G.** Same as D, but HnRNP binding is defined by a CLIP-seq cluster for any hnRNP. **H.** Same as D, but HnRNP binding is defined by a CLIP-seq read for any hnRNP.

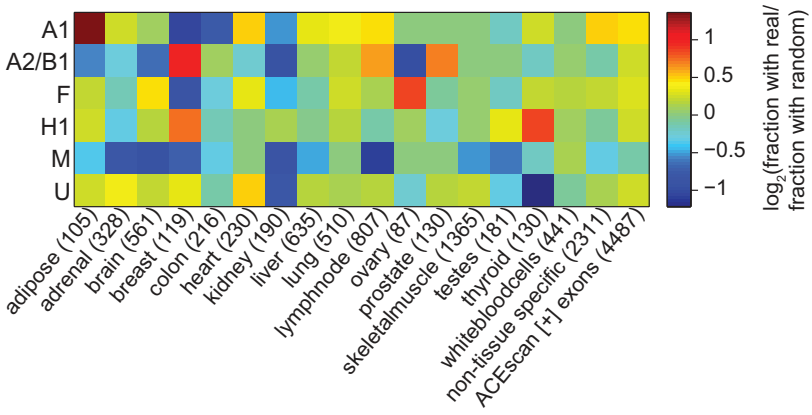


Figure 3.10: HnRNP protein binding near tissue specific splicing events

Heatmap of the log ratio of hnRNP binding versus random background binding near exons which are alternatively spliced in a tissue-specific manner or show conserved alternative splicing between human and mouse (ACEscan [+] exons). No set of exons is significantly over- or under-represented for hnRNP binding ($P < 0.001$, Fisher exact test).

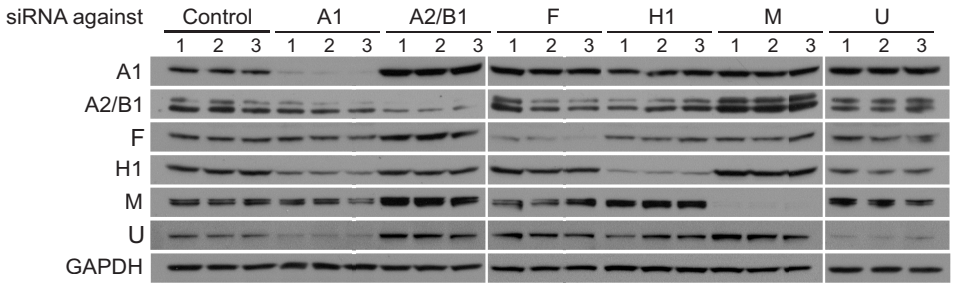


Figure 3.11: Cross-regulation of hnRNP protein levels upon knockdown

Western blots for all 6 hnRNP proteins across all siRNA treated triplicate conditions. GAPDH was used as a loading control.

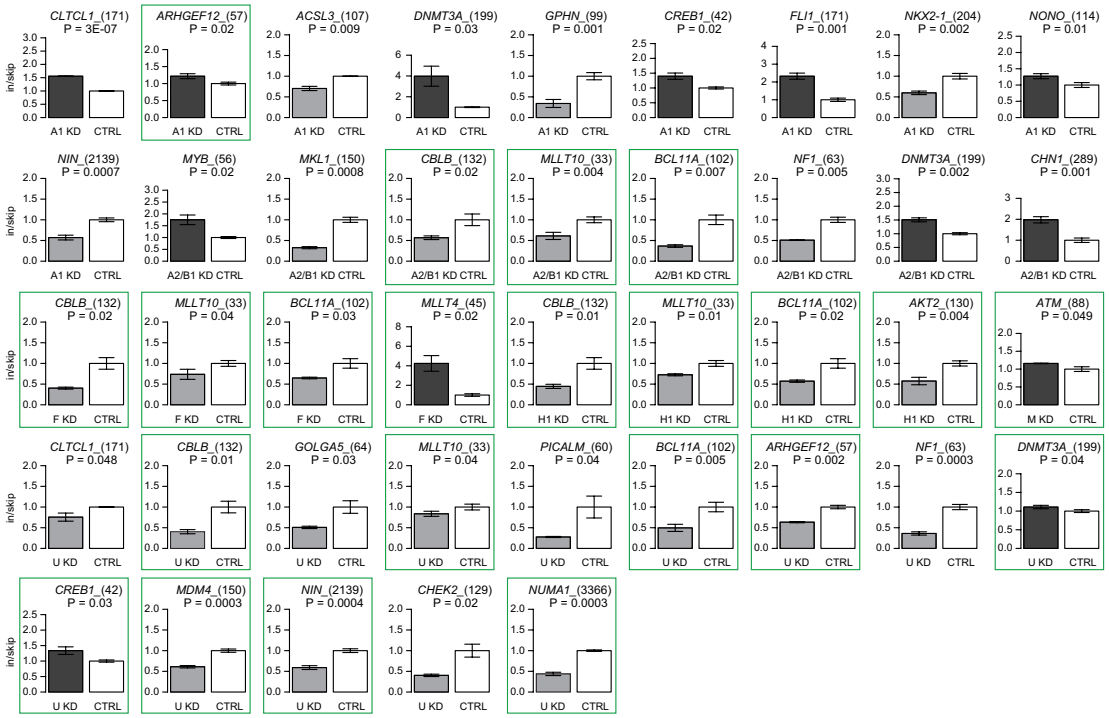


Figure 3.12: Cancer-associated splicing validations by Fluidigm qRT-PCR

Significant splicing changes for 41 cancer-associated events as predicted by the splicing microarray and successfully validated on the Fluidigm micro-fluidic multi-plex qRT-PCR platform. The name of the affected gene along with the length of the cassette exon in parenthesis, and the p-value determined by Student's *t*-test relative to control ($P < 0.05$) is shown. Green boxes are validations of edges shown in Figure 3.6B.

Table 3.1: CLIP-seq data analysis counts

CLIP-seq generates millions of reads and identifies thousands of binding sites for members of the hnRNP protein family. Original raw read count, number of Bowtie mapped reads, number of binding sites (CDCs), and total gene targets containing at least one CDC are listed for each hnRNP.

	A1	A2/B1	F	H1	M	U
Raw reads	7,424,077	21,136,009	41,149,356	(Katz et al., 2010)	10,816,641	27,202,723
Mapped reads	1,778,279 (24%)	16,659,630 (79%)	24,347,839 (59%)	3,856,295	4,727,761 (44%)	11,408,515 (42%)
Reads at unique position in genes	503,416	311,052	586,005	1,674,958	1,540,171	625,481
CDCs	1,956	10,193	18,951	33,209	4,437	17,175
Gene Targets	1,207	4,480	6,479	8,966	1,570	6,057

Table 3.2: hnRNP binding near regulated events

Binding for each hnRNP protein (columns) to each set of hnRNP-regulated events (rows). Binding is defined as an event that contains a CDC for a particular hnRNP on or within 2kb of the changing exon. Values represent a ratio of the fraction of changing events with binding compared to the fraction of unchanged events with binding by each hnRNP for each knock down. Larger numbers demonstrate more binding to events that are differentially spliced relative to unchanged events. The gray shading highlights the 3 largest ratios.

		Binding					
		A1	**F**	**M**	**U**	**A2/B1**	**H1**
Set of regulated events	**A1**	1.398	1.366	1.344	1.205	1.294	1.143
	F	1.569	1.379	1.101	1.250	1.537	1.250
	M	2.248	1.306	4.364	1.337	1.539	1.182
	U	1.969	1.447	1.822	1.399	1.723	1.427
	A2/B1	2.337	1.160	1.823	1.183	1.585	1.198
	H1	1.795	1.367	1.254	1.401	1.449	1.344

Table 3.3: siRNA sequences

siRNA and antisense oligonucleotide (ASO) sequences used for hnRNP knock downs in human 293T cells and primary fibroblasts.

Target gene	siRNA sequence 5' to 3'	Source
hnRNP A1	CCACUUAACUGUGAAAAAGAUUU CUUUGGUGGUGGUCGUGGAUU	Venables et al., 2008, Thermo Scientific
hnRNP A2/B1	GUUCAGAGUUCUAGGAGUG GGAGAGUAGUUGAGCCAAA GAACAAUGGGGAAAGCUUA GCAAGACCUCAUUCAAUUG	Zefeng Wang (University of North Carolina), Dharmacon
hnRNP F	GGAAUGUAUGACCACAGAU	Jens Lykke-Anderson (UCSD)
hnRNP H1	CAAACAACGUUGAAAUGGA UGAAAAGGCUCUAAAGAAA GUUCGCAACUCAUGAAGAU GAACACAGAUAUGUAGAAC	Zefeng Wang (University of North Carolina), Dharmacon
hnRNP M	GAAGUCCUAAACAAGCAUAUU AUAUGCCAAUCCAACUAAAUU	Venables et al., 2008, Thermo Scientific
hnRNP U	AGGAUAUUAUUGAAUACCCAA CUGGCCGUGGUAGUUACUCAA	Qiagen, SI02780540 Qiagen, SI02781002
Non-targeting control	UGGUUUACAUGUCGACUAA	Thermo Scientific, D-001810-01-05
ASO	**ASO sequence 5’ to 3’**	**Source**
hnRNP A2/B1	AAGAAACTGTCTCAAAGGCA	ISIS Pharmaceuticals
Non-targeting control	CTCAGTAACATTGACACCAC	ISIS Pharmaceuticals

Table 3.4: RT-PCR and qRT-PCR primer sequences

RT-PCR and qRT-PCR forward and reverse primer sequences (5' to 3') used for validation of hnRNP regulated splicing events. Predicted size (in nucleotides) of the included event (In) and skipping event (Sk) are listed in the final two columns for RT-PCR experiments.

Gene	Forward Primer 5' to 3'	Reverse Primer 5' to 3'	In	Sk
APLP2	AGAATCCTACTGAACCCGGC	ATCAACATCATTGGTTGGCA	269	101
ASPH	TCGAAGATGAAGCAAAAGAACA	CTTCCACGTGGTAACTATGCTC	220	91
C16orf13	GCTCTTCAGAGCAGCAGGAC	CTCCAGGAGGGCTGTGTC	166	98
CAPRIN2	TGATTTTCTTCAAGAGCCGTT	GCTTGAAAGGGTGCTGTGTT	267	162
CASC4	TCCAGTCCTCTTCAGCGTTT	GCAGGATCCATTTGAAGCTC	323	155
CCR6	CCAAAGTCACCAGAGGGAAA	CTGGCTTAGGAATGGGATCA	319	259
DHX32	TTATGGAAACGTGCCTGTCA	GTATCTCTGCCTGGCTCTGG	479	236
DST	ACAAACATGGAACTGCGTGA	TGGCAGCTGCAGTTGTTATC	403	292
EIF4EBP3	GGAGAACTCTCCTGCCCTTT	AGGATACACAGGAGGCATCG	937	178
EML5	ATGCTGTTCACATAGCACCC	GTCGACGGTAGGTGAGCTGT	283	129
FAM126A	TGCAGTAACCAGCATGTCAA	CTAGACGCAGCCCTGGAATA	425	129
G2AD	AGCTGGCCAAACTGCTCTAC	CAGGTCTCGGCACATCTCA	269	144
GAPDH	ATGTTCGTCATGGGTGTGAA	AGTTGTCATGGATGACCTTGG	qRT-PCR	
GAPVD1	ACAGGAACAGACCTTGGTGG	CCGTTTAATGGCATTCCTGT	212	131
HNRNPA1	TGGGGATGGCTATAATGGAT	CGCCATAGCCACCTTGGTTTC	356	200
HNRNPA2/B1 spliced 3'UTR	TCTGCTGCCACAAAGACTGTA	GCAGCAAGACACCTTCCATT	qRT-PCR	
HNRNPDL	ACTTATGGCAAGGCATCTCG	TCTTCAATGTCGTCCTGCAA	231	126
HNRNPH1 intron removal	GAGAAGGCAGACCAAGTGGC	ACTTGAATACTTCAACATATCTGTGTC	qRT-PCR	
HNRNPH1 intron retention	GAGAAGGCAGACCAAGTGGC	TTCAGAAGGGGTGAGACCTCTA	qRT-PCR	
IDS	CATCAGTGTGCCGTATGGTC	GCCGACCTGTGTATCCAAAT	227	97
IVNS1ABP	AGCATCTGGGAGAATGGAGA	CATCATCACTGCCAAACACC	274	133
KIAA0146	TGGAAGTGATTTGTCGGATG	CTGCTCACACATGGCAACTT	646	294

Table3.4 (continued): RT-PCR and qRT-PCR primer sequences

Gene	Forward Primer 5' to 3'	Reverse Primer 5' to 3'	In	Sk
KITLG	TCATTCAAGAGCCCAGAACC	GCCCAGTGTAGGCTGGAGT	272	188
KRIT1	ATCTCGGTGGTCCAACTCAG	TGGCAGTATTCTTTGGACGA	429	281
LAS1L	TTGAACAGTTGGCAGCTTTG	AACTGAGAGAACGGCTTTGG	136	85
MACF1	CCTACTCGTTCCAGCTCCAG	GCAAGGGATGTCCGACTAGA	272	161
MCL1	AGACCTTACGACGGGTTGG	ACATTCCTGATGCCACCTTC	361	133
MEMO1	GCGGATCCTAGTAATCTCTTTGTG	CCAATGGGATGTCTTCCAC	218	141
METTL8	TTCAACCTCCCAGATTCCAG	AGGACTCGCACAGCTGAGTT	294	139
MFF	CGAGCAGTTGGCAGACTAAA	ATGAGGATTAGAAGTGGCGG	291	132
MTF2	GAGCATGTTCTGGAGGCATT	TTGTATATGGGCCAGGTGGA	280	109
MYCBP2	CTCCAAGCCCTTTCTCAGTG	GGCTCATACGTCCATCAGGT	375	195
MYO18A	GCAGGCTGACCTAAAGTTGG	CCTTGGAAGGTCCCTTGTTT	233	188
MYOIB	AAAAGCGCTGTAAGGAAGCA	TTCTTTCCAGCATTGGCTCT	277	103
NEXN	AACGCAGAATTGAGCAGGAT	TCCTTGAGAGATGGTCGTTG	286	244
NF1	CGAAGTGTGTGCCACTGTTT	GGTGAGACAATGGCAGGATT	179	116
NONO	CTCCGAGGAGATACCAGTCG	GTTGCCCATTTGCAGGTATT	346	232
NRCAM	TCAAACCATACAGCAGAAGCA	ACTTGCATTGCCTTCTGGAG	160	103
OSBPL6	TCAAAGAAAGACAAGCGGGT	AAGCGAACTGGTGACATGGT	185	110
PICALM	GCCCAATGATCTGCTTGATT	ATTAAGGCCAGCTGAAGGGT	292	142
PLCXD2	TTGGATCATCTAAACCGGAAA	AACTTCACGCAGGAGCTGTT	315	186
PPP1R7	AGTCGCAGGAGATGATGGAG	TCCCTATGCGATAGTGATTCAA	237	108
PTPN3	TGGAAATCCTGTGTTGAGCA	TCCGTTTCCGTGATGTAGGT	338	203
RBM15	CACTGGCCAAATCTGAAGAAG	AGGCCCATGTAAACTCCACA	256	145
RBM6	GGGATTCTCGACCTGCTAAC	TGCTGGGTTTCTCCTCACTT	1359	80
SLC15A4	CGTTAGGTGGCATTGCCTAT	GGAATCCTCAAATGAAGACTCTGT	430	261
SLC35B3	AAGTCCTGTGGCTGGTACCTT	AATTTGGTGCAGTTGTGCTG	346	191

Table3.4 (continued): RT-PCR and qRT-PCR primer sequences

Gene	Forward Primer 5' to 3'	Reverse Primer 5' to 3'	In	Sk
SOX6	GCTGCTGCTTCTGGACTCA	CATCTTTGCTCCAGGTGACA	213	90
SPAG9	GCTCTTATGGTGTCTTGTAATCG	GGACGATTTCCTGGTACACCT	149	110
STRAP	GTTGATTTGGCCTTCAGTGG	GCGTGAAATCCACAGTCTTG	277	144
SYNCRIP	CGATACCATCGGACAGGATT	TTGAAGAACTGCCAATGCAC	290	130
TMEM53	TGGACTACACCATCGAGATCC	CTCTGGCCAGGACTACTTCG	663	541
TTC7A	CTGTCGGAGGCTTTTGTCAT	TTCCTCCTCCCTCTCTGTCA	151	93
USPL1	ACGTTGCAACTAGGGTGGAG	CAAGCAGGGCAATACTCATCT	294	127
WDR85	CTGGTCGAGGTCCAAAGAAA	TCCTCCCCAACCCTCTCTAT	295	201
WDHD1	TCTTGTGGATCCTCAGTGGC	CATCATCATCCAAGTCTTCCC	193	100
ZCCHC17	AAGGCCTGAGACCATGGAA	GCCCCATAGTCTGTCACCAT	151	81

3.6 References

Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem *72*, 291-336.

Blanchette, M., and Chabot, B. (1999). Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization. EMBO J *18*, 1939-1952.

Blanchette, M., Green, R.E., MacArthur, S., Brooks, A.N., Brenner, S.E., Eisen, M.B., and Rio, D.C. (2009). Genome-wide analysis of alternative pre-mRNA splicing and RNA-binding specificities of the Drosophila hnRNP A/B family members. Mol Cell *33*, 438-449.

Burd, C.G., and Dreyfuss, G. (1994). RNA binding specificity of hnRNP A1: significance of hnRNP A1 high-affinity binding sites in pre-mRNA splicing. EMBO J *13*, 1197-1204.

Caputi, M., and Zahler, A.M. (2001). Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family. J Biol Chem *276*, 43850-43859.

Caputi, M., and Zahler, A.M. (2002). SR proteins and hnRNP H regulate the splicing of the HIV-1 tev-specific exon 6D. EMBO J *21*, 845-855.

Carpenter, B., MacKay, C., Alnabulsi, A., MacKay, M., Telfer, C., Melvin, W.T., and Murray, G.I. (2006). The roles of heterogeneous nuclear ribonucleoproteins in tumour development and progression. Biochim Biophys Acta *1765*, 85-100.

Chen, C.D., Kobayashi, R., and Helfman, D.M. (1999). Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene. Genes Dev *13*, 593-606.

Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature *460*, 479-486.

Datar, K.V., Dreyfuss, G., and Swanson, M.S. (1993). The human hnRNP M proteins: identification of a methionine/arginine-rich repeat motif in ribonucleoproteins. Nucleic acids research *21*, 439-446.

David, C.J., Chen, M., Assanah, M., Canoll, P., and Manley, J.L. (2010). HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. Nature *463*, 364-368.

Dreyfuss, G., Kim, V.N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. Nat Rev Mol Cell Biol *3*, 195-205.

Dreyfuss, G., Matunis, M.J., Pinol-Roma, S., and Burd, C.G. (1993). hnRNP proteins and the biogenesis of mRNA. Annu Rev Biochem *62*, 289-321.

Du, H., Cline, M.S., Osborne, R.J., Tuttle, D.L., Clark, T.A., Donohue, J.P., Hall, M.P., Shiue, L., Swanson, M.S., Thornton, C.A., and Ares, M., Jr. (2010). Aberrant alternative

splicing and extracellular matrix gene expression in mouse models of myotonic dystrophy. Nat Struct Mol Biol *17*, 187-193.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat Rev Cancer *4*, 177-183.

Gardina, P.J., Clark, T.A., Shimada, B., Staples, M.K., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S., Davies, C., Williams, A., and Turpaz, Y. (2006). Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics *7*, 325.

Gehman, L.T., Stoilov, P., Maguire, J., Damianov, A., Lin, C.H., Shiue, L., Ares, M., Jr., Mody, I., and Black, D.L. (2011). The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. Nat Genet *43*, 706-711.

Golan-Gerstl, R., Cohen, M., Shilo, A., Suh, S.S., Bakacs, A., Coppola, L., and Karni, R. (2011). Splicing Factor hnRNP A2/B1 Regulates Tumor Suppressor Gene Splicing and Is an Oncogenic Driver in Glioblastoma. Cancer Res *71*, 4464-4472.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M., Ulrich, A., Wardle, G.S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129-141.

Han, S.P., Tang, Y.H., and Smith, R. (2010). Functional diversity of the hnRNPs: past, present and perspectives. Biochem J *430*, 379-392.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589.

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., Jr., and Yeo, G.W. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Rep *1*, 167-178.

Hutchison, S., LeBel, C., Blanchette, M., and Chabot, B. (2002). Distinct sets of adjacent heterogeneous nuclear ribonucleoprotein (hnRNP) A1/A2 binding sites control 5' splice site selection in the hnRNP A1 mRNA precursor. J Biol Chem *277*, 29745-29752.

Ip, J.Y., Schmidt, D., Pan, Q., Ramani, A.K., Fraser, A.G., Odom, D.T., and Blencowe, B.J. (2011). Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. Genome research *21*, 390-401.

Kashima, T., and Manley, J.L. (2003). A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. Nat Genet *34*, 460-463.

Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods *7*, 1009-1015.

Kiledjian, M., and Dreyfuss, G. (1992). Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box. The EMBO journal *11*, 2655-2664.

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol *17*, 909-915.

Lagier-Tourenne, C., Polymenidou, M., and Cleveland, D.W. (2010). TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. Hum Mol Genet *19*, R46-64.

Lapuk, A., Marr, H., Jakkula, L., Pedro, H., Bhattacharya, S., Purdom, E., Hu, Z., Simpson, K., Pachter, L., Durinck, S., Wang, N., Parvin, B., Fontenay, G., Speed, T., Garbe, J., Stampfer, M., Bayandorian, H., Dorton, S., Clark, T.A., Schweitzer, A., Wyrobek, A., Feiler, H., Spellman, P., Conboy, J., and Gray, J.W. (2010). Exon-level microarray analyses identify alternative splicing programs in breast cancer. Mol Cancer Res *8*, 961-974.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., Darnell, J.C., and Darnell, R.B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464-469.

Llorian, M., Schwartz, S., Clark, T.A., Hollander, D., Tan, L.Y., Spellman, R., Gordon, A., Schweitzer, A.C., de la Grange, P., Ast, G., and Smith, C.W. (2010). Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. Nat Struct Mol Biol *17*, 1114-1123.

Martinez-Contreras, R., Cloutier, P., Shkreta, L., Fisette, J.F., Revil, T., and Chabot, B. (2007). hnRNP proteins and splicing control. Adv Exp Med Biol *623*, 123-147.

Martinez-Contreras, R., Fisette, J.F., Nasim, F.U., Madden, R., Cordeau, M., and Chabot, B. (2006). Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. PLoS Biol *4*, e21.

Mayeda, A., Munroe, S.H., Caceres, J.F., and Krainer, A.R. (1994). Function of conserved domains of hnRNP A1 and other hnRNP A/B proteins. EMBO J *13*, 5483-5495.

McGlincy, N.J., Tan, L.Y., Paul, N., Zavolan, M., Lilley, K.S., and Smith, C.W. (2010). Expression proteomics of UPF1 knockdown in HeLa cells reveals autoregulation of hnRNP A2/B1 mediated by alternative splicing resulting in nonsense-mediated mRNA decay. BMC Genomics *11*, 565.

Melko, M., and Bardoni, B. (2010). The role of G-quadruplex in RNA metabolism: involvement of FMRP and FMR2P. Biochimie *92*, 919-926.

Misquitta-Ali, C.M., Cheng, E., O'Hanlon, D., Liu, N., McGlade, C.J., Tsao, M.S., and Blencowe, B.J. (2011). Global profiling and molecular characterization of alternative splicing events misregulated in lung cancer. Mol Cell Biol *31*, 138-150.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods *5*, 621-628.

Mukherjee, N., Corcoran, D.L., Nusbaum, J.D., Reid, D.W., Georgiev, S., Hafner, M., Ascano, M., Jr., Tuschl, T., Ohler, U., and Keene, J.D. (2011). Integrative Regulatory Mapping Indicates that the RNA-Binding Protein HuR Couples Pre-mRNA Processing and mRNA Stability. Mol Cell.

Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E., and Ares, M., Jr. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev *21*, 708-718.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res *37*, e123.

Polymenidou, M., Lagier-Tourenne, C., Hutt, K.R., Huelga, S.C., Moran, J., Liang, T.Y., Ling, S.C., Sun, E., Wancewicz, E., Mazur, C., Kordasiewicz, H., Sedaghat, Y., Donohue, J.P., Shiue, L., Bennett, C.F., Yeo, G.W., and Cleveland, D.W. (2011). Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nat Neurosci *14*, 459-468.

Rossbach, O., Hung, L.H., Schreiner, S., Grishina, I., Heiner, M., Hui, J., and Bindereif, A. (2009). Auto- and cross-regulation of the hnRNP L proteins by alternative splicing. Mol Cell Biol *29*, 1442-1451.

Sanford, J.R., Wang, X., Mort, M., Vanduyn, N., Cooper, D.N., Mooney, S.D., Edenberg, H.J., and Liu, Y. (2009). Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. Genome Res *19*, 381-394.

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., and Ares, M., Jr. (2006). Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLoS Comput Biol *2*, e4.

Swanson, M.S., and Dreyfuss, G. (1988). Classification and purification of proteins of heterogeneous nuclear ribonucleoprotein particles by RNA-binding specificities. Mol Cell Biol *8*, 2237-2241.

Venables, J.P., Koh, C.S., Froehlich, U., Lapointe, E., Couture, S., Inkel, L., Bramard, A., Paquet, E.R., Watier, V., Durand, M., Lucier, J.F., Gervais-Bird, J., Tremblay, K., Prinos, P., Klinck, R., Elela, S.A., and Chabot, B. (2008). Multiple and specific mRNA processing targets for the major human hnRNP proteins. Mol Cell Biol *28*, 6033-6043.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873-881.

Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.S., Zhang, C., Yeo, G., Black, D.L., Sun, H., Fu, X.D., and Zhang, Y. (2009). Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. Mol Cell *36*, 996-1006.

Yeo, G.W., Coufal, N.G., Liang, T.Y., Peng, G.E., Fu, X.D., and Gage, F.H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat Struct Mol Biol *16*, 130-137.

4 RBP interactome analysis reveals functional RBP-RNA complexes

4.1 Background

RNA binding proteins (RBPs) interact directly and indirectly with RNA molecules to regulate their processing. Many RBPs interact with premature messenger RNAs (pre-mRNAs) in the nucleus during the process of transcription, 3' end formation and splicing, and accompany processed mRNA into the cytoplasm, where other RBPs become recruited to the RBP-mRNA complex prior to and during translation and degradation. In recent years, RBPs have become increasingly associated with a numerous and startling diversity of human diseases (Lukong et al., 2008).

RBPs have been scrutinized for decades, yet our attention has been focused largely on proteins with known or predicted RNA-binding domains (RBDs). Several studies have been undertaken to identify the complete repertoire of RBPs. In budding yeast *Saccharomyces cerevisiae*, protein microarrays identified a significant number of previously unannotated RBPs, including unconventional RNA interactions with a number of known enzymes (Scherrer et al., 2010; Tsvetanova et al., 2010). More recently, UV cross-linking of proteins to RNA, followed by oligo(dT) capture and mass spectrometry has been used to identify human proteins which bind directly to mRNA. This approach has identified nearly 900 RBPs in human HeLa cells and 555 in mouse embryonic stem cells, including more than 300 and 250 respectively that were not previously annotated as RBPs (Castello et al., 2012; Kwon et al., 2013). However, a major drawback for these studies was that oligo(dT) beads were utilized to isolate proteins which were bound to polyadenylated mature messenger RNAs, thereby precluding the identification of

proteins that interact with pre-mature, unspliced mRNAs. Another disadvantage of the majority of the previous studies is that they were not able to identify proteins that were simultaneously bound to specific RBPs of interest. To circumvent these limitations, we devised a strategy to identify protein complexes that interact with both pre-mRNA and mRNA, as well as proteins that are simultaneously bound to RNAs that are interacting with RBPs of interest, namely the heterogeneous nuclear ribonucleoparticle (hnRNP) protein family.

The hnRNP proteins represent a major class of RBPs that are of the most abundant and ubiquitously expressed proteins in the cell. As denoted by their name, hnRNP proteins are known to bind to RNAs to affect many steps of RNA processing. HnRNP protein-RNA interactomes revealed by UV crosslinking and immunoprecipitation followed by deep-sequencing (CLIP-seq) experiments have shown that hnRNP proteins are highly cooperative and redundant in their regulation of RNA targets (Huelga et al., 2012). Additionally, common RNA targets of hnRNP proteins may be due to the protein-protein interactions between the hnRNP proteins.

Here we explore the protein interactomes for ten of the major hnRNP proteins (hnRNP A0, A1, D, F, H1, K, M, R, U, UL1) as well as a more specific splicing related RBP, RBFOX2, and an RBP, UPF1, involved in the nonsense mediated decay (NMD) pathway as comparisons, using quantitative affinity purified HaloTag fusion mass spec in normal and RNase treated conditions. The HaloTag technology is an optimal technique for purification of protein interactions within complex human systems, and relies on a sepharose-based resin, HaloLink, for capture of the HaloTag protein fusions and their interactors. The rapid binding kinetics of the HaloLink as well as the strong covalent capturing, allows for even weak or transient interactors to be purified using less starting material and at the same time, while minimizing false positives (Daniels et al., 2012).

Subsequent mass spectrometry analysis determines the purified interactors. With this approach we are able to identify unique and shared RNA dependent and independent interactors. Further analysis of shared RNA dependent interactors shows enriched characteristics of known RBPs, identifying 53 interactors to be tested as candidate RNA associated proteins. Finally, through the integration of RBP-RNA targets we confirm that common RBP interactomes can reveal simultaneously bound RNAs.

4.2 Results

4.2.1 Identification of protein interactors of hnRNP, RBFOX2 and UPF1 proteins

We performed unbiased quantitative proteomics to systematically and comprehensively identify proteins in complex with ten hnRNP proteins, RBFOX2 and UPF1, using HaloTag (HT) technology as previously described (Daniels et al., 2012; Deplus et al., 2013). We transfected plasmids each expressing a full-length open-reading frame (ORF) of the RBP fused to HaloTags (HT-RBP) in human 293T cells for 24 hours, lysed the cells, and subjected half of the lysate to RNase digestion and the remaining was left untreated. The protein complexes were covalently captured on an HT affinity resin and interacting proteins were eluted and purified for mass spectrometry (LC/MS/MS) and spectral counting (Figure 4.1A). For each affinity purification, the eluted interactors were run on a gel and silver stained, enabling us to compare the RNase and non-RNase treated samples (silver stain for HT-Control and HT-hnRNPF in Figure 4.1B, with the other HT-RBPs in Figure 4.6A). After spectral counting, normalized spectral abundance counts for all experiments were largely grouped by replicates as shown by hierarchical clustering (Figure 4.6B). Outlier experiments were removed and the remaining replicate experiments were averaged together. Next we defined a set of significantly enriched protein interactors (Figure 4.1C). Within each experiment, the

distribution of $\log_2$ enrichment scores compared to the control experiment was computed for each protein interactor (HnRNPF is presented as an example in Figure 1D, and the others in Figure 4.6C). The set of significantly enriched protein interactors within each experiment was defined using a Z-score metric. Specifically, interactors were significantly enriched if they had a score greater than 1.5 standard deviations (σ) from the mean of the positive values in the distribution (red dashed line, Figure 4.1D). Interactor enrichment scores were computed for both the RNase-treated and the untreated conditions.

4.2.2 Identification of RNA binding protein interactors

To determine the number of other putative RBPs that are in complex in an RNA-dependent manner with these major RBPs, we compiled a comprehensive list of putative RBPs from the PFAM database (http://pfam.sanger.ac.uk/), selecting all proteins in the genome that contained any of the 81 known RNA binding domains (Table 4.1). This list was further filtered to remove proteins containing the RNA binding domains specific for tRNAs, snoRNAs and rRNAs, with the remaining 443 genes containing RNA binding domains predicted to bind pre-mRNA and mRNA sequences. An additional 40 putative RBPs, not containing canonical RNA binding motifs were added based on previous associations with RNA regulation and binding. These include nonsense mediated decay (NMD) pathway genes like *UPF1* and exon junction complex (EJC) genes like *MAGOH*. Our primary list of 483 RBPs was then combined with the RBP interactome defined in HeLa cells using interactome capture (Castello et al., 2012). The RBP interactome contains 845 RBPs, the majority of which overlapped our primary RBP list based on domains. The intersection of these two lists yielded a final list of 1072 RBPs (Table 4.2). 54 different RNA binding domains (RBDs) are represented in this list with 50% of the proteins (528) in the list containing at least one RBD.

For each HT-RBP, the enrichment scores for the untreated sample was compared to the RNase treated sample, for all significantly enriched protein interactors (Figure 4.1E, and others in Figure 4.6E). Expectedly, the HT-RBP was often the most enriched protein precipitated in the experiments (red dot, Figure 4.1E). We found that RNAase treatment retrieved known RNA-dependent interactors, such as the exon-junction complex protein MAGOH (green dot, Figure 4.1E), which only appeared in the untreated sample, implying an RNase-sensitive (i.e. RNA dependent) interaction with hnRNPF. Overall, we found that 25% of the protein interactors of HT-RBPs are other RBPs (blue dots, Figure 4.1E), and interestingly, the fraction of interactors that were known RBPs were roughly similar in both the untreated and RNase treated samples (Figure 4.6D). hnRNPA0 is one exception where the difference between the fraction of RBP interactors in the RNase and untreated samples is not equivalent, although the number of peptides interactors identified is unusually low in those experiments. The representation of other RBPs in our RBP-interactomes recapitulates other RBP-interactome experiments that identify many previously annotated as RNA binding protein interactors (Gregersen et al., 2014). Our results indicate that particular putative RBPs directly interact with each other, independent of RNA.

4.2.3 RNA independent and dependent interactors

Next we computed an RNA dependence score (RDS) for all protein interactors, that reflects the relative degree of how direct or independent of RNA the protein interaction is by taking the difference between the enrichment in RNAse treated sample and the untreated sample. Positive RDS (+RDS) protein interactors are less direct and more dependent on RNA, while negative RDS (–RDS) interactors are more direct and less dependent on RNA. Using this metric, we establish a set of 728 to 2013 interactors

for each RBP, representing the union of targets enriched in the RNase and untreated samples. For each RBP, a violin plot of the RDS distribution was generated (Figure 4.1F). For each RBP we observe a different preference for the types of protein interactions it makes - some proteins, like hnRNP D and hnRNP F have many more interactors that are more dependent on RNA (+RDS), while other proteins like hnRNP A0, and hnRNP A1 have many more interactors that are less dependent on RNA (–RDS) (Figure 4.1F, and Figure 4.7A). In general, a similar fraction of RBPs are represented in the RNA independent and dependent interactions (Figure 4.7B).

With this RNA dependence score, we first examined the interactions between the HT-RBPs. We found both RNA-dependent as well as more RNA independent interactions among the hnRNP proteins, with hnRNP H1 having overall more RNA independent interactions with other hnRNP proteins, and also with RBFOX2 and UPF1; and hnRNP F showing interactions with hnRNPs, as well as RBFOX2 that are slightly sensitive to RNase (Figure 4.7C). Curiously, hnRNPH1 and hnRNPF have been implicated in many RNA processing events together, since their binding motif and RNA targets are quite similar (Caputi and Zahler, 2001; Huelga et al., 2012), although hnRNPH1 is nuclear localized, while hnRNPF is more cytoplasmic (Honore et al., 2004). Consistent with the slightly more RNA dependent interactions of hnRNPF with other hnRNP proteins, our observation suggests that hnRNPF and hnRNPH1 may have redundant roles but likely do not bind their RNA substrates together. Additionally, we also recognized that hnRNPA1 and hnRNPK are often insignificant interactors in many HT-RBP experiments (grey, Figure 4.7C), although these proteins tend to be highly abundant, and are purified with the HT-control experiments, so they are therefore not considered enriched as interactors in the HT-RBP experiments by our conservative metric (see negative quadrants of Figure 4.6E). Despite not being significantly detected

with our very stringent filtering metrics in these HT-RBP experiments, we are able to independently validate interactions between hnRNPA1 and other hnRNPs using BRET analysis (Figure 4.7D). These results reveal that many of the hnRNP proteins are physically interacting in close proximity in live cells. Consistent with this finding, we noted that approximately 60% of all interactors, RNA dependent and even independent, are shared or interacting with more than one of our 12 HT-RBPs (Figure 4.1G).

For the RNA independent and dependent interactors, we assessed the gene ontology terms that were found enriched for each HT-RBP. For all HT-RBPs, in both RNA independent and dependent interactors, terms associated with RNA processing are among the most statistically significantly enriched ontologies identified (Figure 4.2A and B), which is expected for RNA-associated protein interactors. We also observed significantly enriched ontology terms that seem to be unique to particular hnRNPs (Bonferroni corrected $p<10^{-6}$, Figure 4.8A, B). Specifically for RNA dependent interactions, principal component analysis revealed two distinct groups of HT-RBPs, one set (hnRNPs A1, A0, D, K, R, UL1) is enriched for splicing and 3' end processing, while the other set (hnRNPs F, H1, M, U and UPF1 and RBFOX2) is more enriched for cell cycle related ontologies (Figure 4.2C). In support of these findings, hnRNPM has been shown to positively regulate cell cycle progression (Chen et al., 2014), while hnRNPD is known to strongly bind AU-rich elements in 3' untranslated regions to regulate RNA stability, thus interacting with many other 3' end protein interactors (Zhang et al., 1993). While hnRNPA1 (He et al., 2005) and hnRNPD (Vazquez-Chantada et al., 2010) have also been implicated in cell cycle, this association was through the regulation of RNAs related to cell cycle, and not protein interactors.

To take a closer look at splicing related interactors, we extracted all 611 proteins that are annotated as "splicing" and "spliceosomal-related" gene ontologies

(GO:0000245, GO:0000375, GO:0048025, GO:0000398, GO:0000387, GO:0000381, GO:0043484, GO:0008380) and for each HT-RBP we measured the fraction of RNA dependent interactors that were splicing and spliceosomal related (Figure 4.2D). All sets of HT-RBP RNA-dependent interactors, showed enrichment for spliceosomal interactors ($p < 0.00001$) compared to a background control (the mean fraction of spliceosomal interactors from 100 random samples of the max interactor set size, 1374), and proteins falling in the splicing related PCA delineation, hnRNPs D, K, R and UL1 showed slight increased fractions of these spliceosomal interactors. Additionally, since spliceosomal proteins are preferentially localized to the nucleus in eukaryotic cells, we have increased confidence that our RNA dependent interactors are not limited to proteins on mRNAs, but also pre-mRNAs. Since RNA dependent interactors are likely to interact directly with RNA, we examined the occurrence of 81 known RNA binding domains (RBDs, Table 1) in the set of RNA dependent interactors for each HT-RBP (Figure 4.2E). For all HT-RBPs, known RBDs are statistically significantly enriched in the interactors ($p < 0.0001$) compared to a background control interactors (two randomly sampled sets of interactors, of the max interactor set size, 1374). The most common represented domains are RRMs, KH, DEAD and LSM domains.

4.2.4 Shared interactors are enriched for candidate RNA associated proteins

The identification of RNA processing-related protein interactors as well as RNA binding domains suggested that most of the proteins interacting with HT-RBPs are likely also *bona fide* RBPs, especially for those that interact in an RNA dependent fashion. Additionally, of the interactors for each HT-RBP, we identify up to 40% of the HeLa mRNA interactome (Castello et al., 2012) (Figure 4.8C), leaving on average 70% of our RNA dependent interactors to be more closely examined as candidate RNA-associated proteins.

We observed that there are more than a thousand RNA-dependent proteins that interact with only 1 of the 12 HT-RBPs, with slightly more than half of that number that interact with 2 to 12 of our HT-RBPs (Figure 4.3A). As the number of interacting HT-RBPs increase, the fraction of known RBPs represented in the set of interactors also increase. In fact, we observe a dramatic enrichment that 70% of RNA dependent interactors interacting in all 12 HT-RBP experiments are known RBPs (Figure 4.3B). Next we ordered our nearly 3,300 +RDS interactors by the increasing numbers of HT-RBPs that they are found to interact with. With a window that encompasses 100 protein interactors, we enumerated the number that overlapped our compiled list of RBPs. For example, of the proteins that only interact with 1 HT-RBP, an average of 10 proteins (of the 100, or 10%) are known RBPs. As the window is moved from right to left towards proteins that interact with an increasing number of HT-RBPs (Figure 4.3C,D and E), we observed an increasing number of proteins (up to 50%) that are known RBPs. In fact, at the cutoff of 7 HT-RBPs, the number of known RBPs is consistently higher compared to the shuffled list of RNA-dependent interactors (Figure 4.3C). Since our current estimate that ~5% of the genome encodes proteins that are known RBPs, our whole list of proteins is already statistically significantly enriched for known RBPs (~10-20%) (Figure 4.3C, grey dashed line), thus our cutoff of 7 is very conservative. Moreover, when we enumerate the occurrence of any of 81 common RBDs within the same moving window of 100 interactors through the ranked list, we also observe an increasing number of proteins (up to 30%) that contain these known RBDs also correlated with an increase in the number of HT-RBP they interact with (Figure 4.3D). Again, at a cutoff of 7 interacting HT-RBPs, interactors consistently have more RBDs compared to the shuffled list of proteins. Finally, we analyzed the isoelectric point across our ranked interactors and also observe a significant increase in isoelectric point for proteins that interact with more than

7 of the HT-RBPs (Figure 4.3E). Next, we separated our interactors into two distinct groups, those interacting with at least 7 HT-RBPs and those interacting with less than 7 HT-RBPs and analyzed the distribution of isoelectric point for these two sets separately. Through this analysis, we also see an increase in proteins with significantly higher isoelectric point for proteins interacting with at least 7 HT-RBPs (Figure 4.3F). Based on these analyses, we name these proteins that interact with 7 or more of our HT-RBPs in an RNA dependent manner “super interactors” (SI).

Furthermore, while 93 of our 196 super interactors are known RBPs (47%, Figure 4.4B), our analysis suggests that SI that are not previously known RBPs (103, Figure 4.4C) are likely novel candidate RNA-associated proteins. Closer examination of the 103 putative RBPs reveals that despite not being originally annotated as an RBP, 50 of these interactors have a previous association with RNA through spliceosomal or ribosomal complexes (Figure 4.4C, red outlined nodes). This leaves 53 proteins of the SI set not previously associated with RNA to be validated as candidate RNA-associated proteins. Currently crosslinking immunoprecipitation studies are being conducted to confirm interaction with RNA for these putative novel RBPs.

4.2.5 Identification of known and novel functional RBP-RNA complexes

As validation of our super interactor subset and RNA dependence scores we identify many RNA dependent SI’s that are known to be components of the exon-junction complex (EJC) and nonsense-mediated decay (NMD) complex (Figure 4.4A). Core EJC components as well as poly(A)-binding protein (PABPC1) are all found to be strong RNA dependent interactors in more than 10 HT-RBP experiments (colored green, Figure 4.4A). This complex is also associated with the NMD complex through interaction with UPF3 (UPF1 RNA dependent interactor), which directly interacts with UPF2 (UPF1 RNA independent interactor), and UPF2 interacts with UPF1 (blue, Figure 4.4A) on the

RNA. UPF1 inhibits translation in *cis* and has direct interaction with endonuclease SMG6, and with the SMG7-SMG5 complex. In addition to UPF2 and 3, UPF1 shows more RNA independent interactions with NMD proteins SMG6 and SMG7 (purple, Figure 4.4A). Testament to the specificity of our assay, SMG5 (grey, Figure 4.4A) is not detected in any of our HT-RBP experiments. In addition, we have also confirmed recently reported RNA independent interactions of UPF1 with STAU1 and STAU2, and STRAP (Flury et al., 2014).

These NMD and EJC complex interactions are not unique to HT-UPF1, but are also found as interactors to other HT-hnRNP proteins. Interestingly, we see many of these NMD complex proteins interacting with hnRNPF. hnRNPF has been implicated in the regulation of alternative polyadenylation at the 3' end of genes (Alkan et al., 2006) and has shown to have strong binding preferences in the introns and 3'UTRs of RNA transcripts (Huelga et al., 2012). The interactions of hnRNPF with these NMD complex proteins suggest that hnRNPF may also play a role in the regulation of NMD. The identification of these known interactions confirms that our assays and analysis accurately identifies RNA dependent and independent RBP interactions on RNAs.

In addition to known interactions, we have identified several novel interactions, which allow us to associate putative novel functions to the HT-RBP based on the known functions of its protein interactors. As one example, we have identified that RBFOX2 interacts with a majority of the core members of the CCR4-NOT deadenylation complex, including more RNA dependent interactions with CNOT1, CNOT3, CNOT6, CNOT6L, CNOT8, CNOT10, AGO2, and PABP1, and more RNA independent interactions with CNOT7 and PAN2. These interactions predict that RBFOX2 has a role in RNA stability and decay. We have previously observed RBFOX2 binding to its RNA targets is enriched for 3'UTR binding (Lovci et al., 2013), and RNA targets with more binding in the

3'UTR show an increase in expression upon up regulation of RBFOX2 (data not shown), further implicating RBFOX2 in RNA stability. Experimental validations of this and other putative novel RBP functions are currently underway.

4.2.6 Intersection of RBP-protein and RBP-RNA interactomes

RNA dependent interactors presumably occur when proteins interact with the same RNA substrates, so we asked whether proteins with significantly overlapping sets of RNA dependent interactors also have commonly shared RNA targets (Figure 4.5A). In 2012 we had published genome-wide endogenous RNA targets in human 293T cells for 5 of the hnRNPs that we identified protein interactors for (Huelga et al., 2012). Additionally, we have recently published RNA targets in human 293T cells for RBFOX2 (Lovci et al., 2013). First we examined the number of shared RNA dependent interactors for each pair of HT-RBPs (grey shading, Figure 4.9A). Among the most highly significantly overlapping set of RNA-dependent interactors is hnRNPM and hnRNPU (Figure 4.5B left, p-value $<10^{-29}$) and hnRNPM and RBFOX2 (Figure 4.5B, right, p-value $<10^{-43}$). Interestingly, we do see statistically significant overlaps of RNA targets between hnRNPM and hnRNPU, and hnRNPM and RBFOX2 (Figure 4.5C, p-value $<10^{-100}$). This suggests that hnRNPM and hnRNPU as well as hnRNPM and RBFOX2 are indeed binding a common set of RNAs, and thus have many common RNA dependent protein interactors, suggesting that the two proteins may act together on their targets. Furthermore, identification of shared protein interactors may allow us to predict if the RNA targets are also significantly overlapping. When we expanded this analysis to all pairwise combinations of proteins we had RNA targets for, we do see a positive correlation of the fraction of overlapping RNA targets and protein interactors (Figure 4.5D). Z-score analysis confirms that this positive correlation is slight but statistically significant (p<0.05, Figure 4.5D, inset).

Finally, we also identified a significantly larger fraction of all protein interactors that were also CLIP targets for hnRNPF and HRNNPH, and hnRNPU, RBFOX2 ($p < 10^{-100}$, Figure 4.9B). This is suggestive of a model where RNA and protein targets may be in close proximity within the cell, allowing for interaction to take place as soon as an RNA target is translated the hnRNP can also bind the protein. RNA targets in some cases may act as guides that bring along future protein interactors of the translated mRNA molecule. Experimental validations identifying co-localized RNA targets and protein interactors will be necessary to confirm these observations.

4.3 Discussion

We have identified RBP-interactomes for 10 major hnRNP proteins as well as RBFOX2 and UPF1 using a HT purification approach followed by mass spectrometry analysis. Using a combination of RNase treated and untreated samples, we are able to score interactions that are more dependent on RNA. This analysis has confirmed many known RNA dependent and independent interactions that have been previously identified, as well has implicated novel roles for hnRNPF in NMD and RBFOX2 in RNA stability.

Integrative analysis of RNA dependent interactors has identified a set of “super interactors” that interact with 7 or more of our 10 HT-RBPs in an RNA-dependent manner, with characteristics of known RNA binding proteins, containing RBDs and higher isoelectric points. Our approach has therefore identified additional candidate RNA associated proteins, adding to the growing list of proteins that interact with RNA. Finally, by combining interacting proteins and RNA targets identified previously (Huelga et al., 2012; Lovci et al., 2013), we are able to detect a positive correlation between overlaps of interactors and targets, which may suggest simultaneous binding to RNA targets.

Further experimental exploration and analysis of these stimulating findings will be necessary.

Taken together, we have integrated several datasets for multiple RBPs and multiple data types (mass spectrometry and CLIP-seq) to reveal an unprecedented number of RBP-RNA complexes. Identification of these specific interaction complexes will allow us to more carefully detail the functions of these RBPs. This approach presents the first steps toward a complete understanding of RBP function. There are approximately 1000 RBPs in humans, many which still have unknown functions, so we are beginning efforts to expand this interactome approach to more RBPs. As our understanding of the functions of RBP-RNA complexes improves, we can develop an improved understanding of RNA processing and gene regulation important for human disease.

4.4 Methods

4.4.1 Halo-Tag Affinity Purification

Full length Halo tag fusion vectors for hnRNP A0, A1, D0, F, H1, K, M, R, U, UL, RBFOX2, and UPF1 were each created and transfected into human 293T cells. Halo-tag affinity purification was performed as previously described (Deplus et al., 2013) with and without RNase treatment in replicate. Additional control experiments with Halo-tag only were also done with and without RNase treatment in replicate. Affinity purified complexes were than analyzed by nano LC/MS/MS (MSBioworks) as previously described (Deplus et al., 2013).

4.4.2 Mass Spectrometry data analysis

Spectral counting was performed and normalized spectral abundance factors determined (dNSAF values) as previously described (Deplus et al., 2013). False

positives/contaminants were removed by comparing to control experiments. Any potential protein interactor with a dNSAF value that was not 2x greater than its value in any of the control corresponding experiments was assigned a 0 value.

For equivalent comparison to the RBP RNA targets, the lists of protein-protein interactors were filtered for those targets that are expressed at a detectable RNA level in 293T cell RNA-seq data (Huelga et al., 2012). Replicates were clustered, outlier datasets (UPF1_RNASE_2, M_RNASE_1, H_RNASE_1, D0_1) were removed, and the remaining replicates were averaged together. An enrichment score was calculated for each protein interactor by calculating a log2 ratio of the dNSAF value in the RBP purification experiment compared to the Control experiment:

Enrichment Score = log2[(dNSAF RBP + 0.000001)/(dNSAF Control + 0.000001)]

A psuedocount of 0.000001 representing the lowest measurable dNSAF value was added within the enrichment score to prevent losing targets with zero values in the Control experiment (divide by zero errors). Enrichment distributions were created for each data set and a mean and standard deviation was computed for interactors with enrichment scores greater than 0 (enriched in RBP experiment). The final set of protein-protein interactors was determined using all protein interactors with enrichment greater than the $\mu–1.5\sigma$ in each dataset.

RNA dependence score was calculated for each protein interactor as the difference in enrichment between the +RNAse and the –RNase experiment (–RNAse - +RNase). If there is no enrichment in the –RNAse experiment the RNA dependence score is just the enrichment in the (-)RNAse experiment, and if there is no enrichment in the (+)RNase experiment the RNA dependence score is the negative value of the +RNase enrichment score. RNA dependence scores that are more positive are representative of protein interactors that are less direct and more dependent on RNA,

while RNA dependence scores that are more negative are representative of protein interactors that are more direct and less dependent on RNA.

4.4.3 RNA target analysis

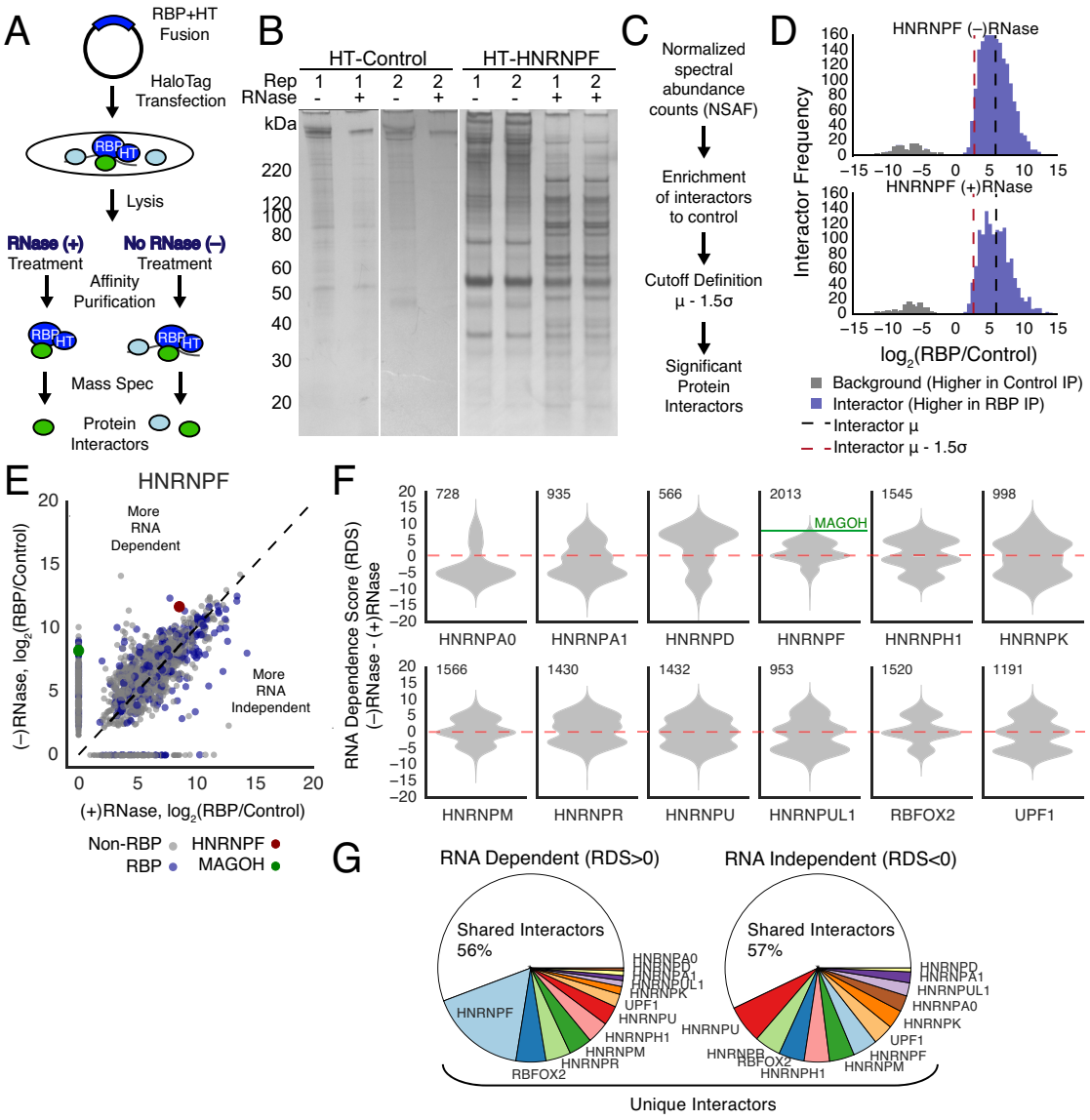
The set of RNA targets for hnRNP A1, F, H1, M and U were obtained from CLIP-seq data published in (Huelga et al., 2012). RNA targets for RBFOX2 were obtained from iCLIP data published in (Lovci et al., 2013).

4.5 RBP-RNA interactome acknowledgement

Chapter 4 is in preparation to be submitted for publication in *Cell*. This dissertation author was the primary researcher/author who directed the experiments, analyzed the data, and wrote the text. Co-authors who have also contributed to this work include Kristopher Brannan (Yeo Lab, UC San Diego), Mike Washburn (Stowers Institute), and Danette Daniels (Promega).

Figure 4.1: Identification of enriched RBP-protein interactors that are more RNA dependent or independent

A. Halo-Tag IP Mass Spec experimental procedure. RBP-Halo Tag fusion protein constructs are transfected into human 293T cells in replicate. Halo-Tag fusion proteins are affinity purified from cells, lysed, and half the lysate is treated with RNase . Purified products are subjected to LC/MS/MS to identify protein interactors in RNase and non-RNase treated conditions.
B. Silver stain gel of replicate (Rep) Halo-Tagged hnRNPF (HT-hnRNPF) protein complex isolations from RNase treated (+) and untreated (–) cells. Replicate control isolation with Halo-Tag only (HT-Control) also from RNase treated (+) and untreated (–) cells. See Figure 4.6A for additional silver stain gels.
C. Analysis flowchart for post processing of MS data. After determining normalized spectral abundance counts (NSAF), enrichment of interactors compared to the control are filtered and computed, and those that pass a μ–1.5σ cutoff are designated significant protein interactors
D. Distribution of enrichment scores compared to control for hnRNPF affinity purifications in RNase treated and untreated conditions. Grey data points (enrichment <0) represent background or contaminant data enriched in Control experiments compared to the hnRNPF experiment and are removed from the analysis. For protein interactors with an enrichment score greater than zero (enriched in hnRNPF experiment compared to Control, blue) a mean (μ, black dashed line) and standard deviation (σ) is computed. A protein interactor is significantly enriched if the enrichment score is greater than 1.5 times the standard deviation (1.5σ, red dashed line).
E. Scatter plot of enrichment scores for significant protein targets in both (+)RNase treated and (–)RNase untreated experiments. Blue protein interactors are known RNA binding proteins (RBPs), grey targets are non-RBP proteins, the red interactor is the affinity purified protein hnRNPF, and the green interactor is the RNA dependent super interactor, MAGOH.
F. Violin plots of RNA Dependence scores (RDS) or the difference between the enrichment in the non-RNase and RNase treated sample, (–)RNase – (+)RNase.. The red line delineates interactors that are more RNA dependent (+RDS) and independent (–RDS), with some Halo-Tagged RBPs showing many more RNA dependent or independent interactions. The number in the corner represents the total interactors that are either RNA dependent of independent. The green line for hnRNP F corresponds to the green dot labeled in Figure 4.1E, the RNA dependent super interactor, MAGOH.
G. Pie charts displaying the percent of interactions that are shared RNA dependent or RNA independent between 2 or more Halo-Tagged RBPs. Interactors that are uniquely interacting with only one of the Halo-tagged RBPs are colored.


A
RBP+HT Fusion
HaloTag Transfection
Lysis
RNase (+) Treatment
No RNase (–) Treatment
Affinity Purification
Mass Spec
Protein Interactors
B
HT-Control
HT-HNRNPF
Rep
RNase
kDa
220
120
100
80
60
50
40
30
20
C
Normalized spectral abundance counts (NSAF)
Enrichment of interactors to control
Cutoff Definition μ - 1.5σ
Significant Protein Interactors
D
HNRNPF (–)RNase
HNRNPF (+)RNase
Interactor Frequency
log2(RBP/Control)
Background (Higher in Control IP)
Interactor (Higher in RBP IP)
Interactor μ
Interactor μ - 1.5σ
E
HNRNPF
More RNA Dependent
More RNA Independent
(–)RNase, log2(RBP/Control)
(+)RNase, log2(RBP/Control)
Non-RBP
RBP
HNRNPF
MAGOH
F
RNA Dependence Score (RDS) (–)RNase - (+)RNase
728
935
566
2013
MAGOH
1545
998
HNRNPA0
HNRNPA1
HNRNPD
HNRNPF
HNRNPH1
HNRNPK
1566
1430
1432
953
1520
1191
HNRNPM
HNRNPR
HNRNPU
HNRNPUL1
RBFOX2
UPF1
G
RNA Dependent (RDS>0)
RNA Independent (RDS<0)
Shared Interactors 56%
Shared Interactors 57%
Unique Interactors

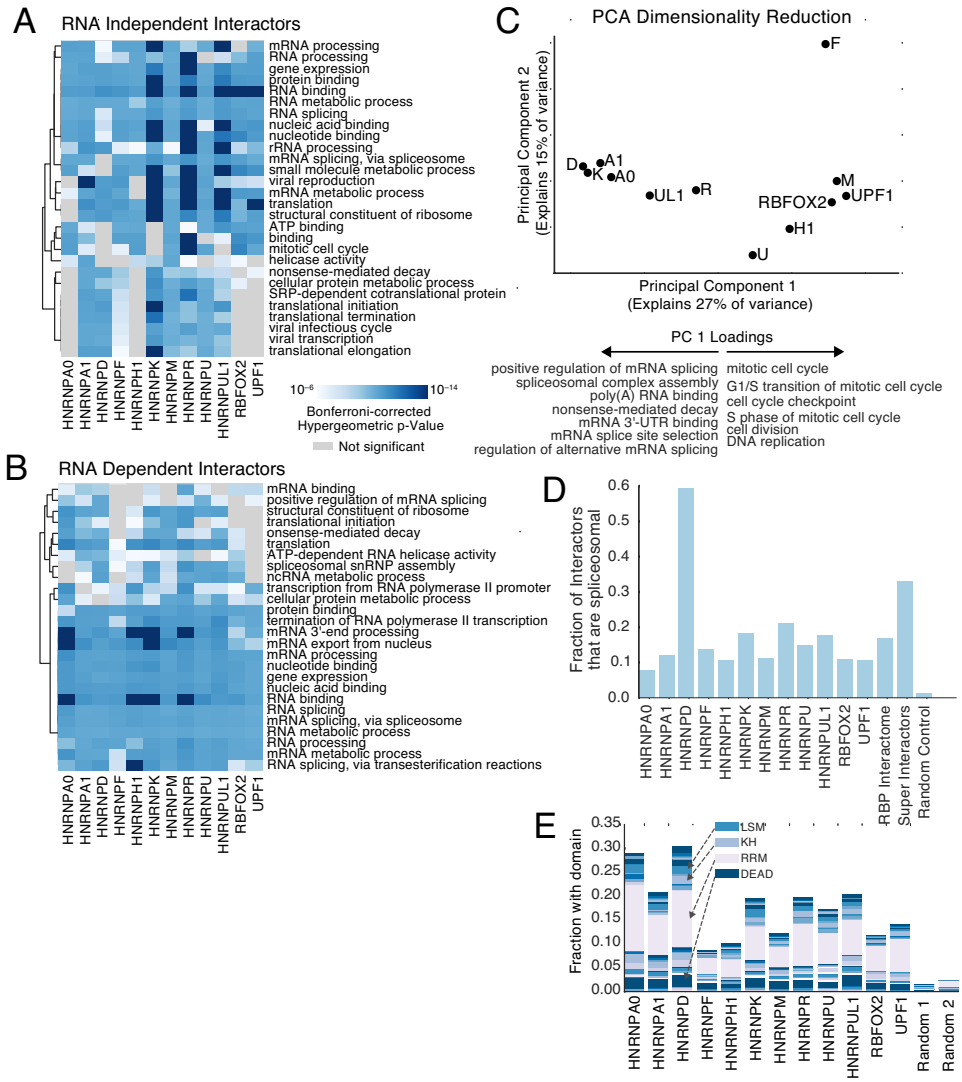


Figure 4.2. Characterization of RNA independent and dependent interactors

Common gene ontology terms among (**A**) RNA independent protein interactors (protein interactors with –RDS), or (**B**) RNA dependent protein interactors (protein interactors with +RDS). Values are bonferroni-corrected hypergeometric *p*-values. The darker the blue, the more significantly enriched the GO term within the set of protein interactors. Grey values are not significant in the given experiment. **C.** Principal component analysis (PCA) dimensionality reduction of RNA dependent interactor enriched gene ontology terms. On the x-axis, principal component (PC 1), splicing and cell cycle gene ontologies separate the HT-RBPs. **D.** Fraction of 611 spliceosomal genes that interact with each Halo-Tagged RBP dependent on RNA, appear in the mRNA RBP interactome (Castello et al., 2012), are a super interactor (interacting with >7 Halo-Tagged-RBPs) or exist in a background control. This control fraction is the mean fraction of spliceosomal interactors from 100 random samples of the max interactor set size, 1374. All sets of interactors are enriched for spliceosomal interactors compared to the control ($p < 0.00001$). **E.** Fraction of RNA dependent interactors that contain any of 81 known RNA binding domains. The RRM, KH, DEAD and LSM domains are among the most commonly represented domains.

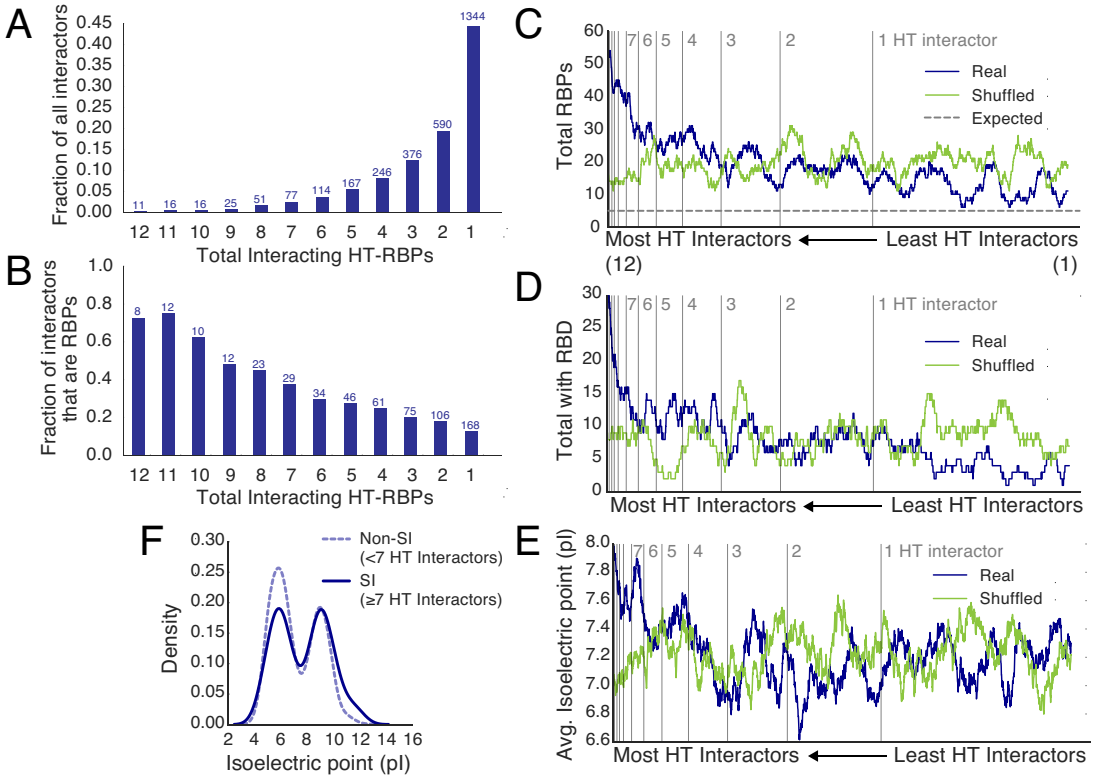


Figure 4.3: Super interactors are enriched for characteristics of known RBPs

A. Bar chart displays the fraction of all RNA dependent interacting proteins that come up in 1 (unique interactor), and 2 to 12 Halo-Tag-RBP experiments (shared interactor). The number of interactors is given at the top of each bar. **B.** Bar chart displays the fraction of unique (1 HT-RBP) and shared (2-12 HT-RBPs) RNA dependent interacting proteins that are RBPs. The number of RBP interactors is given at the top of each bar.
C. Correlation between the number of Halo-Tag-RBP RNA dependent interactors and RBP interactors. Interactors were ranked by the number of HT-RBPs they interacted with (*x* axis, 12 to 1 denoted by vertical grey bars) and the total number of RBP interactors found in the next 100 genes from the ranked list was plotted (*y* axis, dark blue line). As a control a random shuffling of the ranks was also plotted (green line) as well as the genome-wide expected count of RBPs (grey dashed line). **D.** Correlation between the number of Halo-Tag-RBP RNA dependent interactors and common RNA binding domains. Interactors were ranked by the number of HT-RBPs they interacted with (*x* axis, 12 to 1 denoted by vertical grey bars) and the total number of interactors containing RBDs found in the next 100 genes from the ranked list was plotted (*y* axis, dark blue line). As a control a random shuffling of the ranks was also plotted (green line).
E. Correlation between the number of Halo-Tag-RBP RNA dependent interactors and interactor isoelectric point. Interactors were ranked by the number of HT-RBPs they interacted with (*x* axis, 12 to 1 denoted by vertical grey bars) and the average isoelectric point found in the next 100 genes from the ranked list were plotted (*y* axis, dark blue line). As a control a random shuffling of the ranks was also plotted (green line).
F. Density of the calculated isoelectric points (pI) of RNA dependent super interacting proteins (solid blue line), and non-super interacting proteins (dashed line).

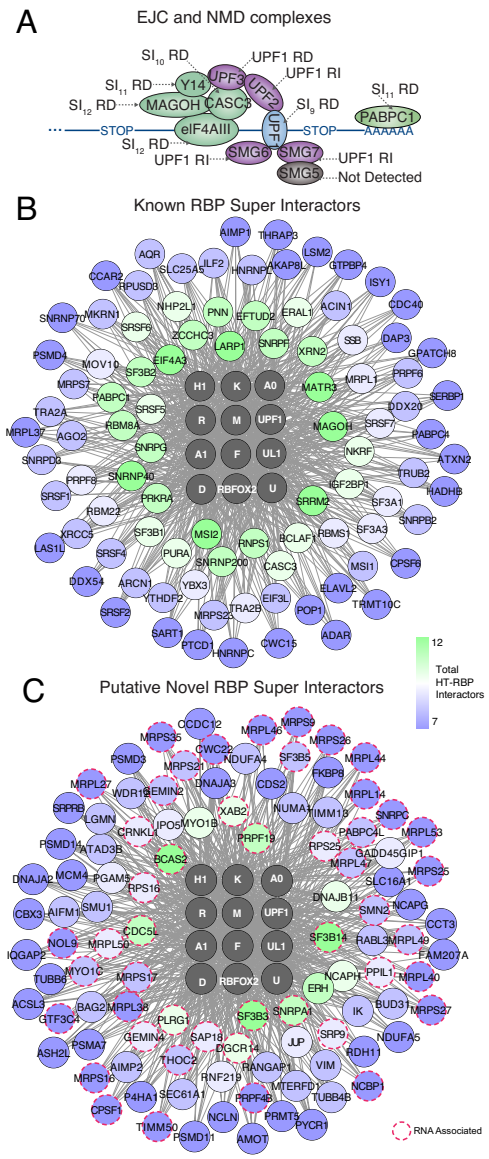


Figure 4.4: RNA dependent super interactors reveal known RBP-RBP complexes and candidate RNA associated proteins

A. The protein interactions of the exon-junction complex (EJC) and nonsense-mediated decay (NMD) complex bound to RNA. Super interacting (SI) proteins are colored green and are labeled with the number of Halo-tagged RBPs they interact with. Detectable UPF1 (in blue) RNA dependent (RD) and independent (RI) interacting proteins are colored in purple. NMD component, SMG5 is undetectable in our MS experiments. Figure adapted from (Kervestin and Jacobson, 2012). **B**. Super Interactor proteins (interacting with 7 or more of the Halo-tagged RBPs) that are known RBPs. The more green the node, the more Halo-tagged RBPs it interacts with. **C**. Super Interactor proteins (interacting with 7 or more of the Halo-tagged RBPs) that are candidate RNA associated proteins. The more green the node, the more Halo-tagged RBPs it interacts with. Interactors with red-dashed circles represent interactors that have shown previous associations with RNA, mainly through interactions with spliceosomal complexes.

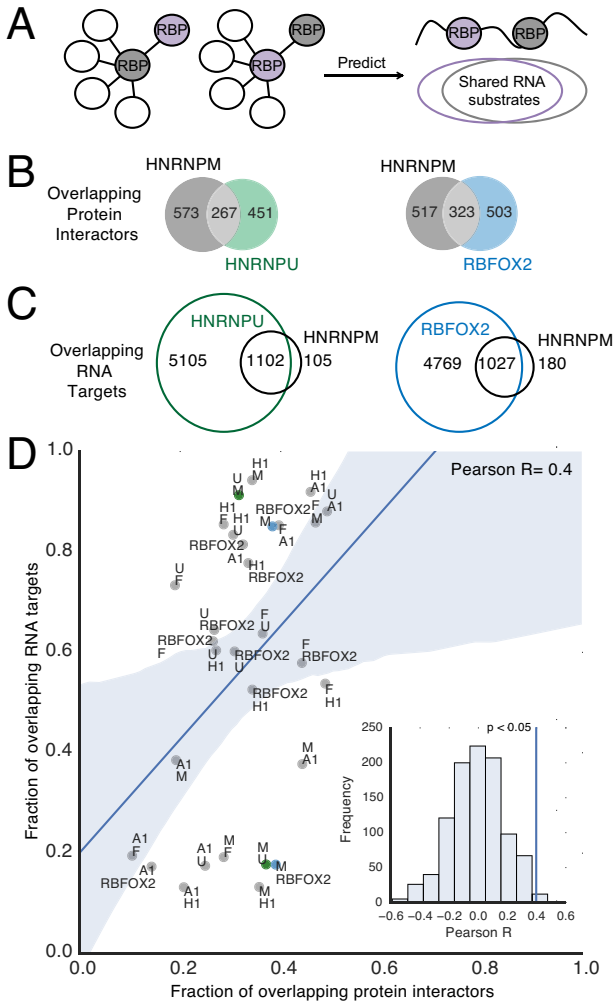
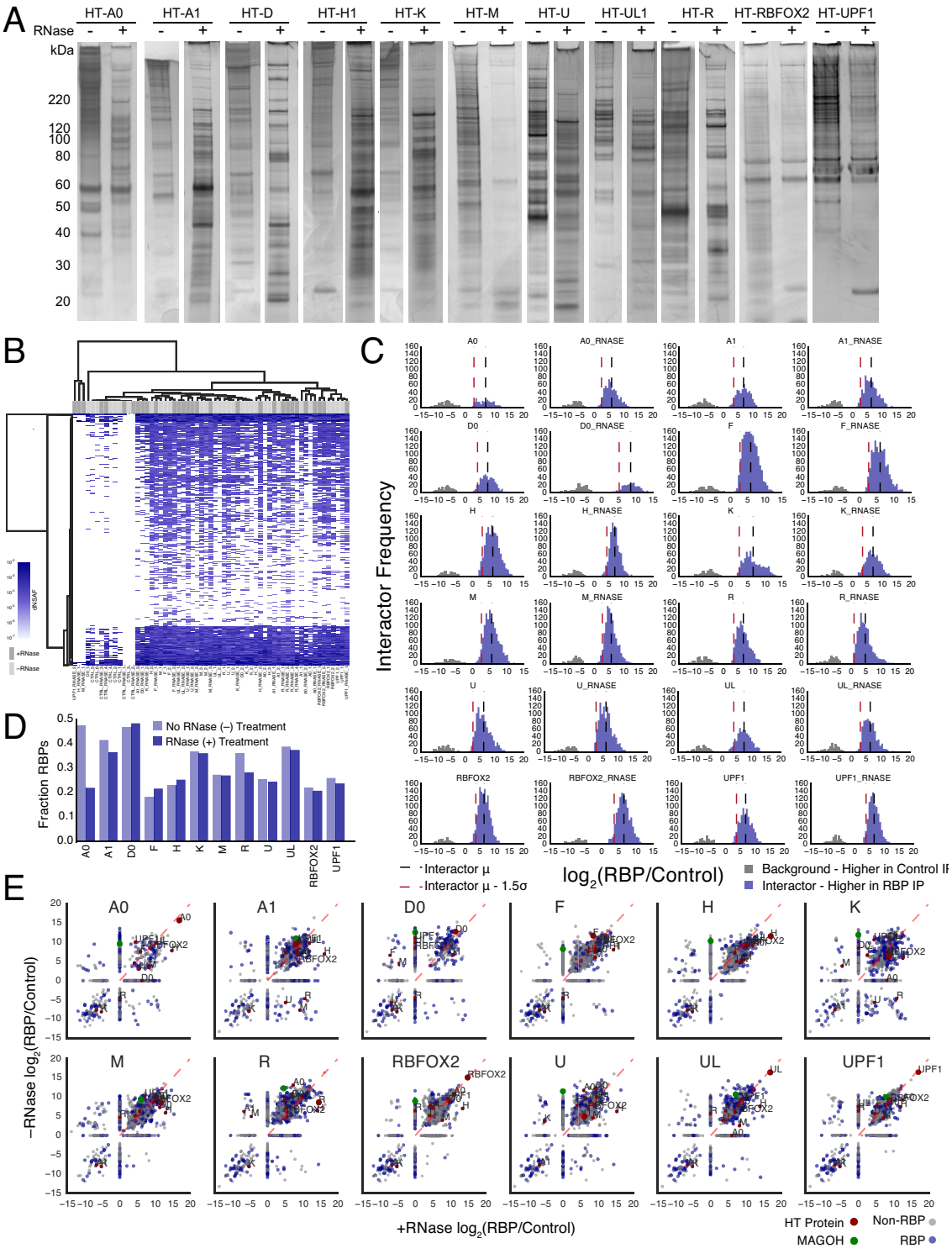


Figure 4.5: Intersection of protein interactors and RNA targets

A. Model describing when two RBPs (purple and grey) interact with a common set of RNA dependent interactors (white), and each other, we can predict that the two RBPs also have a shared set of RNA substrates. **B**. Comparisons of the overlap of RNA dependent protein interactors between hnRNPM and hnRNPU (green, $p < 10^{-29}$), and between hnRNPM and RBFOX2 (blue, p-value $< 10^{-43}$).. Values listed are the total overlapping and non-overlapping interactors. These two overlapping sets are among the top most significant overlaps, between all HT-RBPs (Figure 4.9A).
C. Venn diagram representing overlap of RNA targets determined by CLIP-seq (Huelga et al., 2012; Lovci et al., 2013). Overlaps are significant by chi-square ($p < 10^{-100}$).
D. Linear regression plot of the fraction of overlapping protein interactors (x-axis) versus the fraction of overlapping RNA targets (y-axis). Each point is labeled with the pair that were overlapped, with the fraction representing the total overlapping, divided by the total targets or interactors of the protein listed on the bottom. Blue points and green points correspond to the overlaps in (**B**) and (**C**). Blue line represents fitted regression line, while blue shading represents 95% confidence intervals, Pearson R=0.4. Inset shows the distribution of Pearson R correlation values for overlapping protein fractions and overlapping RNA fractions shuffled 1000 times. Blue line denotes observed Pearson R of 0.4 ($p < 0.5$).

Figure 4.6: Identification of RBP interactomes for 12 RBPs

A. Silver stain gel for each Halo-Tagged RBP protein complex isolations from RNase treated (+) and untreated (–) cells. Replicate control isolation with Halo-Tag only (HT-Control) and HT-hnRNPF also from RNase treated (+) and untreated (–) cells is shown in Figure 4.1B.
B. Hierarchical clustering of contaminant filtered dNSAF values. Dark grey labels RNase treated samples, while light grey labels non-RNase treated samples. Outliers on far left are removed from downstream analyses. Remaining replicates cluster well and are averaged in downstream analysis.
C. Distribution of enrichment scores compared to control for each Halo-Tagged RBP in RNase treated and untreated conditions. Grey data points (enrichment <0) represent background or contaminant data enriched in Control experiments compared to Halo-Tag-RBP experiment and are removed from the analysis. For protein interactors with enrichment score greater than zero (enriched in Halo-Tag-RBP experiment compared to Control, blue) a mean (μ, black dashed line) and standard deviation (σ) is computed. A protein interactor is significantly enriched if the enrichment score is greater than 1.5 times the standard deviation (1.5σ, red dashed line).
D. Fraction of significant interactors that are RBPs in each experiment.
E. Scatter plots of enrichment scores for all protein interactors in both (+)RNase treated and (–)RNase untreated experiments. Blue protein interactors are known RBPs (RBP), grey targets are non-RBP proteins, and the green interactor is the RNA dependent super interactor, MAGOH. Red interactors label all 12 affinity purified RBPs within each experiment. Significant interactors appear only in the upper right quadrant of the scatter plots, while the remaining quadrants show interactors that were purified in Control experiments. Note that commonly, hnRNPK and hnRNPA1 appear in the Control experiments and thus are not often enriched interactors.


A
HT-A0
HT-A1
HT-D
HT-H1
HT-K
HT-M
HT-U
HT-UL1
HT-R
HT-RBFOX2
HT-UPF1
RNase
kDa
220
120
100
80
60
50
40
30
20
B
C
Interactor Frequency
A0
A0_RNASE
A1
A1_RNASE
D0
D0_RNASE
F
F_RNASE
H
H_RNASE
K
K_RNASE
M
M_RNASE
R
R_RNASE
U
U_RNASE
UL
UL_RNASE
RBFOX2
RBFOX2_RNASE
UPF1
UPF1_RNASE
Interactor μ
Interactor μ - 1.5σ
log2(RBP/Control)
Background - Higher in Control IP
Interactor - Higher in RBP IP
D
Fraction RBPs
No RNase (–) Treatment
RNase (+) Treatment
E
−RNase log2(RBP/Control)
+RNase log2(RBP/Control)
HT Protein
Non-RBP
MAGOH
RBP

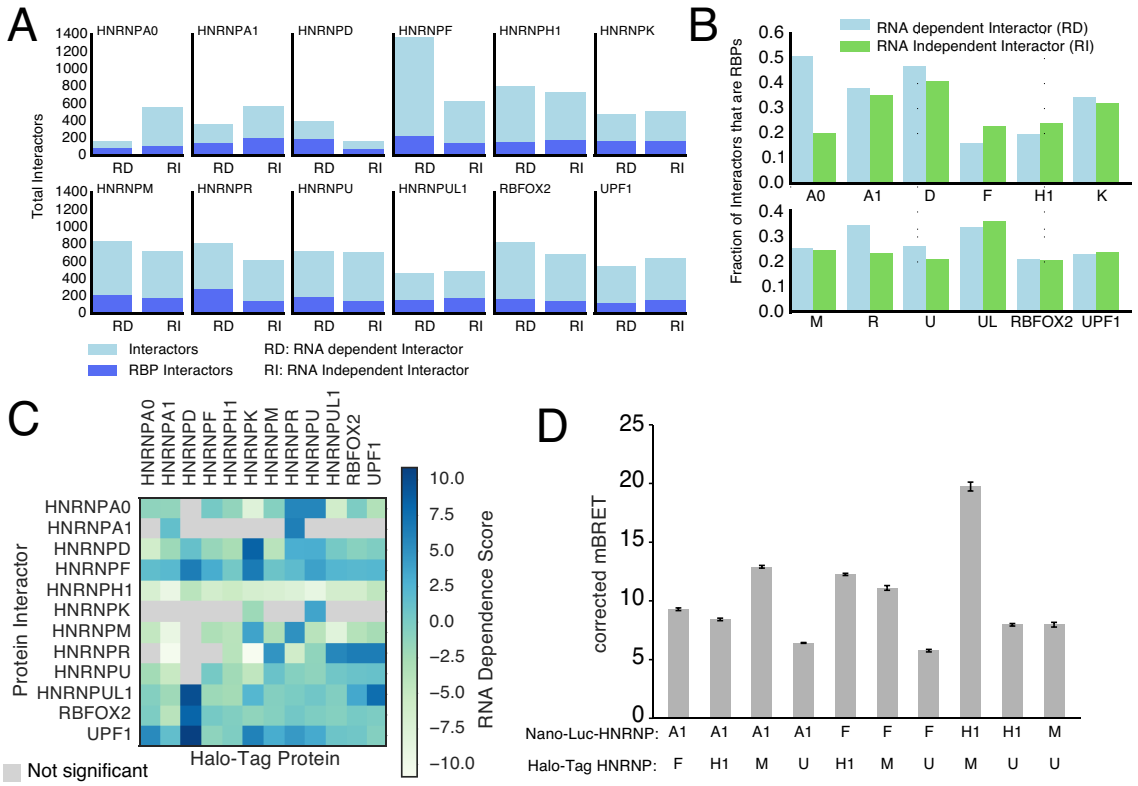


Figure 4.7: RBP interactions

A. Bar charts displaying the proportion of total RNA independent (RI) and RNA dependent (RD) interactors that are RBPs.
B. Bar charts displaying the fraction of total RNA independent (RI) and RNA dependent (RD) interactors that are RBPs.
C. Heatmap of RNA dependency score between the Halo-Tagged RBPs purified. More blue values represent more RNA dependent interactions. Grey values represent a non-significant interaction.
D. hnRNP interaction validations monitored using BRET assay. Values are corrected BRET measurements. X-axis represents different pairs of hnRNP interactions tested NanoLuc-hnRNP/Halo-Tag hnRNP.

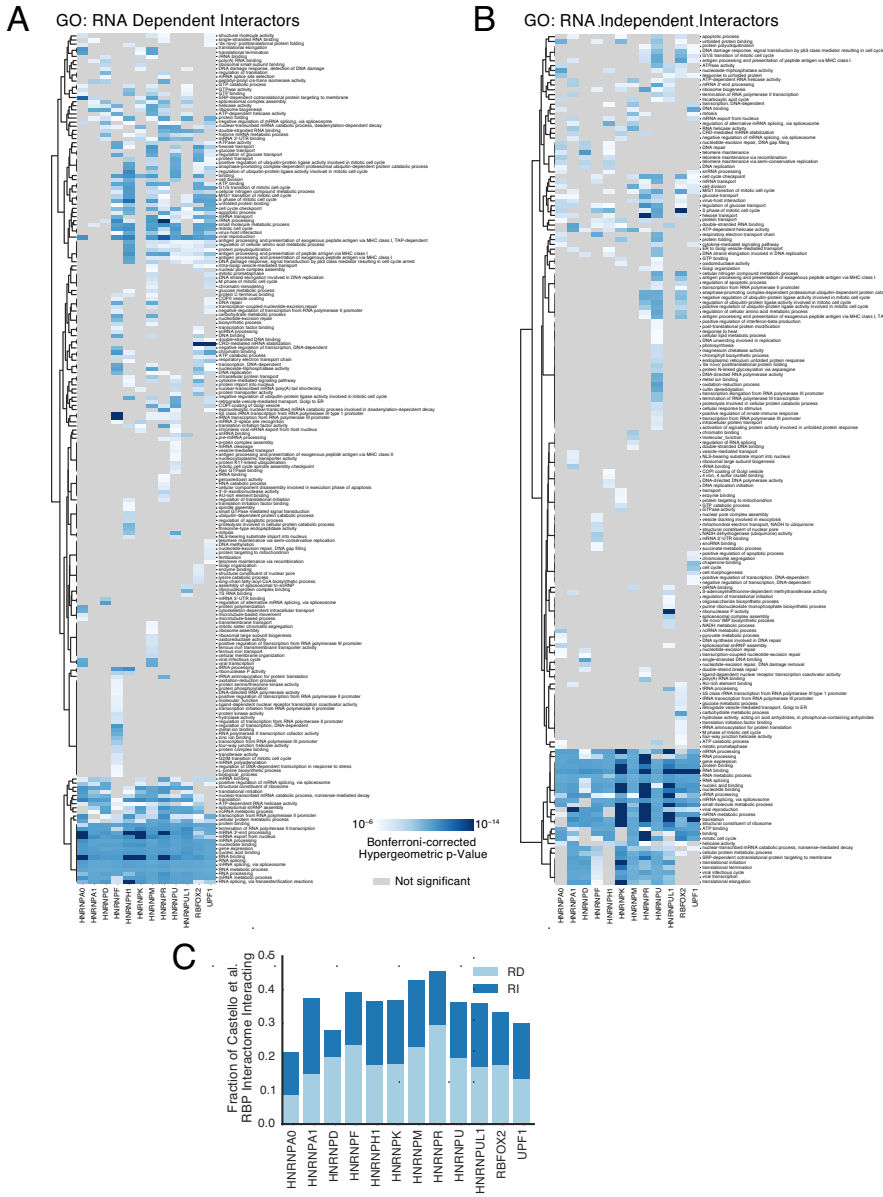


Figure 4.8: Gene Ontology analysis of interactors

All enriched gene ontology terms among (**A**) RNA dependent protein interactors (protein interactors with +RDS), or (**B**) RNA dependent protein interactors (protein interactors with –RDS). Values are bonferroni-corrected hypergeometric *p*-values. The darker the blue, the more significantly enriched the GO term within the set of protein interactors. Grey values are not significant in the given experiment.
C. Fraction of more RNA dependent interactors (RD, dark blue) and more RNA independent interactors (RI, light blue) that come out in the mRNA RBP interactome (Castello et al., 2012).

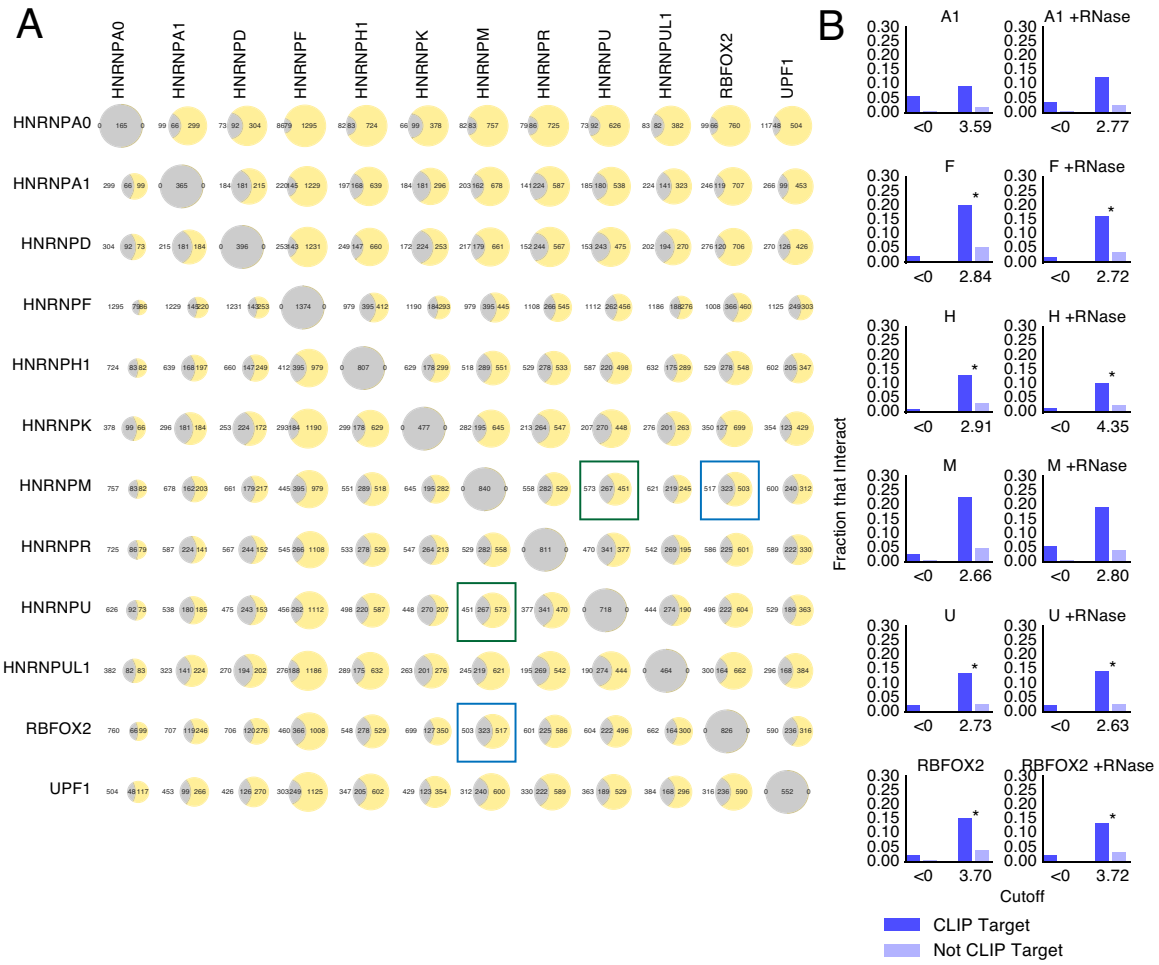


Figure 4.9. Pairwise HT-RBP interactor overlap and RNA target overlap

A. Comparisons of the overlap of RNA dependent protein interactors between all pairs of HT-RBP proteins. Circles represent the amount of interactors for the HT-RBP listed in the column (sized according to the relative number of interactors), with the grey shading representing the amount of overlap with the HT-RBP listed in the row. Values listed are the total overlapping and non-overlapping interactors. Overlaps boxed in are among the top most significant overlaps, between hnRNPM and hnRNPU (green, $p < 10^{-29}$), and between hnRNPM and RBFOX2 (blue, p-value $< 10^{-43}$).
B. Fraction of non-significant (<0) and significant (>1.5σ) protein interactors that are also hnRNP RNA targets (CLIP). Values with a * are significant by Chi-square test ($p < 10^{-100}$).

Table 4.1: PFAM RNA binding associated domains

PFAM ID	Domain Description	PFAM ID	Domain Description
PF00013	KH_1	PF05383	La domain
PF00035	dsrm	PF05731	TROVE
PF00076	RRM_1	PF05741	zf-nanos
PF00098	zf-CCHC	PF06859	Bicoid-interacting protein 3 (Bin3)
PF00181	Ribosomal_L2	PF07497	Rho termination factor, RNA-binding
PF00270	DEAD	PF07498	Rho termination factor, N-terminal
PF00298	Ribosomal_L11	PF07650	KH_2
PF00567	TUDOR	PF08081	RBM1CTR (NUC064) family
PF00575	S1	PF08156	NOP5NT (NUC127) domain
PF00658	PABP	PF08210	APOBEC-like N-terminal domain
PF00806	PUF	PF08228	RNase P subunit Pop3
PF01142	tRNA pseudouridine synthase D	PF08373	RAP domain
PF01423	LSM domain	PF08492	SRP72 RNA-binding domain
PF01585	G-patch	PF08517	Ataxin-1 and HBP1 module (AXH)
PF01588	tRNA_bind	PF08675	RNA binding domain
PF01798	Nop	PF08710	nsp9 replicase
PF01805	Surp module	PF08772	Nin one binding, NOB1 Zn-ribbon
PF02081	Tryptophan RNA-binding attenuator	PF08777	RNA binding motif
PF02170	PAZ domain	PF09105	Elongation factor SelB, winged helix
PF02171	Piwi	PF09106	Elongation factor SelB, winged helix
PF02272	DHHA1 domain	PF09107	Elongation factor SelB, winged helix
PF02290	Signal recognition particle 14kD	PF09157	Pseudouridine synthase II TruB, C-term
PF02599	Global regulator protein family	PF09162	Tap, RNA-binding
PF02792	Mago nashi protein	PF09246	PHAT
PF03123	CAT_RBD	PF09369	Domain of unknown func DUF1998
PF03368	Double stranded RNA binding	PF09387	Mitochondrial RNA binding MRP
PF03501	Plectin/S10 domain	PF09738	Double stranded RNA binding
PF03726	PNPase	PF10133	Predicted RNA-binding protein
PF03880	DbpA	PF10258	PHAX RNA-binding domain
PF04032	RNAse P subunit domain	PF10330	Putative Sin3 binding protein
PF04059	RRM_2	PF10373	Est1 DNA/RNA binding domain
PF04135	Nucleolar RNA-binding protein	PF10378	Putative RRM domain
PF04266	ASCH domain	PF10458	Valyl tRNA synthetase tRNA
PF04410	Gar1	PF10597	U5-snRNA binding site 2 of PrP8
PF04557	tRNA_synt_1c_R2	PF10598	RNA recognition spliceosomal PrP8
PF04558	tRNA_synt_1c_R1	PF11435	RNA binding protein She2p
PF04774	HABP4_PAI-RBP1	PF11717	RNA activity-knot chromodomain
PF04801	Sin-like protein conserved	PF12009	Telomerase ribonucleoprotein
PF04845	PurA ssDNA and RNA-binding	PF12171	Zinc-finger double-stranded RNA
PF04857	CAF1 family ribonuclease	PF12212	Piwi/Ago/Zwille siRNA-binding
PF04926	PAP_RNA-bind	PF12220	U1 small nuclear ribonucleoprotein
PF05172	MPPN (rrm-like) domain	PF12235	Fragile X-related 1,C terminal
PF05253	U11-48K-like CHHC zinc finger	PF12287	Cytoplasmic activation/proliferation

Table 4.2. Compiled set of 1072 putative RNA binding proteins

For each RBP, the Ensembl ID, gene symbol and gene description is given, along with the PFAM IDs for any of the known RNA binding domains (**Table 4.1**) contained in the protein. In the final column, a 1 is listed if the protein was identified in the mRNA-RBP interactome (Castello et al., 2012) and 0 if not.

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000090861	AARS	alanyl-tRNA synthetase		0
ENSG00000108270	AATF	apoptosis antagonizing transcription factor		1
ENSG00000204574	ABCF1	ATP-binding cassette, sub-family F (GCN20), member 1		1
ENSG00000146109	ABT1	activator of basal transcription 1		1
ENSG00000167315	ACAA2	acetyl-CoA acyltransferase 2		1
ENSG00000100813	ACIN1	apoptotic chromatin condensation inducer 1		1
ENSG00000130402	ACTN4	actinin, alpha 4		1
ENSG00000164113	ADAD1	adenosine deaminase domain containing 1 (testis-specific)	PF00035	0
ENSG00000140955	ADAD2	adenosine deaminase domain containing 2	PF00035	0
ENSG00000160710	ADAR	adenosine deaminase, RNA-specific	PF00035	1
ENSG00000197381	ADARB1	adenosine deaminase, RNA-specific, B1	PF00035	0
ENSG00000185736	ADARB2	adenosine deaminase, RNA-specific, B2	PF00035	1
ENSG00000087274	ADD1	adducin 1 (alpha)		1
ENSG00000156110	ADK	adenosine kinase		1
ENSG00000164252	AGGF1	angiogenic factor with G patch and FHA domains 1	PF01585	0
ENSG00000124942	AHNAK	AHNAK nucleoprotein		1
ENSG00000111732	AICDA	activation-induced cytidine deaminase	PF08210	0
ENSG00000164022	AIMP1	aminoacyl tRNA synthetase complex-interacting multifunctional protein 1		0
ENSG00000121057	AKAP1	A kinase (PRKA) anchor protein 1	PF00567, PF00013	1
ENSG00000105127	AKAP8	A kinase (PRKA) anchor protein 8		1
ENSG00000011243	AKAP8L	A kinase (PRKA) anchor protein 8-like		1
ENSG00000059573	ALDH18A1	aldehyde dehydrogenase 18 family, member A1		1
ENSG00000119711	ALDH6A1	aldehyde dehydrogenase 6 family, member A1		1
ENSG00000091542	ALKBH5	alkB, alkylation repair homolog 5 (E. coli)		1
ENSG00000259224	SLC35G6	solute carrier family 35, member G6		1
ENSG00000131503	ANKHD1	ankyrin repeat and KH domain containing 1	PF00013	1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000090861	AARS	alanyl-tRNA synthetase		0
ENSG00000108270	AATF	apoptosis antagonizing transcription factor		1
ENSG00000204574	ABCF1	ATP-binding cassette, sub-family F (GCN20), member 1		1
ENSG00000146109	ABT1	activator of basal transcription 1		1
ENSG00000167315	ACAA2	acetyl-CoA acyltransferase 2		1
ENSG00000100813	ACIN1	apoptotic chromatin condensation inducer 1		1
ENSG00000130402	ACTN4	actinin, alpha 4		1
ENSG00000164113	ADAD1	adenosine deaminase domain containing 1 (testis-specific)	PF00035	0
ENSG00000140955	ADAD2	adenosine deaminase domain containing 2	PF00035	0
ENSG00000160710	ADAR	adenosine deaminase, RNA-specific	PF00035	1
ENSG00000197381	ADARB1	adenosine deaminase, RNA-specific, B1	PF00035	0
ENSG00000185736	ADARB2	adenosine deaminase, RNA-specific, B2	PF00035	1
ENSG00000087274	ADD1	adducin 1 (alpha)		1
ENSG00000156110	ADK	adenosine kinase		1
ENSG00000164252	AGGF1	angiogenic factor with G patch and FHA domains 1	PF01585	0
ENSG00000124942	AHNAK	AHNAK nucleoprotein		1
ENSG00000111732	AICDA	activation-induced cytidine deaminase	PF08210	0
ENSG00000164022	AIMP1	aminoacyl tRNA synthetase complex-interacting multifunctional protein 1		0
ENSG00000121057	AKAP1	A kinase (PRKA) anchor protein 1	PF00567, PF00013	1
ENSG00000105127	AKAP8	A kinase (PRKA) anchor protein 8		1
ENSG00000011243	AKAP8L	A kinase (PRKA) anchor protein 8-like		1
ENSG00000059573	ALDH18A1	aldehyde dehydrogenase 18 family, member A1		1
ENSG00000119711	ALDH6A1	aldehyde dehydrogenase 6 family, member A1		1
ENSG00000091542	ALKBH5	alkB, alkylation repair homolog 5 (E. coli)		1
ENSG00000259224	SLC35G6	solute carrier family 35, member G6		1
ENSG00000131503	ANKHD1	ankyrin repeat and KH domain containing 1	PF00013	1
ENSG00000132466	ANKRD17	ankyrin repeat domain 17	PF00013	1
ENSG00000182718	ANXA2	annexin A2		1
ENSG00000164062	APEH	N-acylaminoacyl-peptide hydrolase		1
ENSG00000166181	API5	apoptosis inhibitor 5		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000111701	APOBEC1	apolipoprotein B mRNA editing enzyme, catalytic polypeptide 1	PF08210	0
ENSG00000124701	APOBEC2	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 2	PF08210	0
ENSG00000128383	APOBEC3A	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3A	PF08210	0
ENSG00000179750	APOBEC3B	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3B	PF08210	0
ENSG00000244509	APOBEC3C	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3C	PF08210	1
ENSG00000243811	APOBEC3D	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3D	PF08210	0
ENSG00000128394	APOBEC3F	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3F	PF08210	0
ENSG00000239713	APOBEC3G	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G	PF08210	0
ENSG00000100298	APOBEC3H	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3H	PF08210	0
ENSG00000173627	APOBEC4	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 4 (putative)	PF08210	0
ENSG00000021776	AQR	aquarius homolog (mouse)		1
ENSG00000095139	ARCN1	archain 1		1
ENSG00000076928	ARHGEF1	Rho guanine nucleotide exchange factor (GEF) 1		1
ENSG00000182196	ARL6IP4	ADP-ribosylation-like factor 6 interacting protein 4		1
ENSG00000138303	ASCC1	activating signal cointegrator 1 complex subunit 1	PF00013	0
ENSG00000112249	ASCC3	activating signal cointegrator 1 complex subunit 3	PF00270	0
ENSG00000130707	ASS1	argininosuccinate synthase 1		1
ENSG00000152234	ATP5A1	ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit 1, cardiac muscle		1
ENSG00000165629	ATP5C1	ATP synthase, H+ transporting, mitochondrial F1 complex, gamma polypeptide 1		1
ENSG00000124788	ATXN1	ataxin 1	PF08517	0
ENSG00000224470	ATXN1L	ataxin 1-like	PF08517	0
ENSG00000204842	ATXN2	ataxin 2		1
ENSG00000168488	ATXN2L	ataxin 2-like		1
ENSG00000198563	DDX39B	DEAD (Asp-Glu-Ala-Asp) box polypeptide 39B		0
ENSG00000204438	GPANK1	G patch domain and ankyrin repeats 1	PF01585	0
ENSG00000107949	BCCIP	BRCA2 and CDKN1A interacting protein		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000186666	BCDIN3D	BCDIN3 domain containing	PF06859	0
ENSG00000029363	BCLAF1	BCL2-associated transcription factor 1		1
ENSG00000122870	BICC1	bicaudal C homolog 1 (Drosophila)	PF00013	1
ENSG00000197299	BLM	Bloom syndrome, RecQ helicase-like	PF00270	0
ENSG00000165733	BMS1	BMS1 homolog, ribosome assembly protein (yeast)		1
ENSG00000152430	BOLL	bol, boule-like (Drosophila)	PF08777, PF00076	0
ENSG00000170727	BOP1	block of proliferation 1		1
ENSG00000113460	BRIX1	BRX1, biogenesis of ribosomes, homolog (S. cerevisiae)		1
ENSG00000130303	BST2	bone marrow stromal cell antigen 2		1
ENSG00000145741	BTF3	basic transcription factor 3		1
ENSG00000137656	BUD13	BUD13 homolog (S. cerevisiae)		1
ENSG00000112578	BYSL	bystin-like		1
ENSG00000082153	BZW1	basic leucine zipper and W2 domains 1		1
ENSG00000175573	C11orf68	chromosome 11 open reading frame 68		1
ENSG00000119705	SLIRP	SRA stem-loop interacting RNA binding protein	PF00076	1
ENSG00000087302	C14orf166	chromosome 14 open reading frame 166		1
ENSG00000196943	NOP9	NOP9 nucleolar protein homolog (yeast)		1
ENSG00000100802	C14orf93	chromosome 14 open reading frame 93		1
ENSG00000188549	C15orf52	chromosome 15 open reading frame 52		1
ENSG00000103550	C16orf88	chromosome 16 open reading frame 88		1
ENSG00000172171	TEFM	transcription elongation factor, mitochondrial		1
ENSG00000074356	C17orf85	chromosome 17 open reading frame 85		1
ENSG00000143633	C1orf131	chromosome 1 open reading frame 131		1
ENSG00000168275	COA6	cytochrome c oxidase assembly factor 6 homolog (S. cerevisiae)		1
ENSG00000143793	C1orf35	chromosome 1 open reading frame 35		1
ENSG00000162642	C1orf52	chromosome 1 open reading frame 52		1
ENSG00000160679	CHTOP	chromatin target of PRMT1		1
ENSG00000100220	C22orf28	chromosome 22 open reading frame 28		1
ENSG00000184220	CMSS1	cms1 ribosomal small subunit homolog (yeast)	PF00270	1
ENSG00000163946	FAM208A	family with sequence similarity 208, member A		1
ENSG00000173769	TOPAZ1	testis and ovary specific PAZ domain containing 1		0
ENSG00000123838	C4BPA	complement component 4 binding protein, alpha		1
ENSG00000084092	NOA1	nitric oxide associated 1		1
ENSG00000146540	C7orf50	chromosome 7 open reading frame 50		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000198917	C9orf114	chromosome 9 open reading frame 114		1
ENSG00000179218	CALR	calreticulin		1
ENSG00000127022	CANX	calnexin		1
ENSG00000135387	CAPRIN1	cell cycle associated protein 1	PF12287	1
ENSG00000108349	CASC3	cancer susceptibility candidate 3		1
ENSG00000153113	CAST	calpastatin		1
ENSG00000060339	CCAR1	cell division cycle and apoptosis regulator 1		1
ENSG00000181378	CCDC108	coiled-coil domain containing 108		1
ENSG00000007080	CCDC124	coiled-coil domain containing 124		1
ENSG00000185298	CCDC137	coiled-coil domain containing 137		1
ENSG00000108588	CCDC47	coiled-coil domain containing 47		1
ENSG00000126653	NSRP1	nuclear speckle splicing regulatory protein 1		1
ENSG00000133773	CCDC59	coiled-coil domain containing 59		1
ENSG00000152133	CCDC75	coiled-coil domain containing 75	PF01585	0
ENSG00000110104	CCDC86	coiled-coil domain containing 86		1
ENSG00000105321	CCDC9	coiled-coil domain containing 9		1
ENSG00000115484	CCT4	chaperonin containing TCP1, subunit 4 (delta)		1
ENSG00000146731	CCT6A	chaperonin containing TCP1, subunit 6A (zeta 1)		1
ENSG00000117877	CD3EAP	CD3e molecule, epsilon associated protein		1
ENSG00000168438	CDC40	cell division cycle 40 homolog (S. cerevisiae)		1
ENSG00000179604	CDC42EP4	CDC42 effector protein (Rho GTPase binding) 4		1
ENSG00000115816	CEBPZ	CCAAT/enhancer binding protein (C/EBP), zeta		1
ENSG00000149187	CELF1	CUGBP, Elav-like family member 1	PF00076	1
ENSG00000048740	CELF2	CUGBP, Elav-like family member 2	PF00076	0
ENSG00000159409	CELF3	CUGBP, Elav-like family member 3	PF00076	0
ENSG00000101489	CELF4	CUGBP, Elav-like family member 4	PF00076	0
ENSG00000161082	CELF5	CUGBP, Elav-like family member 5	PF00076	0
ENSG00000140488	CELF6	CUGBP, Elav-like family member 6	PF00076	0
ENSG00000173575	CHD2	chromodomain helicase DNA binding protein 2		1
ENSG00000170004	CHD3	chromodomain helicase DNA binding protein 3		1
ENSG00000085872	CHERP	calcium homeostasis endoplasmic reticulum protein	PF01805, PF01585	1
ENSG00000099622	CIRBP	cold inducible RNA binding protein	PF00076	1
ENSG00000141076	CIRH1A	cirrhosis, autosomal recessive 1A (cirhin)		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000145354	CISD2	CDGSH iron sulfur domain 2		1
ENSG00000136026	CKAP4	cytoskeleton-associated protein 4		1
ENSG00000179335	CLK3	CDC-like kinase 3		1
ENSG00000074201	CLNS1A	chloride channel, nucleotide-sensitive, 1A		1
ENSG00000169714	CNBP	CCHC-type zinc finger, nucleic acid binding protein	PF00098	1
ENSG00000125107	CNOT1	CCR4-NOT transcription complex, subunit 1		1
ENSG00000080802	CNOT4	CCR4-NOT transcription complex, subunit 4	PF00076	0
ENSG00000198791	CNOT7	CCR4-NOT transcription complex, subunit 7	PF04857	0
ENSG00000155508	CNOT8	CCR4-NOT transcription complex, subunit 8	PF04857	0
ENSG00000187955	COL14A1	collagen, type XIV, alpha 1		1
ENSG00000102879	CORO1A	coronin, actin binding protein, 1A		1
ENSG00000137449	CPEB2	cytoplasmic polyadenylation element binding protein 2	PF00076	1
ENSG00000107864	CPEB3	cytoplasmic polyadenylation element binding protein 3	PF00076	1
ENSG00000113742	CPEB4	cytoplasmic polyadenylation element binding protein 4	PF00076	1
ENSG00000085719	CPNE3	copine III		1
ENSG00000111605	CPSF6	cleavage and polyadenylation specific factor 6, 68kDa	PF00076	1
ENSG00000149532	CPSF7	cleavage and polyadenylation specific factor 7, 59kDa	PF00076	1
ENSG00000099942	CRKL	v-crk sarcoma virus CT10 oncogene homolog (avian)-like		1
ENSG00000060138	CSDA	cold shock domain protein A		1
ENSG00000009307	CSDE1	cold shock domain containing E1, RNA-binding		1
ENSG00000159176	CSRP1	cysteine and glycine-rich protein 1		1
ENSG00000160213	CSTB	cystatin B (stefin B)		1
ENSG00000101138	CSTF1	cleavage stimulation factor, 3' pre-RNA, subunit 1, 50kDa		1
ENSG00000101811	CSTF2	cleavage stimulation factor, 3' pre-RNA, subunit 2, 64kDa	PF00076	1
ENSG00000177613	CSTF2T	cleavage stimulation factor, 3' pre-RNA, subunit 2, 64kDa, tau variant	PF00076	1
ENSG00000044115	CTNNA1	catenin (cadherin-associated protein), alpha 1, 102kDa		1
ENSG00000150316	CWC15	CWC15 spliceosome-associated protein homolog (S. cerevisiae)		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000132676	DAP3	death associated protein 3		1
ENSG00000115866	DARS	aspartyl-tRNA synthetase		1
ENSG00000188120	DAZ1	deleted in azoospermia 1	PF00076	0
ENSG00000205944	DAZ2	deleted in azoospermia 2	PF00076	0
ENSG00000187191	DAZ3	deleted in azoospermia 3	PF00076	0
ENSG00000205916	DAZ4	deleted in azoospermia 4	PF00076	0
ENSG00000071626	DAZAP1	DAZ associated protein 1	PF00076	1
ENSG00000092345	DAZL	deleted in azoospermia-like	PF00076	0
ENSG00000138231	DBR1	debranching enzyme homolog 1 (S. cerevisiae)		1
ENSG00000164934	DCAF13	DDB1 and CUL4 associated factor 13		1
ENSG00000161634	DCD	dermcidin		1
ENSG00000011465	DCN	decorin		1
ENSG00000151065	DCP1B	DCP1 decapping enzyme homolog B (S. cerevisiae)		0
ENSG00000079785	DDX1	DEAD (Asp-Glu-Ala-Asp) box helicase 1	PF00270	1
ENSG00000178105	DDX10	DEAD (Asp-Glu-Ala-Asp) box polypeptide 10	PF00270	1
ENSG00000100201	DDX17	DEAD (Asp-Glu-Ala-Asp) box helicase 17	PF00270	1
ENSG00000088205	DDX18	DEAD (Asp-Glu-Ala-Asp) box polypeptide 18	PF00270	1
ENSG00000168872	DDX19A	DEAD (Asp-Glu-Ala-Asp) box polypeptide 19A	PF00270	0
ENSG00000157349	DDX19B	DEAD (Asp-Glu-Ala-Asp) box polypeptide 19B	PF00270	0
ENSG00000064703	DDX20	DEAD (Asp-Glu-Ala-Asp) box polypeptide 20	PF00270	0
ENSG00000165732	DDX21	DEAD (Asp-Glu-Ala-Asp) box helicase 21		1
ENSG00000174243	DDX23	DEAD (Asp-Glu-Ala-Asp) box polypeptide 23	PF00270	0
ENSG00000089737	DDX24	DEAD (Asp-Glu-Ala-Asp) box polypeptide 24	PF00270	1
ENSG00000109832	DDX25	DEAD (Asp-Glu-Ala-Asp) box helicase 25	PF00270	0
ENSG00000124228	DDX27	DEAD (Asp-Glu-Ala-Asp) box polypeptide 27	PF00270	1
ENSG00000182810	DDX28	DEAD (Asp-Glu-Ala-Asp) box polypeptide 28	PF00270	1
ENSG00000125485	DDX31	DEAD (Asp-Glu-Ala-Asp) box polypeptide 31		1
ENSG00000123136	DDX39A	DEAD (Asp-Glu-Ala-Asp) box polypeptide 39A	PF00270	1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000215301	DDX3X	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked	PF00270	1
ENSG00000067048	DDX3Y	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked	PF00270	0
ENSG00000152670	DDX4	DEAD (Asp-Glu-Ala-Asp) box polypeptide 4	PF00270	0
ENSG00000183258	DDX41	DEAD (Asp-Glu-Ala-Asp) box polypeptide 41	PF00270	1
ENSG00000198231	DDX42	DEAD (Asp-Glu-Ala-Asp) box polypeptide 42	PF00270	0
ENSG00000080007	DDX43	DEAD (Asp-Glu-Ala-Asp) box polypeptide 43	PF07650, PF00270, PF00013	0
ENSG00000145833	DDX46	DEAD (Asp-Glu-Ala-Asp) box polypeptide 46	PF00270	0
ENSG00000213782	DDX47	DEAD (Asp-Glu-Ala-Asp) box polypeptide 47	PF00270	1
ENSG00000105671	DDX49	DEAD (Asp-Glu-Ala-Asp) box polypeptide 49	PF00270	1
ENSG00000108654	DDX5	DEAD (Asp-Glu-Ala-Asp) box helicase 5	PF00270	1
ENSG00000107625	DDX50	DEAD (Asp-Glu-Ala-Asp) box polypeptide 50	PF00270	1
ENSG00000185163	DDX51	DEAD (Asp-Glu-Ala-Asp) box polypeptide 51	PF00270	1
ENSG00000141141	DDX52	DEAD (Asp-Glu-Ala-Asp) box polypeptide 52	PF00270	1
ENSG00000184735	DDX53	DEAD (Asp-Glu-Ala-Asp) box polypeptide 53	PF00270, PF00013	0
ENSG00000123064	DDX54	DEAD (Asp-Glu-Ala-Asp) box polypeptide 54	PF00270	1
ENSG00000111364	DDX55	DEAD (Asp-Glu-Ala-Asp) box polypeptide 55	PF00270	1
ENSG00000136271	DDX56	DEAD (Asp-Glu-Ala-Asp) box helicase 56	PF00270	1
ENSG00000107201	DDX58	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58	PF00270	0
ENSG00000118197	DDX59	DEAD (Asp-Glu-Ala-Asp) box polypeptide 59	PF00270	0
ENSG00000110367	DDX6	DEAD (Asp-Glu-Ala-Asp) box helicase 6	PF00270	1
ENSG00000137628	DDX60	DEAD (Asp-Glu-Ala-Asp) box polypeptide 60	PF00270	0
ENSG00000181381	DDX60L	DEAD (Asp-Glu-Ala-Asp) box polypeptide 60-like	PF00270	0
ENSG00000124795	DEK	DEK oncogene		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000128191	DGCR8	DiGeorge syndrome critical region gene 8	PF00035	0
ENSG00000109606	DHX15	DEAH (Asp-Glu-Ala-His) box polypeptide 15	PF00270	1
ENSG00000204560	DHX16	DEAH (Asp-Glu-Ala-His) box polypeptide 16	PF00270	1
ENSG00000067248	DHX29	DEAH (Asp-Glu-Ala-His) box polypeptide 29	PF00270	0
ENSG00000132153	DHX30	DEAH (Asp-Glu-Ala-His) box polypeptide 30	PF00270	1
ENSG00000005100	DHX33	DEAH (Asp-Glu-Ala-His) box polypeptide 33	PF00270	1
ENSG00000174953	DHX36	DEAH (Asp-Glu-Ala-His) box polypeptide 36	PF00270	1
ENSG00000163214	DHX57	DEAH (Asp-Glu-Ala-Asp/His) box polypeptide 57	PF00270	1
ENSG00000108771	DHX58	DEXH (Asp-Glu-X-His) box polypeptide 58	PF00270	0
ENSG00000067596	DHX8	DEAH (Asp-Glu-Ala-His) box polypeptide 8	PF00270, PF00575	1
ENSG00000135829	DHX9	DEAH (Asp-Glu-Ala-His) box polypeptide 9	PF00270, PF00035	1
ENSG00000131504	DIAPH1	diaphanous homolog 1 (Drosophila)		1
ENSG00000100697	DICER1	dicer 1, ribonuclease type III	PF00270, PF02170, PF03368	0
ENSG00000117597	DIEXF	digestive organ expansion factor homolog (zebrafish)		1
ENSG00000086189	DIMT1	DIM1 dimethyladenosine transferase 1 homolog (S. cerevisiae)		1
ENSG00000130826	DKC1	dyskeratosis congenita 1, dyskerin		1
ENSG00000132837	DMGDH	dimethylglycine dehydrogenase		1
ENSG00000104129	DNAJC17	DnaJ (Hsp40) homolog, subfamily C, member 17	PF00076	0
ENSG00000105821	DNAJC2	DnaJ (Hsp40) homolog, subfamily C, member 2		1
ENSG00000168724	DNAJC21	DnaJ (Hsp40) homolog, subfamily C, member 21	PF12171	1
ENSG00000256453	DND1	dead end homolog 1 (zebrafish)	PF00076	0
ENSG00000067334	DNTTIP2	deoxynucleotidyltransferase, terminal, interacting protein 2		1
ENSG00000203909	DPPA5	developmental pluripotency associated 5		0
ENSG00000113360	DROSHA	drosha, ribonuclease type III	PF00035	1
ENSG00000096696	DSP	desmoplakin		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000167264	DUS2L	dihydrouridine synthase 2-like, SMM1 homolog (S. cerevisiae)		0
ENSG00000144048	DUSP11	dual specificity phosphatase 11 (RNA/RNP complex 1-interacting)		1
ENSG00000128951	DUT	deoxyuridine triphosphatase		1
ENSG00000197102	DYNC1H1	dynein, cytoplasmic 1, heavy chain 1		1
ENSG00000144635	DYNC1LI1	dynein, cytoplasmic 1, light intermediate chain 1		1
ENSG00000198919	DZIP3	DAZ interacting protein 3, zinc finger		1
ENSG00000117395	EBNA1BP2	EBNA1 binding protein 2		1
ENSG00000107223	EDF1	endothelial differentiation-related factor 1		1
ENSG00000156508	EEF1A1	eukaryotic translation elongation factor 1 alpha 1		1
ENSG00000167658	EEF2	eukaryotic translation elongation factor 2		1
ENSG00000108883	EFTUD2	elongation factor Tu GTP binding domain containing 2		1
ENSG00000173674	EIF1AX	eukaryotic translation initiation factor 1A, X-linked		1
ENSG00000055332	EIF2AK2	eukaryotic translation initiation factor 2-alpha kinase 2	PF00035	0
ENSG00000092847	EIF2C1	eukaryotic translation initiation factor 2C, 1	PF02171, PF02170	0
ENSG00000123908	EIF2C2	eukaryotic translation initiation factor 2C, 2	PF02171, PF02170	1
ENSG00000126070	EIF2C3	eukaryotic translation initiation factor 2C, 3	PF02171, PF02170	1
ENSG00000134698	EIF2C4	eukaryotic translation initiation factor 2C, 4	PF02171, PF02170	0
ENSG00000134001	EIF2S1	eukaryotic translation initiation factor 2, subunit 1 alpha, 35kDa	PF00575	1
ENSG00000125977	EIF2S2	eukaryotic translation initiation factor 2, subunit 2 beta, 38kDa		1
ENSG00000107581	EIF3A	eukaryotic translation initiation factor 3, subunit A		1
ENSG00000106263	EIF3B	eukaryotic translation initiation factor 3, subunit B	PF00076	0
ENSG00000184110	EIF3C	eukaryotic translation initiation factor 3, subunit C		1
ENSG00000100353	EIF3D	eukaryotic translation initiation factor 3, subunit D		1
ENSG00000130811	EIF3G	eukaryotic translation initiation factor 3, subunit G	PF00076	1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000147677	EIF3H	eukaryotic translation initiation factor 3, subunit H		1
ENSG00000100129	EIF3L	eukaryotic translation initiation factor 3, subunit L		1
ENSG00000161960	EIF4A1	eukaryotic translation initiation factor 4A1	PF00270	1
ENSG00000156976	EIF4A2	eukaryotic translation initiation factor 4A2	PF00270	1
ENSG00000141543	EIF4A3	eukaryotic translation initiation factor 4A3	PF00270	1
ENSG00000063046	EIF4B	eukaryotic translation initiation factor 4B	PF00076	1
ENSG00000114867	EIF4G1	eukaryotic translation initiation factor 4 gamma, 1		1
ENSG00000110321	EIF4G2	eukaryotic translation initiation factor 4 gamma, 2		1
ENSG00000075151	EIF4G3	eukaryotic translation initiation factor 4 gamma, 3		1
ENSG00000106682	EIF4H	eukaryotic translation initiation factor 4H	PF00076	1
ENSG00000158417	EIF5B	eukaryotic translation initiation factor 5B		1
ENSG00000006744	ELAC2	elaC homolog 2 (E. coli)		1
ENSG00000066044	ELAVL1	ELAV (embryonic lethal, abnormal vision, Drosophila)-like 1 (Hu antigen R)	PF00076	1
ENSG00000107105	ELAVL2	ELAV (embryonic lethal, abnormal vision, Drosophila)-like 2 (Hu antigen B)	PF00076	1
ENSG00000196361	ELAVL3	ELAV (embryonic lethal, abnormal vision, Drosophila)-like 3 (Hu antigen C)	PF00076	0
ENSG00000162374	ELAVL4	ELAV (embryonic lethal, abnormal vision, Drosophila)-like 4	PF00076	0
ENSG00000126749	EMG1	EMG1 nucleolar protein homolog (S. cerevisiae)		1
ENSG00000074800	ENO1	enolase 1, (alpha)		1
ENSG00000120658	ENOX1	ecto-NOX disulfide-thiol exchanger 1	PF00076	0
ENSG00000165675	ENOX2	ecto-NOX disulfide-thiol exchanger 2	PF00076	0
ENSG00000132591	ERAL1	Era G-protein-like 1 (E. coli)		1
ENSG00000163161	ERCC3	excision repair cross-complementing rodent repair deficiency, complementation group 3		0
ENSG00000117419	ERI3	ERI1 exoribonuclease family member 3		1
ENSG00000089048	ESF1	ESF1, nucleolar pre-rRNA processing protein, homolog (S. cerevisiae)		1
ENSG00000104413	ESRP1	epithelial splicing regulatory protein 1	PF00076	0
ENSG00000103067	ESRP2	epithelial splicing regulatory protein 2		1
ENSG00000120705	ETF1	eukaryotic translation termination factor 1		1
ENSG00000182944	EWSR1	Ewing sarcoma breakpoint region 1	PF00076	1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000171824	EXOSC10	exosome component 10		1
ENSG00000123737	EXOSC9	exosome component 9		1
ENSG00000092820	EZR	ezrin		1
ENSG00000048828	FAM120A	family with sequence similarity 120A		1
ENSG00000184083	FAM120C	family with sequence similarity 120C		1
ENSG00000105058	FAM32A	family with sequence similarity 32, member A		1
ENSG00000112773	FAM46A	family with sequence similarity 46, member A		1
ENSG00000119812	FAM98A	family with sequence similarity 98, member A		1
ENSG00000187790	FANCM	Fanconi anemia, complementation group M	PF00270	0
ENSG00000169710	FASN	fatty acid synthase		1
ENSG00000164896	FASTK	Fas-activated serine/threonine kinase	PF08373	0
ENSG00000138399	FASTKD1	FAST kinase domains 1	PF08373	1
ENSG00000118246	FASTKD2	FAST kinase domains 2	PF08373	1
ENSG00000124279	FASTKD3	FAST kinase domains 3	PF08373	1
ENSG00000215251	FASTKD5	FAST kinase domains 5	PF08373	0
ENSG00000105202	FBL	fibrillarin		1
ENSG00000119616	FCF1	FCF1 small subunit (SSU) processome component homolog (S. cerevisiae)		1
ENSG00000160752	FDPS	farnesyl diphosphate synthase		1
ENSG00000145216	FIP1L1	FIP1 like 1 (S. cerevisiae)		1
ENSG00000100442	FKBP3	FK506 binding protein 3, 25kDa		1
ENSG00000004478	FKBP4	FK506 binding protein 4, 59kDa		1
ENSG00000196924	FLNA	filamin A, alpha		1
ENSG00000162076	FLYWCH2	FLYWCH family member 2		1
ENSG00000102081	FMR1	fragile X mental retardation 1		1
ENSG00000102531	FNDC3A	fibronectin type III domain containing 3A		1
ENSG00000075420	FNDC3B	fibronectin type III domain containing 3B		1
ENSG00000109536	FRG1	FSHD region gene 1		1
ENSG00000075618	FSCN1	fascin homolog 1, actin-bundling protein (Strongylocentrotus purpuratus)		1
ENSG00000108592	FTSJ3	FtsJ homolog 3 (E. coli)		1
ENSG00000137200	FTSJD2	FtsJ methyltransferase domain containing 2	PF01585	0
ENSG00000162613	FUBP1	far upstream element (FUSE) binding protein 1	PF07650, PF00013	1
ENSG00000107164	FUBP3	far upstream element (FUSE) binding protein 3	PF07650, PF00013	1
ENSG00000089280	FUS	fused in sarcoma	PF00076	1
ENSG00000114416	FXR1	fragile X mental retardation, autosomal homolog 1	PF12235, PF00013	1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000129245	FXR2	fragile X mental retardation, autosomal homolog 2	PF12235, PF00013	1
ENSG00000145907	G3BP1	GTPase activating protein (SH3 domain) binding protein 1	PF00076	1
ENSG00000138757	G3BP2	GTPase activating protein (SH3 domain) binding protein 2	PF00076	1
ENSG00000089597	GANAB	glucosidase, alpha; neutral AB		1
ENSG00000109534	GAR1	GAR1 ribonucleoprotein homolog (yeast)	PF04410	1
ENSG00000082516	GEMIN5	gem (nuclear organelle) associated protein 5		1
ENSG00000168827	GFM1	G elongation factor, mitochondrial 1		1
ENSG00000108010	GLRX3	glutaredoxin 3		1
ENSG00000105373	GLTSCR2	glioma tumor suppressor candidate region gene 2		1
ENSG00000204628	GNB2L1	guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1		1
ENSG00000134697	GNL2	guanine nucleotide binding protein-like 2 (nucleolar)		1
ENSG00000163938	GNL3	guanine nucleotide binding protein-like 3 (nucleolar)		1
ENSG00000130119	GNL3L	guanine nucleotide binding protein-like 3 (nucleolar)-like		1
ENSG00000173230	GOLGB1	golgin B1		1
ENSG00000092978	GPATCH2	G patch domain containing 2	PF01585	0
ENSG00000198746	GPATCH3	G patch domain containing 3	PF01585	0
ENSG00000160818	GPATCH4	G patch domain containing 4	PF01585	1
ENSG00000186566	GPATCH8	G patch domain containing 8	PF12171, PF01585	0
ENSG00000068394	GPKOW	G patch domain and KOW motifs	PF01585	0
ENSG00000177885	GRB2	growth factor receptor-bound protein 2		1
ENSG00000030582	GRN	granulin		1
ENSG00000132463	GRSF1	G-rich RNA sequence binding factor 1	PF00076	1
ENSG00000105447	GRWD1	glutamate-rich WD repeat containing 1		1
ENSG00000103342	GSPT1	G1 to S phase transition 1		1
ENSG00000189369	GSPT2	G1 to S phase transition 2		1
ENSG00000197265	GTF2E2	general transcription factor IIE, polypeptide 2, beta 34kDa		1
ENSG00000125651	GTF2F1	general transcription factor IIF, polypeptide 1, 74kDa		1
ENSG00000100226	GTPBP1	GTP binding protein 1		1
ENSG00000105793	GTPBP10	GTP-binding protein 10 (putative)		1
ENSG00000107937	GTPBP4	GTP binding protein 4		1
ENSG00000170627	GTSF1	gametocyte specific factor 1	PF05253	0
ENSG00000124196	GTSF1L	gametocyte specific factor 1-like	PF05253	0

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000189060	H1F0	H1 histone family, member 0		1
ENSG00000130956	HABP4	hyaluronan binding protein 4	PF04774	0
ENSG00000138029	HADHB	hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase/enoyl-CoA hydratase (trifunctional protein), beta subunit		1
ENSG00000105856	HBP1	HMG-box transcription factor 1	PF08517	0
ENSG00000143321	HDGF	hepatoma-derived growth factor		1
ENSG00000115677	HDLBP	high density lipoprotein binding protein	PF00013	1
ENSG00000119285	HEATR1	HEAT repeat containing 1		1
ENSG00000163312	HELQ	helicase, POLQ-like	PF00270	0
ENSG00000198265	HELZ	helicase with zinc finger		1
ENSG00000138646	HERC5	HECT and RLD domain containing E3 ubiquitin protein ligase 5		1
ENSG00000162669	HFM1	HFM1, ATP-dependent DNA helicase homolog (S. cerevisiae)	PF00270	0
ENSG00000184357	HIST1H1B	histone cluster 1, H1b		1
ENSG00000187837	HIST1H1C	histone cluster 1, H1c		1
ENSG00000158406	HIST1H4H	histone cluster 1, H4h		1
ENSG00000071794	HLTF	helicase-like transcription factor		1
ENSG00000189403	HMGB1	high mobility group box 1		1
ENSG00000164104	HMGB2	high mobility group box 2		1
ENSG00000177733	HNRNPA0	heterogeneous nuclear ribonucleoprotein A0	PF00076	1
ENSG00000135486	HNRNPA1	heterogeneous nuclear ribonucleoprotein A1	PF00076	1
ENSG00000139675	HNRNPA1L2	heterogeneous nuclear ribonucleoprotein A1-like 2	PF00076	0
ENSG00000122566	HNRNPA2B1	heterogeneous nuclear ribonucleoprotein A2/B1	PF00076	1
ENSG00000170144	HNRNPA3	heterogeneous nuclear ribonucleoprotein A3	PF00076	1
ENSG00000197451	HNRNPAB	heterogeneous nuclear ribonucleoprotein A/B	PF00076	1
ENSG00000092199	HNRNPC	heterogeneous nuclear ribonucleoprotein C (C1/C2)	PF00076	1
ENSG00000179172	HNRNPCL1	heterogeneous nuclear ribonucleoprotein C-like 1	PF00076	0
ENSG00000138668	HNRNPD	heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA binding protein 1, 37kDa)	PF00076	1
ENSG00000169813	HNRNPF	heterogeneous nuclear ribonucleoprotein F	PF00076	1
ENSG00000169045	HNRNPH1	heterogeneous nuclear ribonucleoprotein H1 (H)	PF00076	1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000126945	HNRNPH2	heterogeneous nuclear ribonucleoprotein H2 (H')	PF00076	1
ENSG00000096746	HNRNPH3	heterogeneous nuclear ribonucleoprotein H3 (2H9)	PF00076	1
ENSG00000165119	HNRNPK	heterogeneous nuclear ribonucleoprotein K	PF00013	1
ENSG00000104824	HNRNPL	heterogeneous nuclear ribonucleoprotein L	PF00076	1
ENSG00000099783	HNRNPM	heterogeneous nuclear ribonucleoprotein M	PF00076	1
ENSG00000125944	HNRNPR	heterogeneous nuclear ribonucleoprotein R	PF00076	1
ENSG00000153187	HNRNPU	heterogeneous nuclear ribonucleoprotein U (scaffold attachment factor A)		1
ENSG00000105323	HNRNPUL1	heterogeneous nuclear ribonucleoprotein U-like 1		1
ENSG00000214753	HNRNPUL2	heterogeneous nuclear ribonucleoprotein U-like 2		1
ENSG00000152795	HNRPDL	heterogeneous nuclear ribonucleoprotein D-like	PF00076	1
ENSG00000143889	HNRPLL	heterogeneous nuclear ribonucleoprotein L-like	PF00076	1
ENSG00000132541	HRSP12	heat-responsive protein 12		1
ENSG00000080824	HSP90AA1	heat shock protein 90kDa alpha (cytosolic), class A member 1		1
ENSG00000096384	HSP90AB1	heat shock protein 90kDa alpha (cytosolic), class B member 1		1
ENSG00000204388	HSPA1B	heat shock 70kDa protein 1B		1
ENSG00000109971	HSPA8	heat shock 70kDa protein 8		1
ENSG00000113013	HSPA9	heat shock 70kDa protein 9 (mortalin)		1
ENSG00000106211	HSPB1	heat shock 27kDa protein 1		1
ENSG00000144381	HSPD1	heat shock 60kDa protein 1 (chaperonin)		1
ENSG00000115541	HSPE1	heat shock 10kDa protein 1 (chaperonin 10)		1
ENSG00000102241	HTATSF1	HIV-1 Tat specific factor 1	PF00076	1
ENSG00000086758	HUWE1	HECT, UBA and WWE domain containing 1, E3 ubiquitin protein ligase		1
ENSG00000163565	IFI16	interferon, gamma-inducible protein 16		1
ENSG00000115267	IFIH1	interferon induced with helicase C domain 1	PF00270	0
ENSG00000119922	IFIT2	interferon-induced protein with tetratricopeptide repeats 2		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000159217	IGF2BP1	insulin-like growth factor 2 mRNA binding protein 1	PF00013, PF00076	1
ENSG00000073792	IGF2BP2	insulin-like growth factor 2 mRNA binding protein 2	PF00013, PF00076	1
ENSG00000136231	IGF2BP3	insulin-like growth factor 2 mRNA binding protein 3	PF00013, PF00076	1
ENSG00000143621	ILF2	interleukin enhancer binding factor 2, 45kDa		1
ENSG00000129351	ILF3	interleukin enhancer binding factor 3, 90kDa	PF00035	1
ENSG00000132305	IMMT	inner membrane protein, mitochondrial		1
ENSG00000177971	IMP3	IMP3, U3 small nucleolar ribonucleoprotein, homolog (yeast)		1
ENSG00000106348	IMPDH1	IMP (inosine 5'-monophosphate) dehydrogenase 1		0
ENSG00000143319	ISG20L2	interferon stimulated exonuclease gene 20kDa-like 2		1
ENSG00000240682	ISY1	ISY1 splicing factor homolog (S. cerevisiae)		1
ENSG00000121774	KHDRBS1	KH domain containing, RNA binding, signal transduction associated 1	PF00013	1
ENSG00000112232	KHDRBS2	KH domain containing, RNA binding, signal transduction associated 2	PF00013	0
ENSG00000131773	KHDRBS3	KH domain containing, RNA binding, signal transduction associated 3		0
ENSG00000088247	KHSRP	KH-type splicing regulatory protein	PF07650, PF00013	1
ENSG00000080608	KIAA0020	KIAA0020		1
ENSG00000168502	SOGA2	SOGA family member 2		1
ENSG00000116299	KIAA1324	KIAA1324		1
ENSG00000158941	KIAA1967	KIAA1967		1
ENSG00000129250	KIF1C	kinesin family member 1C		1
ENSG00000129347	KRI1	KRI1 homolog (S. cerevisiae)		1
ENSG00000111615	KRR1	KRR1, small subunit (SSU) processome component, homolog (yeast)	PF00013	1
ENSG00000111057	KRT18	keratin 18		1
ENSG00000126777	KTN1	kinectin 1 (kinesin receptor)		1
ENSG00000155506	LARP1	La ribonucleoprotein domain family, member 1	PF05383	1
ENSG00000138709	LARP1B	La ribonucleoprotein domain family, member 1B	PF05383	0
ENSG00000161813	LARP4	La ribonucleoprotein domain family, member 4	PF05383	1
ENSG00000107929	LARP4B	La ribonucleoprotein domain family, member 4B	PF05383	1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000166173	LARP6	La ribonucleoprotein domain family, member 6	PF05383	0
ENSG00000174720	LARP7	La ribonucleoprotein domain family, member 7	PF08777, PF00076, PF05383	1
ENSG00000001497	LAS1L	LAS1-like (S. cerevisiae)		1
ENSG00000100097	LGALS1	lectin, galactoside-binding, soluble, 1		1
ENSG00000131981	LGALS3	lectin, galactoside-binding, soluble, 3		1
ENSG00000131914	LIN28A	lin-28 homolog A (C. elegans)	PF00098	0
ENSG00000187772	LIN28B	lin-28 homolog B (C. elegans)		0
ENSG00000139233	LLPH	LLP homolog, long-term synaptic facilitation (Aplysia)		1
ENSG00000123384	LRP1	low density lipoprotein receptor-related protein 1		1
ENSG00000138095	LRPPRC	leucine-rich pentatricopeptide repeat containing		1
ENSG00000108829	LRRC59	leucine rich repeat containing 59		1
ENSG00000175324	LSM1	LSM1 homolog, U6 small nuclear RNA associated (S. cerevisiae)	PF01423	1
ENSG00000181817	LSM10	LSM10, U7 small nuclear RNA associated	PF01423	0
ENSG00000155858	LSM11	LSM11, U7 small nuclear RNA associated	PF01423	0
ENSG00000257103	LSM14A	LSM14A, SCD6 homolog A (S. cerevisiae)		1
ENSG00000204392	LSM2	LSM2 homolog, U6 small nuclear RNA associated (S. cerevisiae)	PF01423	1
ENSG00000170860	LSM3	LSM3 homolog, U6 small nuclear RNA associated (S. cerevisiae)	PF01423	1
ENSG00000130520	LSM4	LSM4 homolog, U6 small nuclear RNA associated (S. cerevisiae)	PF01423	1
ENSG00000106355	LSM5	LSM5 homolog, U6 small nuclear RNA associated (S. cerevisiae)	PF01423	0
ENSG00000164167	LSM6	LSM6 homolog, U6 small nuclear RNA associated (S. cerevisiae)	PF01423	0
ENSG00000130332	LSM7	LSM7 homolog, U6 small nuclear RNA associated (S. cerevisiae)	PF01423	0
ENSG00000183011	LSMD1	LSM domain containing 1	PF01423	0
ENSG00000111144	LTA4H	leukotriene A4 hydrolase		1
ENSG00000146963	LUC7L2	LUC7-like 2 (S. cerevisiae)		1
ENSG00000108848	LUC7L3	LUC7-like 3 (S. cerevisiae)		1
ENSG00000145220	LYAR	Ly1 antibody reactive homolog (mouse)		1
ENSG00000162385	MAGOH	mago-nashi homolog, proliferation-associated (Drosophila)		0

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000264176	MAGOH3P	mago-nashi homolog 2, proliferation-associated (Drosophila), pseudogene		0
ENSG00000111196	MAGOHB	mago-nashi homolog B (Drosophila)	PF02792	0
ENSG00000198042	MAK16	MAK16 homolog (S. cerevisiae)		1
ENSG00000047849	MAP4	microtubule-associated protein 4		1
ENSG00000072518	MARK2	MAP/microtubule affinity-regulating kinase 2		1
ENSG00000015479	MATR3	matrin 3		1
ENSG00000103495	MAZ	MYC-associated zinc finger protein (purine-binding transcription factor)		1
ENSG00000152601	MBNL1	muscleblind-like splicing regulator 1		1
ENSG00000139793	MBNL2	muscleblind-like splicing regulator 2		1
ENSG00000076770	MBNL3	muscleblind-like splicing regulator 3		1
ENSG00000146701	MDH2	malate dehydrogenase 2, NAD (mitochondrial)		1
ENSG00000169057	MECP2	methyl CpG binding protein 2 (Rett syndrome)		1
ENSG00000146834	MEPCE	methylphosphate capping enzyme	PF06859	1
ENSG00000111142	METAP2	methionyl aminopeptidase 2		1
ENSG00000127804	METTL16	methyltransferase like 16		1
ENSG00000254726	MEX3A	mex-3 homolog A (C. elegans)	PF00013	0
ENSG00000183496	MEX3B	mex-3 homolog B (C. elegans)	PF00013	0
ENSG00000176624	MEX3C	mex-3 homolog C (C. elegans)	PF00013	1
ENSG00000181588	MEX3D	mex-3 homolog D (C. elegans)	PF00013	1
ENSG00000140259	MFAP1	microfibrillar-associated protein 1		1
ENSG00000148773	MKI67	antigen identified by monoclonal antibody Ki-67		1
ENSG00000155438	MKI67IP	MKI67 (FHA domain) interacting nucleolar phosphoprotein	PF00076	1
ENSG00000133606	MKRN1	makorin ring finger protein 1		1
ENSG00000055609	MLL3	myeloid/lymphoid or mixed-lineage leukemia 3		1
ENSG00000155363	MOV10	Mov10, Moloney leukemia virus 10, homolog (mouse)		1
ENSG00000124383	MPHOSPH10	M-phase phosphoprotein 10 (U3 small nucleolar ribonucleoprotein)		1
ENSG00000129282	MRM1	mitochondrial rRNA methyltransferase 1 homolog (S. cerevisiae)		1
ENSG00000169288	MRPL1	mitochondrial ribosomal protein L1		1
ENSG00000174547	MRPL11	mitochondrial ribosomal protein L11		1
ENSG00000172172	MRPL13	mitochondrial ribosomal protein L13		1
ENSG00000137547	MRPL15	mitochondrial ribosomal protein L15		1
ENSG00000116221	MRPL37	mitochondrial ribosomal protein L37		1
ENSG00000182154	MRPL41	mitochondrial ribosomal protein L41		1
ENSG00000086504	MRPL28	mitochondrial ribosomal protein L28		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000114686	MRPL3	mitochondrial ribosomal protein L3		1
ENSG00000106591	MRPL32	mitochondrial ribosomal protein L32		1
ENSG00000154719	MRPL39	mitochondrial ribosomal protein L39		1
ENSG00000105364	MRPL4	mitochondrial ribosomal protein L4		1
ENSG00000198015	MRPL42	mitochondrial ribosomal protein L42		1
ENSG00000055950	MRPL43	mitochondrial ribosomal protein L43		1
ENSG00000174100	MRPL45	mitochondrial ribosomal protein L45		1
ENSG00000181991	MRPS11	mitochondrial ribosomal protein S11		1
ENSG00000116898	MRPS15	mitochondrial ribosomal protein S15		1
ENSG00000181610	MRPS23	mitochondrial ribosomal protein S23		1
ENSG00000062582	MRPS24	mitochondrial ribosomal protein S24		1
ENSG00000147586	MRPS28	mitochondrial ribosomal protein S28		1
ENSG00000112996	MRPS30	mitochondrial ribosomal protein S30		1
ENSG00000102738	MRPS31	mitochondrial ribosomal protein S31		1
ENSG00000144029	MRPS5	mitochondrial ribosomal protein S5		1
ENSG00000125445	MRPS7	mitochondrial ribosomal protein S7		1
ENSG00000053372	MRTO4	mRNA turnover 4 homolog (S. cerevisiae)		1
ENSG00000135097	MSI1	musashi homolog 1 (Drosophila)	PF00076	0
ENSG00000153944	MSI2	musashi homolog 2 (Drosophila)	PF00076	1
ENSG00000147649	MTDH	metadherin		1
ENSG00000103248	MTHFSD	methenyltetrahydrofolate synthetase domain containing	PF00076	1
ENSG00000085760	MTIF2	mitochondrial translational initiation factor 2		1
ENSG00000107951	MTPAP	mitochondrial poly(A) polymerase		1
ENSG00000132382	MYBBP1A	MYB binding protein (P160) 1a		1
ENSG00000104177	MYEF2	myelin expression factor 2	PF08777, PF00076	0
ENSG00000047597	NA	neuroacanthocytosis		1
ENSG00000164134	NAA15	N(alpha)-acetyltransferase 15, NatA auxiliary subunit		1
ENSG00000128534	NAA38	N(alpha)-acetyltransferase 38, NatC auxiliary subunit	PF01423	1
ENSG00000188613	NANOS1	nanos homolog 1 (Drosophila)	PF05741	0
ENSG00000188425	NANOS2	nanos homolog 2 (Drosophila)	PF05741	0
ENSG00000187556	NANOS3	nanos homolog 3 (Drosophila)	PF05741	0
ENSG00000205531	NAP1L4	nucleosome assembly protein 1-like 4		1
ENSG00000135372	NAT10	N-acetyltransferase 10 (GCN5-related)		1
ENSG00000114503	NCBP2	nuclear cap binding protein subunit 2, 20kDa	PF00076	1
ENSG00000170935	NCBP2L	nuclear cap binding protein subunit 2-like		0
ENSG00000115053	NCL	nucleolin		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000160194	NDUFV3	NADH dehydrogenase (ubiquinone) flavoprotein 3, 10kDa		1
ENSG00000086102	NFX1	nuclear transcription factor, X-box binding 1		1
ENSG00000129460	NGDN	neuroguidin, EIF4E binding protein		1
ENSG00000182768	NGRN	neugrin, neurite outgrowth associated		1
ENSG00000145912	NHP2	NHP2 ribonucleoprotein homolog (yeast)		1
ENSG00000100138	NHP2L1	NHP2 non-histone chromosome protein 2-like 1 (S. cerevisiae)	PF08228	1
ENSG00000132603	NIP7	nuclear import 7 homolog (S. cerevisiae)		1
ENSG00000186416	NKRF	NFKB repressing factor	PF01585	1
ENSG00000179873	NLRP11	NLR family, pyrin domain containing 11		1
ENSG00000169251	NMD3	NMD3 homolog (S. cerevisiae)		1
ENSG00000239672	NME1	NME/NM23 nucleoside diphosphate kinase 1		1
ENSG00000141101	NOB1	NIN1/RPN12 binding protein 1 homolog (S. cerevisiae)	PF08772	0
ENSG00000188976	NOC2L	nucleolar complex associated 2 homolog (S. cerevisiae)		1
ENSG00000173145	NOC3L	nucleolar complex associated 3 homolog (S. cerevisiae)		1
ENSG00000184967	NOC4L	nucleolar complex associated 4 homolog (S. cerevisiae)		1
ENSG00000115761	NOL10	nucleolar protein 10		1
ENSG00000130935	NOL11	nucleolar protein 11		1
ENSG00000256872	NOL12	nucleolar protein 12		1
ENSG00000165271	NOL6	nucleolar protein family 6 (RNA-associated)		1
ENSG00000225921	NOL7	nucleolar protein 7, 27kDa		1
ENSG00000198000	NOL8	nucleolar protein 8	PF00076	1
ENSG00000166197	NOLC1	nucleolar and coiled-body phosphoprotein 1		1
ENSG00000147140	NONO	non-POU domain containing, octamer-binding	PF00076	1
ENSG00000182117	NOP10	NOP10 ribonucleoprotein homolog (yeast)	PF04135	0
ENSG00000087269	NOP14	NOP14 nucleolar protein homolog (yeast)		1
ENSG00000048162	NOP16	NOP16 nucleolar protein homolog (yeast)		1
ENSG00000111641	NOP2	NOP2 nucleolar protein homolog (yeast)		1
ENSG00000101361	NOP56	NOP56 ribonucleoprotein homolog (yeast)	PF08156, PF01798	1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000055044	NOP58	NOP58 ribonucleoprotein homolog (yeast)	PF08156, PF01798	1
ENSG00000142546	NOSIP	nitric oxide synthase interacting protein		1
ENSG00000139910	NOVA1	neuro-oncological ventral antigen 1	PF00013	0
ENSG00000104967	NOVA2	neuro-oncological ventral antigen 2	PF00013	0
ENSG00000181163	NPM1	nucleophosmin (nucleolar phosphoprotein B23, numatrin)		1
ENSG00000107833	NPM3	nucleophosmin/nucleoplasmin 3		1
ENSG00000181019	NQO1	NAD(P)H dehydrogenase, quinone 1		1
ENSG00000164346	NSA2	NSA2 ribosome biogenesis homolog (S. cerevisiae)		1
ENSG00000037474	NSUN2	NOP2/Sun RNA methyltransferase family, member 2		1
ENSG00000130305	NSUN5	NOP2/Sun domain family, member 5		1
ENSG00000167005	NUDT21	nudix (nucleoside diphosphate linked moiety X)-type motif 21		1
ENSG00000108256	NUFIP2	nuclear fragile X mental retardation protein interacting protein 2		1
ENSG00000163002	NUP35	nucleoporin 35kDa	PF05172	0
ENSG00000137804	NUSAP1	nucleolar and spindle associated protein 1		1
ENSG00000143748	NVL	nuclear VCP-like		1
ENSG00000162231	NXF1	nuclear RNA export factor 1	PF09162	1
ENSG00000185554	NXF2	nuclear RNA export factor 2	PF09162	0
ENSG00000185945	NXF2B	nuclear RNA export factor 2B	PF09162	0
ENSG00000147206	NXF3	nuclear RNA export factor 3	PF09162	0
ENSG00000126952	NXF5	nuclear RNA export factor 5	PF09162	0
ENSG00000135114	OASL	2'-5'-oligoadenylate synthetase-like		1
ENSG00000185624	P4HB	prolyl 4-hydroxylase, beta polypeptide		1
ENSG00000170515	PA2G4	proliferation-associated 2G4, 38kDa		1
ENSG00000070756	PABPC1	poly(A) binding protein, cytoplasmic 1	PF00076, PF00658	1
ENSG00000101104	PABPC1L	poly(A) binding protein, cytoplasmic 1-like	PF00076, PF00658	0
ENSG00000186288	PABPC1L2A	poly(A) binding protein, cytoplasmic 1-like 2A	PF00076	0
ENSG00000184388	PABPC1L2B	poly(A) binding protein, cytoplasmic 1-like 2B	PF00076	0
ENSG00000151846	PABPC3	poly(A) binding protein, cytoplasmic 3	PF00076, PF00658	0
ENSG00000090621	PABPC4	poly(A) binding protein, cytoplasmic 4 (inducible form)	PF00076, PF00658	1
ENSG00000174740	PABPC5	poly(A) binding protein, cytoplasmic 5	PF00076	0
ENSG00000100836	PABPN1	poly(A) binding protein, nuclear 1	PF00076	1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000205022	PABPN1L	poly(A) binding protein, nuclear 1-like (cytoplasmic)	PF00076	0
ENSG00000121274	PAPD5	PAP associated domain containing 5		1
ENSG00000090060	PAPOLA	poly(A) polymerase alpha	PF04926	0
ENSG00000218823	PAPOLB	poly(A) polymerase beta (testis specific)	PF04926	0
ENSG00000115421	PAPOLG	poly(A) polymerase gamma	PF04926	0
ENSG00000140694	PARN	poly(A)-specific ribonuclease	PF08675, PF04857	1
ENSG00000143799	PARP1	poly (ADP-ribose) polymerase 1		1
ENSG00000059378	PARP12	poly (ADP-ribose) polymerase family, member 12		1
ENSG00000166889	PATL1	protein associated with topoisomerase II homolog 1 (yeast)		1
ENSG00000169564	PCBP1	poly(rC) binding protein 1	PF07650, PF00013	1
ENSG00000197111	PCBP2	poly(rC) binding protein 2	PF07650, PF00013	1
ENSG00000183570	PCBP3	poly(rC) binding protein 3	PF07650, PF00013	1
ENSG00000090097	PCBP4	poly(rC) binding protein 4	PF07650, PF00013	1
ENSG00000169174	PCSK9	proprotein convertase subtilisin/kexin type 9		1
ENSG00000148843	PDCD11	programmed cell death 11	PF00575	1
ENSG00000167004	PDIA3	protein disulfide isomerase family A, member 3		1
ENSG00000155660	PDIA4	protein disulfide isomerase family A, member 4		1
ENSG00000089220	PEBP1	phosphatidylethanolamine binding protein 1		1
ENSG00000242265	PEG10	paternally expressed 10		1
ENSG00000100029	PES1	pescadillo ribosomal biogenesis factor 1		1
ENSG00000100410	PHF5A	PHD finger protein 5A		1
ENSG00000156531	PHF6	PHD finger protein 6		1
ENSG00000171056	PINX1	PIN2/TERF1 interacting, telomerase inhibitor 1	PF01585	0
ENSG00000125207	PIWIL1	piwi-like 1 (Drosophila)	PF02171, PF02170	0
ENSG00000197181	PIWIL2	piwi-like 2 (Drosophila)	PF02171, PF02170	0
ENSG00000184571	PIWIL3	piwi-like 3 (Drosophila)	PF02171, PF02170	0
ENSG00000134627	PIWIL4	piwi-like 4 (Drosophila)	PF02171, PF02170	0
ENSG00000067225	PKM	pyruvate kinase, muscle		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000065243	PKN2	protein kinase N2		1
ENSG00000178209	PLEC	plectin	PF03501	1
ENSG00000146453	PNLDC1	poly(A)-specific ribonuclease (PARN)-like domain containing 1	PF04857	0
ENSG00000100941	PNN	pinin, desmosome associated protein		1
ENSG00000115946	PNO1	partner of NOB1 homolog (S. cerevisiae)	PF00013	0
ENSG00000138035	PNPT1	polyribonucleotide nucleotidyltransferase 1	PF00575, PF03726, PF00013	0
ENSG00000100227	POLDIP3	polymerase (DNA-directed), delta interacting protein 3	PF00076	1
ENSG00000051341	POLQ	polymerase (DNA directed), theta		0
ENSG00000047315	POLR2B	polymerase (RNA) II (DNA directed) polypeptide B, 140kDa		1
ENSG00000168002	POLR2G	polymerase (RNA) II (DNA directed) polypeptide G	PF00575	0
ENSG00000058600	POLR3E	polymerase (RNA) III (DNA directed) polypeptide E (80kD)	PF04801	0
ENSG00000099821	POLRMT	polymerase (RNA) mitochondrial (DNA directed)		1
ENSG00000104356	POP1	processing of precursor 1, ribonuclease P/MRP subunit (S. cerevisiae)		1
ENSG00000204531	POU5F1	POU class 5 homeobox 1		1
ENSG00000130810	PPAN	peter pan homolog (Drosophila)		1
ENSG00000109819	PPARGC1A	peroxisome proliferator-activated receptor gamma, coactivator 1 alpha	PF00076	0
ENSG00000155846	PPARGC1B	peroxisome proliferator-activated receptor gamma, coactivator 1 beta	PF00076	0
ENSG00000134283	PPHLN1	periphilin 1		1
ENSG00000196262	PPIA	peptidylprolyl isomerase A (cyclophilin A)		1
ENSG00000166794	PPIB	peptidylprolyl isomerase B (cyclophilin B)		1
ENSG00000084072	PPIE	peptidylprolyl isomerase E (cyclophilin E)	PF00076	1
ENSG00000138398	PPIG	peptidylprolyl isomerase G (cyclophilin G)		1
ENSG00000131013	PPIL4	peptidylprolyl isomerase (cyclophilin)-like 4	PF00076	1
ENSG00000148840	PPRC1	peroxisome proliferator-activated receptor gamma, coactivator-related 1	PF00076	1
ENSG00000117450	PRDX1	peroxiredoxin 1		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000180228	PRKRA	protein kinase, interferon-inducible double stranded RNA dependent activator	PF00035	0
ENSG00000105618	PRPF31	PRP31 pre-mRNA processing factor 31 homolog (S. cerevisiae)	PF01798	1
ENSG00000134186	PRPF38B	PRP38 pre-mRNA processing factor 38 (yeast) domain containing B		1
ENSG00000101161	PRPF6	PRP6 pre-mRNA processing factor 6 homolog (S. cerevisiae)		1
ENSG00000174231	PRPF8	PRP8 pre-mRNA processing factor 8 homolog (S. cerevisiae)	PF10597, PF10598	1
ENSG00000204576	PRR3	proline rich 3		1
ENSG00000130723	PRRC2B	proline-rich coiled-coil 2B		1
ENSG00000117523	PRRC2C	proline-rich coiled-coil 2C		1
ENSG00000164985	PSIP1	PC4 and SFRS1 interacting protein 1		1
ENSG00000100764	PSMC1	proteasome (prosome, macropain) 26S subunit, ATPase, 1		1
ENSG00000159352	PSMD4	proteasome (prosome, macropain) 26S subunit, non-ATPase, 4		1
ENSG00000121390	PSPC1	paraspeckle component 1	PF00076	1
ENSG00000011304	PTBP1	polypyrimidine tract binding protein 1	PF00076	1
ENSG00000117569	PTBP2	polypyrimidine tract binding protein 2	PF00076	1
ENSG00000106246	PTCD1	pentatricopeptide repeat domain 1		1
ENSG00000049883	PTCD2	pentatricopeptide repeat domain 2		1
ENSG00000132300	PTCD3	pentatricopeptide repeat domain 3		1
ENSG00000196396	PTPN1	protein tyrosine phosphatase, non-receptor type 1		1
ENSG00000177469	PTRF	polymerase I and transcript release factor		1
ENSG00000187024	PTRH1	peptidyl-tRNA hydrolase 1 homolog (S. cerevisiae)		1
ENSG00000179950	PUF60	poly-U binding splicing factor 60KDa	PF00076	1
ENSG00000134644	PUM1	pumilio homolog 1 (Drosophila)	PF00806	1
ENSG00000055917	PUM2	pumilio homolog 2 (Drosophila)	PF00806	1
ENSG00000185129	PURA	purine-rich element binding protein A	PF04845	1
ENSG00000146676	PURB	purine-rich element binding protein B	PF04845	1
ENSG00000172733	PURG	purine-rich element binding protein G	PF04845	0
ENSG00000177192	PUS1	pseudouridylate synthase 1		1
ENSG00000091127	PUS7	pseudouridylate synthase 7 homolog (S. cerevisiae)	PF01142	1
ENSG00000129317	PUS7L	pseudouridylate synthase 7 homolog (S. cerevisiae)-like	PF01142	0
ENSG00000241945	PWP2	PWP2 periodic tryptophan protein homolog (yeast)		1
ENSG00000172053	QARS	glutaminyl-tRNA synthetase		0

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000112531	QKI	QKI, KH domain containing, RNA binding	PF00013	0
ENSG00000048991	R3HDM1	R3H domain containing 1		1
ENSG00000179912	R3HDM2	R3H domain containing 2		1
ENSG00000125970	RALY	RNA binding protein, autoantigenic (hnRNP-associated with lethal yellow homolog (mouse))	PF00076	1
ENSG00000184672	RALYL	RALY RNA binding protein-like	PF00076	1
ENSG00000132341	RAN	RAN, member RAS oncogene family		1
ENSG00000161847	RAVER1	ribonucleoprotein, PTB-binding 1	PF00076	1
ENSG00000162437	RAVER2	ribonucleoprotein, PTB-binding 2	PF00076	1
ENSG00000078328	RBFOX1	RNA binding protein, fox-1 homolog (C. elegans) 1	PF00076	0
ENSG00000100320	RBFOX2	RNA binding protein, fox-1 homolog (C. elegans) 2	PF00076	1
ENSG00000167281	RBFOX3	RNA binding protein, fox-1 homolog (C. elegans) 3	PF00076	0
ENSG00000182872	RBM10	RNA binding motif protein 10	PF01585, PF00076	1
ENSG00000185272	RBM11	RNA binding motif protein 11	PF00076	0
ENSG00000244462	RBM12	RNA binding motif protein 12	PF00076	1
ENSG00000183808	RBM12B	RNA binding motif protein 12B	PF00076	1
ENSG00000239306	RBM14	RNA binding motif protein 14	PF00076	1
ENSG00000162775	RBM15	RNA binding motif protein 15	PF00076	1
ENSG00000179837	RBM15B	RNA binding motif protein 15B	PF00076	1
ENSG00000213079	SCAF8	SR-related CTD-associated factor 8	PF00076	0
ENSG00000134453	RBM17	RNA binding motif protein 17	PF01585	0
ENSG00000119446	RBM18	RNA binding motif protein 18	PF00076	0
ENSG00000122965	RBM19	RNA binding motif protein 19	PF00076	1
ENSG00000086589	RBM22	RNA binding motif protein 22	PF00076	1
ENSG00000100461	RBM23	RNA binding motif protein 23	PF00076	0
ENSG00000112183	RBM24	RNA binding motif protein 24	PF00076	0
ENSG00000119707	RBM25	RNA binding motif protein 25	PF00076	1
ENSG00000139746	RBM26	RNA binding motif protein 26	PF00076	1
ENSG00000091009	RBM27	RNA binding motif protein 27		1
ENSG00000106344	RBM28	RNA binding motif protein 28	PF00076	1
ENSG00000102317	RBM3	RNA binding motif (RNP1, RRM) protein 3	PF00076	1
ENSG00000184863	RBM33	RNA binding motif protein 33	PF00076	1
ENSG00000188739	RBM34	RNA binding motif protein 34	PF00076	1
ENSG00000132819	RBM38	RNA binding motif protein 38	PF00076	1
ENSG00000131051	RBM39	RNA binding motif protein 39	PF00076	1
ENSG00000173933	RBM4	RNA binding motif protein 4	PF00098, PF00076	1
ENSG00000089682	RBM41	RNA binding motif protein 41	PF00076	0

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000126254	RBM42	RNA binding motif protein 42	PF00076	1
ENSG00000177483	RBM44	RNA binding motif protein 44	PF00076	0
ENSG00000155636	RBM45	RNA binding motif protein 45	PF00076	1
ENSG00000151962	RBM46	RNA binding motif protein 46	PF00076	0
ENSG00000163694	RBM47	RNA binding motif protein 47	PF00076	1
ENSG00000173914	RBM4B	RNA binding motif protein 4B	PF00098, PF00076	1
ENSG00000003756	RBM5	RNA binding motif protein 5	PF01585, PF00076	0
ENSG00000004534	RBM6	RNA binding motif protein 6	PF01585	1
ENSG00000076053	RBM7	RNA binding motif protein 7	PF00076	1
ENSG00000131795	RBM8A	RNA binding motif protein 8A	PF08777, PF00076	0
ENSG00000153250	RBMS1	RNA binding motif, single stranded interacting protein 1	PF00076	1
ENSG00000076067	RBMS2	RNA binding motif, single stranded interacting protein 2	PF00076	1
ENSG00000144642	RBMS3	RNA binding motif, single stranded interacting protein 3	PF00076	0
ENSG00000147274	RBMX	RNA binding motif protein, X-linked	PF08081, PF00076	1
ENSG00000134597	RBMX2	RNA binding motif protein, X-linked 2	PF00076	1
ENSG00000213516	RBMXL1	RNA binding motif protein, X-linked-like 1	PF08081, PF00076	0
ENSG00000170748	RBMXL2	RNA binding motif protein, X-linked-like 2	PF08081, PF00076	0
ENSG00000175718	RBMXL3	RNA binding motif protein, X-linked-like 3	PF00076	0
ENSG00000234414	RBMY1A1	RNA binding motif protein, Y-linked, family 1, member A1	PF00076, PF08081	0
ENSG00000242875	RBMY1B	RNA binding motif protein, Y-linked, family 1, member B	PF00076, PF08081	0
ENSG00000244395	RBMY1D	RNA binding motif protein, Y-linked, family 1, member D	PF00076, PF08081	0
ENSG00000242389	RBMY1E	RNA binding motif protein, Y-linked, family 1, member E	PF00076, PF08081	0
ENSG00000169800	RBMY1F	RNA binding motif protein, Y-linked, family 1, member F	PF00076, PF08081	0
ENSG00000226941	RBMY1J	RNA binding motif protein, Y-linked, family 1, member J	PF00076, PF08081	0
ENSG00000157110	RBPMS	RNA binding protein with multiple splicing	PF00076	1
ENSG00000166831	RBPMS2	RNA binding protein with multiple splicing 2	PF00076	0
ENSG00000135870	RC3H1	ring finger and CCCH-type domains 1		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000056586	RC3H2	ring finger and CCCH-type domains 2		1
ENSG00000179051	RCC2	regulator of chromosome condensation 2		1
ENSG00000204356	RDBP	RD RNA binding protein	PF00076	1
ENSG00000137710	RDX	radixin		1
ENSG00000004700	RECQL	RecQ protein-like (DNA helicase Q1-like)	PF00270	0
ENSG00000160957	RECQL4	RecQ protein-like 4		0
ENSG00000108469	RECQL5	RecQ protein-like 5	PF00270	0
ENSG00000214022	REPIN1	replication initiator 1		1
ENSG00000148300	REXO4	REX4, RNA exonuclease 4 homolog (S. cerevisiae)		1
ENSG00000174173	TRMT10C	tRNA methyltransferase 10 homolog C (S. cerevisiae)		1
ENSG00000145331	TRMT10A	tRNA methyltransferase 10 homolog A (S. cerevisiae)		1
ENSG00000079841	RIMS1	regulating synaptic membrane exocytosis 1		1
ENSG00000132972	RNF17	ring finger protein 17	PF00567	0
ENSG00000171861	RNMTL1	RNA methyltransferase like 1		1
ENSG00000185946	RNPC3	RNA-binding region (RNP1, RRM) containing 3	PF00076	0
ENSG00000205937	RNPS1	RNA binding protein S1, serine-rich domain	PF00076	1
ENSG00000119314	PTBP3	polypyrimidine tract binding protein 3	PF00076	1
ENSG00000197498	RPF2	ribosome production factor 2 homolog (S. cerevisiae)		1
ENSG00000156313	RPGR	retinitis pigmentosa GTPase regulator		1
ENSG00000147403	RPL10	ribosomal protein L10		1
ENSG00000198755	RPL10A	ribosomal protein L10a		1
ENSG00000167526	RPL13	ribosomal protein L13		1
ENSG00000142541	RPL13A	ribosomal protein L13a		1
ENSG00000188846	RPL14	ribosomal protein L14		1
ENSG00000174748	RPL15	ribosomal protein L15		1
ENSG00000215472	RPL17	ribosomal protein L17		1
ENSG00000105640	RPL18A	ribosomal protein L18a		1
ENSG00000108298	RPL19	ribosomal protein L19		1
ENSG00000122026	RPL21	ribosomal protein L21		1
ENSG00000116251	RPL22	ribosomal protein L22		1
ENSG00000125691	RPL23	ribosomal protein L23		1
ENSG00000198242	RPL23A	ribosomal protein L23a		1
ENSG00000137876	RSL24D1	ribosomal L24 domain containing 1		1
ENSG00000131469	RPL27	ribosomal protein L27		1
ENSG00000108107	RPL28	ribosomal protein L28		1
ENSG00000162244	RPL29	ribosomal protein L29		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000100316	RPL3	ribosomal protein L3		1
ENSG00000156482	RPL30	ribosomal protein L30		1
ENSG00000071082	RPL31	ribosomal protein L31		1
ENSG00000144713	RPL32	ribosomal protein L32		1
ENSG00000136942	RPL35	ribosomal protein L35		1
ENSG00000182899	RPL35A	ribosomal protein L35a		1
ENSG00000130255	RPL36	ribosomal protein L36		1
ENSG00000174444	RPL4	ribosomal protein L4		1
ENSG00000122406	RPL5	ribosomal protein L5		1
ENSG00000089009	RPL6	ribosomal protein L6		1
ENSG00000147604	RPL7	ribosomal protein L7		1
ENSG00000240522	RPL7AP10	ribosomal protein L7a pseudogene 10		1
ENSG00000146223	RPL7L1	ribosomal protein L7-like 1		1
ENSG00000161016	RPL8	ribosomal protein L8		1
ENSG00000089157	RPLP0	ribosomal protein, large, P0		1
ENSG00000163902	RPN1	ribophorin I		1
ENSG00000241370	RPP21	ribonuclease P/MRP 21kDa subunit	PF04032	0
ENSG00000178718	RPP25	ribonuclease P/MRP 25kDa subunit		1
ENSG00000148688	RPP30	ribonuclease P/MRP 30kDa subunit		1
ENSG00000124614	RPS10	ribosomal protein S10	PF03501	1
ENSG00000142534	RPS11	ribosomal protein S11		1
ENSG00000112306	RPS12	ribosomal protein S12		1
ENSG00000115268	RPS15	ribosomal protein S15		1
ENSG00000134419	RPS15A	ribosomal protein S15a		1
ENSG00000187051	RPS19BP1	ribosomal protein S19 binding protein 1		1
ENSG00000140988	RPS2	ribosomal protein S2		1
ENSG00000008988	RPS20	ribosomal protein S20		1
ENSG00000171858	RPS21	ribosomal protein S21		1
ENSG00000138326	RPS24	ribosomal protein S24		1
ENSG00000197728	RPS26P32	ribosomal protein S26 pseudogene 32		1
ENSG00000143947	RPS27A	ribosomal protein S27a		1
ENSG00000233927	RPS28	ribosomal protein S28		1
ENSG00000149273	RPS3	ribosomal protein S3		1
ENSG00000145425	RPS3A	ribosomal protein S3A		1
ENSG00000198034	RPS4X	ribosomal protein S4, X-linked		1
ENSG00000083845	RPS5	ribosomal protein S5		1
ENSG00000137154	RPS6	ribosomal protein S6		1
ENSG00000171863	RPS7	ribosomal protein S7		1
ENSG00000142937	RPS8	ribosomal protein S8		1
ENSG00000170889	RPS9	ribosomal protein S9		1
ENSG00000166133	RPUSD2	RNA pseudouridylate synthase domain containing 2		1
ENSG00000156990	RPUSD3	RNA pseudouridylate synthase domain containing 3		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000165526	RPUSD4	RNA pseudouridylate synthase domain containing 4		1
ENSG00000160214	RRP1	ribosomal RNA processing 1 homolog (S. cerevisiae)		1
ENSG00000052749	RRP12	ribosomal RNA processing 12 homolog (S. cerevisiae)		1
ENSG00000160208	RRP1B	ribosomal RNA processing 1 homolog B (S. cerevisiae)		1
ENSG00000124541	RRP36	ribosomal RNA processing 36 homolog (S. cerevisiae)		1
ENSG00000189306	RRP7A	ribosomal RNA processing 7 homolog A (S. cerevisiae)		1
ENSG00000132275	RRP8	ribosomal RNA processing 8, methyltransferase, homolog (yeast)		1
ENSG00000114767	RRP9	ribosomal RNA processing 9, small subunit (SSU) processome component, homolog (yeast)		1
ENSG00000179041	RRS1	RRS1 ribosome biogenesis regulator homolog (S. cerevisiae)		1
ENSG00000171490	RSL1D1	ribosomal L1 domain containing 1		1
ENSG00000137996	RTCA	RNA 3'-terminal phosphate cyclase		1
ENSG00000137815	RTF1	Rtf1, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae)		1
ENSG00000115310	RTN4	reticulon 4		1
ENSG00000188643	S100A16	S100 calcium binding protein A16		1
ENSG00000196154	S100A4	S100 calcium binding protein A4		1
ENSG00000160633	SAFB	scaffold attachment factor B	PF00076	1
ENSG00000130254	SAFB2	scaffold attachment factor B2	PF00076	1
ENSG00000020577	SAMD4A	sterile alpha motif domain containing 4A		1
ENSG00000155307	SAMSN1	SAM domain, SH3 domain and nuclear localization signals 1		1
ENSG00000205323	SARNP	SAP domain containing ribonucleoprotein		1
ENSG00000175467	SART1	squamous cell carcinoma antigen recognized by T cells		1
ENSG00000075856	SART3	squamous cell carcinoma antigen recognized by T cells 3	PF00076	1
ENSG00000126524	SBDS	Shwachman-Bodian-Diamond syndrome		1
ENSG00000139218	SCAF11	SR-related CTD-associated factor 11		1
ENSG00000104112	SCG3	secretogranin III		1
ENSG00000107651	SEC23IP	SEC23 interacting protein		1
ENSG00000106803	SEC61B	Sec61 beta subunit		1
ENSG00000025796	SEC63	SEC63 homolog (S. cerevisiae)		1
ENSG00000187742	SECISBP2	SECIS binding protein 2		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000138593	SECISBP2L	SECIS binding protein 2-like		1
ENSG00000142864	SERBP1	SERPINE1 mRNA binding protein 1	PF04774	1
ENSG00000149257	SERPINH1	serpin peptidase inhibitor, clade H (heat shock protein 47), member 1, (collagen binding protein 1)		1
ENSG00000099381	SETD1A	SET domain containing 1A	PF00076	0
ENSG00000139718	SETD1B	SET domain containing 1B	PF00076	0
ENSG00000168066	SF1	splicing factor 1	PF00013, PF00098	1
ENSG00000099995	SF3A1	splicing factor 3a, subunit 1, 120kDa	PF01805	1
ENSG00000104897	SF3A2	splicing factor 3a, subunit 2, 66kDa		1
ENSG00000183431	SF3A3	splicing factor 3a, subunit 3, 60kDa		1
ENSG00000115524	SF3B1	splicing factor 3b, subunit 1, 155kDa		0
ENSG00000087365	SF3B2	splicing factor 3b, subunit 2, 145kDa		1
ENSG00000143368	SF3B4	splicing factor 3b, subunit 4, 49kDa	PF00076	1
ENSG00000116560	SFPQ	splicing factor proline/glutamine-rich	PF00076	1
ENSG00000156304	SCAF4	SR-related CTD-associated factor 4	PF00076	0
ENSG00000061936	SFSWAP	splicing factor, suppressor of white-apricot homolog (Drosophila)	PF01805	0
ENSG00000204351	SKIV2L	superkiller viralicidic activity 2-like (S. cerevisiae)	PF00270	0
ENSG00000039123	SKIV2L2	superkiller viralicidic activity 2-like 2 (S. cerevisiae)	PF00270	1
ENSG00000163950	SLBP	stem-loop binding protein		1
ENSG00000141526	SLC16A3	solute carrier family 16, member 3 (monocarboxylic acid transporter 4)		1
ENSG00000005022	SLC25A5	solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 5		1
ENSG00000168003	SLC3A2	solute carrier family 3 (activators of dibasic and neutral amino acid transport), member 2		1
ENSG00000163798	SLC4A1AP	solute carrier family 4 (anion exchanger), member 1, adaptor protein	PF00035	0
ENSG00000137776	SLTM	SAFB-like, transcription modulator	PF00076	1
ENSG00000119953	SMNDC1	survival motor neuron domain containing 1		1
ENSG00000197157	SND1	staphylococcal nuclease and tudor domain containing 1	PF00567	1
ENSG00000163877	SNIP1	Smad nuclear interacting protein 1		1
ENSG00000265236	SNORD84	small nucleolar RNA, C/D box 84	PF00270	1
ENSG00000144028	SNRNP200	small nuclear ribonucleoprotein 200kDa (U5)	PF00270	1
ENSG00000184209	SNRNP35	small nuclear ribonucleoprotein 35kDa (U11/U12)	PF00076	0

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000060688	SNRNP40	small nuclear ribonucleoprotein 40kDa (U5)		1
ENSG00000104852	SNRNP70	small nuclear ribonucleoprotein 70kDa (U1)	PF12220, PF00076	1
ENSG00000077312	SNRPA	small nuclear ribonucleoprotein polypeptide A	PF00076	0
ENSG00000125835	SNRPB	small nuclear ribonucleoprotein polypeptides B and B1	PF01423	0
ENSG00000125870	SNRPB2	small nuclear ribonucleoprotein polypeptide B	PF00076	0
ENSG00000167088	SNRPD1	small nuclear ribonucleoprotein D1 polypeptide 16kDa		1
ENSG00000125743	SNRPD2	small nuclear ribonucleoprotein D2 polypeptide 16.5kDa	PF01423	1
ENSG00000100028	SNRPD3	small nuclear ribonucleoprotein D3 polypeptide 18kDa	PF01423	0
ENSG00000182004	SNRPE	small nuclear ribonucleoprotein polypeptide E	PF01423	0
ENSG00000256968	SNRPEP2	small nuclear ribonucleoprotein polypeptide E pseudogene 2		0
ENSG00000139343	SNRPF	small nuclear ribonucleoprotein polypeptide F	PF01423	0
ENSG00000143977	SNRPG	small nuclear ribonucleoprotein polypeptide G	PF01423	1
ENSG00000128739	SNRPN	small nuclear ribonucleoprotein polypeptide N	PF01423	0
ENSG00000168807	SNTB2	syntrophin, beta 2 (dystrophin-associated protein A1, 59kDa, basic component 2)		1
ENSG00000159140	SON	SON DNA binding protein	PF01585, PF00035	1
ENSG00000154556	SORBS2	sorbin and SH3 domain containing 2		1
ENSG00000123352	SPATS2	spermatogenesis associated, serine-rich 2		1
ENSG00000196141	SPATS2L	spermatogenesis associated, serine-rich 2-like		1
ENSG00000065526	SPEN	spen homolog, transcriptional regulator (Drosophila)		1
ENSG00000068784	SRBD1	S1 RNA binding domain 1	PF00575	0
ENSG00000153914	SREK1	splicing regulatory glutamine/lysine-rich protein 1	PF00076	0
ENSG00000151304	SRFBP1	serum response factor binding protein 1		1
ENSG00000140319	SRP14	signal recognition particle 14kDa (homologous Alu RNA binding protein)	PF02290	1
ENSG00000153037	SRP19	signal recognition particle 19kDa	PF00076	1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000100883	SRP54	signal recognition particle 54kDa		1
ENSG00000167881	SRP68	signal recognition particle 68kDa		1
ENSG00000174780	SRP72	signal recognition particle 72kDa	PF08492	1
ENSG00000135250	SRPK2	SRSF protein kinase 2		1
ENSG00000182934	SRPR	signal recognition particle receptor (docking protein)		1
ENSG00000167978	SRRM2	serine/arginine repetitive matrix 2		1
ENSG00000087087	SRRT	serrate RNA effector molecule homolog (Arabidopsis)		1
ENSG00000136450	SRSF1	serine/arginine-rich splicing factor 1	PF00076	1
ENSG00000188529	SRSF10	serine/arginine-rich splicing factor 10	PF00076	0
ENSG00000116754	SRSF11	serine/arginine-rich splicing factor 11	PF00076	1
ENSG00000154548	SRSF12	serine/arginine-rich splicing factor 12	PF00076	0
ENSG00000161547	SRSF2	serine/arginine-rich splicing factor 2	PF00076	1
ENSG00000112081	SRSF3	serine/arginine-rich splicing factor 3	PF00076	1
ENSG00000116350	SRSF4	serine/arginine-rich splicing factor 4	PF00076	1
ENSG00000100650	SRSF5	serine/arginine-rich splicing factor 5	PF00076	1
ENSG00000124193	SRSF6	serine/arginine-rich splicing factor 6	PF08777, PF00076	1
ENSG00000115875	SRSF7	serine/arginine-rich splicing factor 7	PF00098, PF00076,	1
ENSG00000180771	SRSF8	serine/arginine-rich splicing factor 8		1
ENSG00000111786	SRSF9	serine/arginine-rich splicing factor 9	PF00076	1
ENSG00000138385	SSB	Sjogren syndrome antigen B (autoantigen La)	PF08777, PF00076, PF05383	1
ENSG00000106028	SSBP1	single-stranded DNA binding protein 1, mitochondrial		1
ENSG00000149136	SSRP1	structure specific recognition protein 1		1
ENSG00000124214	STAU1	staufen, RNA binding protein, homolog 1 (Drosophila)	PF00035	0
ENSG00000040341	STAU2	staufen, RNA binding protein, homolog 2 (Drosophila)	PF00035	1
ENSG00000168439	STIP1	stress-induced-phosphoprotein 1		1
ENSG00000196335	STK31	serine/threonine kinase 31	PF00567	0
ENSG00000023734	STRAP	serine/threonine kinase receptor associated protein		1
ENSG00000165209	STRBP	spermatid perinuclear RNA binding protein	PF00035	1
ENSG00000136854	STXBP1	syntaxin binding protein 1		1
ENSG00000113387	SUB1	SUB1 homolog (S. cerevisiae)		1
ENSG00000163541	SUCLG1	succinate-CoA ligase, alpha subunit		1
ENSG00000105705	SUGP1	SURP and G patch domain containing 1	PF01805, PF01585	0

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000064607	SUGP2	SURP and G patch domain containing 2	PF01805, PF01585	1
ENSG00000116030	SUMO1	SMT3 suppressor of mif two 3 homolog 1 (S. cerevisiae)		1
ENSG00000092201	SUPT16H	suppressor of Ty 16 homolog (S. cerevisiae)		1
ENSG00000196235	SUPT5H	suppressor of Ty 5 homolog (S. cerevisiae)		1
ENSG00000109111	SUPT6H	suppressor of Ty 6 homolog (S. cerevisiae)	PF00575	1
ENSG00000156502	SUPV3L1	suppressor of var1, 3-like 1 (S. cerevisiae)		1
ENSG00000148296	SURF6	surfeit 6		1
ENSG00000117614	SYF2	SYF2 homolog, RNA splicing factor (S. cerevisiae)		1
ENSG00000135316	SYNCRIP	synaptotagmin binding, cytoplasmic RNA interacting protein	PF00076	1
ENSG00000131018	SYNE1	spectrin repeat containing, nuclear envelope 1		1
ENSG00000172660	TAF15	TAF15 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 68kDa	PF00076	1
ENSG00000139546	TARBP2	TAR (HIV-1) RNA binding protein 2	PF00035	0
ENSG00000120948	TARDBP	TAR DNA binding protein	PF00076	1
ENSG00000106638	TBL2	transducin (beta)-like 2		1
ENSG00000183751	TBL3	transducin (beta)-like 3		1
ENSG00000136270	TBRG4	transforming growth factor beta regulator 4	PF08373	1
ENSG00000113649	TCERG1	transcription elongation regulator 1		1
ENSG00000100207	TCF20	transcription factor 20 (AR1)		1
ENSG00000070814	TCOF1	Treacher Collins-Franceschetti syndrome 1		1
ENSG00000095627	TDRD1	tudor domain containing 1	PF00567	0
ENSG00000163239	TDRD10	tudor domain containing 10	PF00567, PF00076	0
ENSG00000173809	TDRD12	tudor domain containing 12	PF00567, PF00270	0
ENSG00000083544	TDRD3	tudor domain containing 3	PF00567	1
ENSG00000162782	TDRD5	tudor domain containing 5	PF00567	0
ENSG00000180113	TDRD6	tudor domain containing 6	PF00567	0
ENSG00000196116	TDRD7	tudor domain containing 7	PF00567	0
ENSG00000156414	TDRD9	tudor domain containing 9	PF00567, PF00270	0

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000182134	TDRKH	tudor and KH domain containing	PF07650, PF00567, PF00013	0
ENSG00000129566	TEP1	telomerase-associated protein 1	PF05731	0
ENSG00000135269	TES	testis derived transcript (3 LIM domains)		1
ENSG00000108064	TFAM	transcription factor A, mitochondrial		1
ENSG00000029639	TFB1M	transcription factor B1, mitochondrial		1
ENSG00000100109	TFIP11	tuftelin interacting protein 11	PF01585	0
ENSG00000072274	TFRC	transferrin receptor (p90, CD71)		1
ENSG00000183684	ALYREF	Aly/REF export factor	PF00076	1
ENSG00000054118	THRAP3	thyroid hormone receptor associated protein 3		1
ENSG00000066654	THUMPD1	THUMP domain containing 1		1
ENSG00000116001	TIA1	TIA1 cytotoxic granule-associated RNA binding protein	PF00076	1
ENSG00000151923	TIAL1	TIA1 cytotoxic granule-associated RNA binding protein-like 1	PF00076	1
ENSG00000205542	TMSB4X	thymosin beta 4, X-linked		1
ENSG00000083312	TNPO1	transportin 1		1
ENSG00000100354	TNRC6B	trinucleotide repeat containing 6B		1
ENSG00000078687	TNRC6C	trinucleotide repeat containing 6C		0
ENSG00000079308	TNS1	tensin 1		1
ENSG00000132773	TOE1	target of EGR1, member 1 (nuclear)	PF04857	0
ENSG00000198900	TOP1	topoisomerase (DNA) I		1
ENSG00000100038	TOP3B	topoisomerase (DNA) III beta		1
ENSG00000101150	TPD52L2	tumor protein D52-like 2		1
ENSG00000047410	TPR	translocated promoter region, nuclear basket protein		1
ENSG00000133112	TPT1	tumor protein, translationally-controlled 1		1
ENSG00000164548	TRA2A	transformer 2 alpha homolog (Drosophila)	PF00076	1
ENSG00000136527	TRA2B	transformer 2 beta homolog (Drosophila)	PF00076	1
ENSG00000126602	TRAP1	TNF receptor-associated protein 1		1
ENSG00000121060	TRIM25	tripartite motif containing 25		1
ENSG00000204599	TRIM39	tripartite motif containing 39	PF04032	0
ENSG00000169871	TRIM56	tripartite motif containing 56		1
ENSG00000087077	TRIP6	thyroid hormone receptor interactor 6		1
ENSG00000104907	TRMT1	tRNA methyltransferase 1 homolog (S. cerevisiae)		1
ENSG00000121486	TRMT1L	tRNA methyltransferase 1 homolog (S. cerevisiae)-like		1
ENSG00000099899	TRMT2A	tRNA methyltransferase 2 homolog A (S. cerevisiae)		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000089195	TRMT6	tRNA methyltransferase 6 homolog (S. cerevisiae)		1
ENSG00000180098	TRNAU1AP	tRNA selenocysteine 1 associated protein 1		1
ENSG00000116747	TROVE2	TROVE domain family, member 2	PF05731	0
ENSG00000167112	TRUB2	TruB pseudouridine (psi) synthase homolog 2 (E. coli)		1
ENSG00000167721	TSR1	TSR1, 20S rRNA accumulation, homolog (S. cerevisiae)		1
ENSG00000178952	TUFM	Tu translation elongation factor, mitochondrial		1
ENSG00000149016	TUT1	terminal uridylyl transferase 1, U6 snRNA-specific	PF00076	0
ENSG00000247596	TWF2	twinfilin, actin-binding protein, homolog 2 (Drosophila)		1
ENSG00000136810	TXN	thioredoxin		1
ENSG00000160201	U2AF1	U2 small nuclear RNA auxiliary factor 1	PF00076	1
ENSG00000063244	U2AF2	U2 small nuclear RNA auxiliary factor 2	PF00076	1
ENSG00000130985	UBA1	ubiquitin-like modifier activating enzyme 1		1
ENSG00000137073	UBAP2	ubiquitin associated protein 2		1
ENSG00000143569	UBAP2L	ubiquitin associated protein 2-like		1
ENSG00000103275	UBE2I	ubiquitin-conjugating enzyme E2I		1
ENSG00000185651	UBE2L3	ubiquitin-conjugating enzyme E2L 3		1
ENSG00000103353	UBFD1	ubiquitin family domain containing 1		1
ENSG00000116750	UCHL5	ubiquitin carboxyl-terminal hydrolase L5		1
ENSG00000152332	UHMK1	U2AF homology motif (UHM) kinase 1	PF00076	0
ENSG00000132478	UNK	unkempt homolog (Drosophila)		1
ENSG00000005007	UPF1	UPF1 regulator of nonsense transcripts homolog (yeast)		1
ENSG00000151461	UPF2	UPF2 regulator of nonsense transcripts homolog (yeast)		0
ENSG00000125351	UPF3B	UPF3 regulator of nonsense transcripts homolog B (yeast)		1
ENSG00000142207	URB1	URB1 ribosome biogenesis 1 homolog (S. cerevisiae)		1
ENSG00000138768	USO1	USO1 vesicle docking protein homolog (yeast)		1
ENSG00000103194	USP10	ubiquitin specific peptidase 10		1
ENSG00000055483	USP36	ubiquitin specific peptidase 36		1
ENSG00000183520	UTP11L	UTP11-like, U3 small nucleolar ribonucleoprotein, (yeast)		1
ENSG00000156697	UTP14A	UTP14, U3 small nucleolar ribonucleoprotein, homolog A (yeast)		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000164338	UTP15	UTP15, U3 small nucleolar ribonucleoprotein, homolog (S. cerevisiae)		1
ENSG00000011260	UTP18	UTP18 small subunit (SSU) processome component homolog (yeast)		1
ENSG00000120800	UTP20	UTP20, small subunit (SSU) processome component, homolog (yeast)		1
ENSG00000147679	UTP23	UTP23, small subunit (SSU) processome component, homolog (yeast)		1
ENSG00000132467	UTP3	UTP3, small subunit (SSU) processome component, homolog (S. cerevisiae)		1
ENSG00000174374	WBSCR16	Williams-Beuren syndrome chromosome region 16		1
ENSG00000065183	WDR3	WD repeat domain 3		1
ENSG00000134987	WDR36	WD repeat domain 36		1
ENSG00000163811	WDR43	WD repeat domain 43		1
ENSG00000178252	WDR6	WD repeat domain 6		1
ENSG00000115368	WDR75	WD repeat domain 75		1
ENSG00000165392	WRN	Werner syndrome, RecQ helicase-like	PF00270	0
ENSG00000168334	XIRP1	xin actin-binding repeat containing 1		1
ENSG00000082898	XPO1	exportin 1 (CRM1 homolog, yeast)		0
ENSG00000124571	XPO5	exportin 5		1
ENSG00000079246	XRCC5	X-ray repair complementing defective repair in Chinese hamster cells 5 (double-strand-break rejoining)		1
ENSG00000196419	XRCC6	X-ray repair complementing defective repair in Chinese hamster cells 6		1
ENSG00000114127	XRN1	5'-3' exoribonuclease 1		1
ENSG00000088930	XRN2	5'-3' exoribonuclease 2		1
ENSG00000134684	YARS	tyrosyl-tRNA synthetase	PF01588	1
ENSG00000065978	YBX1	Y box binding protein 1		1
ENSG00000083896	YTHDC1	YTH domain containing 1		1
ENSG00000047188	YTHDC2	YTH domain containing 2	PF00270	1
ENSG00000149658	YTHDF1	YTH domain family, member 1		1
ENSG00000198492	YTHDF2	YTH domain family, member 2		1
ENSG00000185728	YTHDF3	YTH domain family, member 3		1
ENSG00000108953	YWHAE	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide		1
ENSG00000170027	YWHAG	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, gamma polypeptide		1

Table 4.2 (continued): Compiled set of 1072 putative RNA binding proteins

Ensembl ID	Gene Symbol	Gene Description	PFAM Domain IDs	RBP Interact-ome (Castello et al., 2012)
ENSG00000164924	YWHAZ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide		1
ENSG00000135482	ZC3H10	zinc finger CCCH-type containing 10		1
ENSG00000058673	ZC3H11A	zinc finger CCCH-type containing 11A		1
ENSG00000100722	ZC3H14	zinc finger CCCH-type containing 14		1
ENSG00000065548	ZC3H15	zinc finger CCCH-type containing 15		1
ENSG00000158545	ZC3H18	zinc finger CCCH-type containing 18		1
ENSG00000130749	ZC3H4	zinc finger CCCH-type containing 4		1
ENSG00000122299	ZC3H7A	zinc finger CCCH-type containing 7A	PF12171	1
ENSG00000100403	ZC3H7B	zinc finger CCCH-type containing 7B		1
ENSG00000144161	ZC3H8	zinc finger CCCH-type containing 8		1
ENSG00000105939	ZC3HAV1	zinc finger CCCH-type, antiviral 1		1
ENSG00000134744	ZCCHC11	zinc finger, CCHC domain containing 11	PF00098	1
ENSG00000121766	ZCCHC17	zinc finger, CCHC domain containing 17	PF00575	1
ENSG00000177764	ZCCHC3	zinc finger, CCHC domain containing 3	PF00098	1
ENSG00000083223	ZCCHC6	zinc finger, CCHC domain containing 6	PF00098	1
ENSG00000147905	ZCCHC7	zinc finger, CCHC domain containing 7	PF00098	1
ENSG00000033030	ZCCHC8	zinc finger, CCHC domain containing 8	PF00098	1
ENSG00000131732	ZCCHC9	zinc finger, CCHC domain containing 9	PF00098	1
ENSG00000139168	ZCRB1	zinc finger CCHC-type and RNA binding motif 1	PF00098, PF00076	1
ENSG00000133858	ZFC3H1	zinc finger, C3H1-type containing		1
ENSG00000103994	ZFP106	zinc finger protein 106 homolog (mouse)		1
ENSG00000128016	ZFP36	zinc finger protein 36, C3H type, homolog (mouse)		1
ENSG00000185650	ZFP36L1	zinc finger protein 36, C3H type-like 1		1
ENSG00000152518	ZFP36L2	zinc finger protein 36, C3H type-like 2		1
ENSG00000056097	ZFR	zinc finger RNA binding protein	PF12171	1
ENSG00000197114	ZGPAT	zinc finger, CCCH-type with G patch domain	PF01585	0
ENSG00000172667	ZMAT3	zinc finger, matrin-type 3	PF12171	1
ENSG00000162664	ZNF326	zinc finger protein 326		1
ENSG00000167962	ZNF598	zinc finger protein 598		1
ENSG00000173545	ZNF622	zinc finger protein 622	PF12171	1
ENSG00000075292	ZNF638	zinc finger protein 638	PF00076	1
ENSG00000124201	ZNFX1	zinc finger, NFX1-type containing 1		1
ENSG00000132485	ZRANB2	zinc finger, RAN-binding domain containing 2		1
ENSG00000212643	ZRSR1	zinc finger (CCCH type), RNA-binding motif and serine/arginine rich 1		0
ENSG00000169249	ZRSR2	zinc finger (CCCH type), RNA-binding motif and serine/arginine rich 2	PF00076	0
ENSG00000159840	ZYX	zyxin		1

4.6 References

Alkan, S.A., Martincic, K., and Milcarek, C. (2006). The hnRNPs F and H2 bind to similar sequences to influence gene expression. Biochem J *393*, 361-371.

Caputi, M., and Zahler, A.M. (2001). Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family. J Biol Chem *276*, 43850-43859.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., Krijgsveld, J., and Hentze, M.W. (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell *149*, 1393-1406.

Chen, S., Zhang, J., Duan, L., Zhang, Y., Li, C., Liu, D., Ouyang, C., Lu, F., and Liu, X. (2014). Identification of HnRNP M as a novel biomarker for colorectal carcinoma by quantitative proteomics. American journal of physiology Gastrointestinal and liver physiology *306*, G394-403.

Daniels, D.L., Mendez, J., Mosley, A.L., Ramisetty, S.R., Murphy, N., Benink, H., Wood, K.V., Urh, M., and Washburn, M.P. (2012). Examining the complexity of human RNA polymerase complexes using HaloTag technology coupled to label free quantitative proteomics. Journal of proteome research *11*, 564-575.

Deplus, R., Delatte, B., Schwinn, M.K., Defrance, M., Mendez, J., Murphy, N., Dawson, M.A., Volkmar, M., Putmans, P., Calonne, E., Shih, A.H., Levine, R.L., Bernard, O., Mercher, T., Solary, E., Urh, M., Daniels, D.L., and Fuks, F. (2013). TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and SET1/COMPASS. EMBO J *32*, 645-655.

Flury, V., Restuccia, U., Bachi, A., and Muhlemann, O. (2014). Characterization of Phosphorylation- and RNA-Dependent UPF1 Interactors by Quantitative Proteomics. Journal of proteome research.

Gregersen, L.H., Schueler, M., Munschauer, M., Mastrobuoni, G., Chen, W., Kempa, S., Dieterich, C., and Landthaler, M. (2014). MOV10 Is a 5' to 3' RNA Helicase Contributing to UPF1 mRNA Target Degradation by Translocation along 3' UTRs. Mol Cell.

He, Y., Brown, M.A., Rothnagel, J.A., Saunders, N.A., and Smith, R. (2005). Roles of heterogeneous nuclear ribonucleoproteins A and B in cell proliferation. J Cell Sci *118*, 3173-3183.

Honore, B., Baandrup, U., and Vorum, H. (2004). Heterogeneous nuclear ribonucleoproteins F and H/H' show differential expression in normal and selected cancer tissues. Experimental cell research *294*, 199-209.

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., Jr., and Yeo, G.W. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Rep *1*, 167-178.

Kervestin, S., and Jacobson, A. (2012). NMD: a multifaceted response to premature translational termination. Nat Rev Mol Cell Biol *13*, 700-712.

Kwon, S.C., Yi, H., Eichelbaum, K., Fohr, S., Fischer, B., You, K.T., Castello, A., Krijgsveld, J., Hentze, M.W., and Kim, V.N. (2013). The RNA-binding protein repertoire of embryonic stem cells. Nat Struct Mol Biol *20*, 1122-1130.

Lovci, M.T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T.Y., Stark, T.J., Gehman, L.T., Hoon, S., Massirer, K.B., Pratt, G.A., Black, D.L., Gray, J.W., Conboy, J.G., and Yeo, G.W. (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat Struct Mol Biol *20*, 1434-1442.

Lukong, K.E., Chang, K.W., Khandjian, E.W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. Trends Genet *24*, 416-425.

Scherrer, T., Mittal, N., Janga, S.C., and Gerber, A.P. (2010). A screen for RNA-binding proteins in yeast indicates dual functions for many enzymes. PLoS One *5*, e15499.

Tsvetanova, N.G., Klass, D.M., Salzman, J., and Brown, P.O. (2010). Proteome-wide search reveals unexpected RNA-binding proteins in Saccharomyces cerevisiae. PLoS One *5*.

Vazquez-Chantada, M., Fernandez-Ramos, D., Embade, N., Martinez-Lopez, N., Varela-Rey, M., Woodhoo, A., Luka, Z., Wagner, C., Anglim, P.P., Finnell, R.H., Caballeria, J., Laird-Offringa, I.A., Gorospe, M., Lu, S.C., Mato, J.M., and Martinez-Chantar, M.L. (2010). HuR/methyl-HuR and AUF1 regulate the MAT expressed during liver proliferation, differentiation, and carcinogenesis. Gastroenterology *138*, 1943-1953.

Zhang, W., Wagner, B.J., Ehrenman, K., Schaefer, A.W., DeMaria, C.T., Crater, D., DeHaven, K., Long, L., and Brewer, G. (1993). Purification, characterization, and cDNA cloning of an AU-rich element RNA-binding protein, AUF1. Mol Cell Biol *13*, 7652-7665.

5 RNA regulation by 273 RNA binding proteins

5.1 Background

Within eukaryotic cells, similar to histone proteins on DNA, RNAs are coated by RNA binding proteins (RBPs). The RBPs on RNA transcripts act to dictate regulatory decisions for the RNA throughout all steps of RNA processing, and ultimately controlling the composition and fate of the RNA within the cell. Currently there is a minimal level of understanding that relates to which RBPs affect what RNAs, and how they perform their precise role, either individually or in a coordinated fashion.

In chapter 4 we compiled a list of 1072 known RNA binding proteins and have potentially expanded this list by several dozens more. While a small fraction of these proteins have been mostly studied in isolation, or along with related family members (Huelga et al., 2012; Lagier-Tourenne et al., 2010), a more comprehensive analysis of hundreds of different human RNA binding proteins has yet to be conducted. In chapter 3 we assessed the roles of 6 hnRNP proteins by depleting each protein using siRNAs and analyzing the direct and indirect splicing and gene expression changes (Huelga et al., 2012). While similar studies have been performed for larger sets of proteins, their RNA regulation analyses were limited to a selected set of genes for qPCR. In 2013, a study was published that depleted 81 major RNA binding proteins and assed the splicing changes for a small set of 47 exons in fibroblasts, IPSCs and IPSC-derived fibroblasts (Venables et al., 2013). Although this study explored several different cell types, they focused on only well-studied RNA binding proteins, and left out any uncharacterized proteins that contained RNA binding domains. When high-throughput depletion studies are conducted, it often becomes a challenge to validate all depletion experiments. In

many cases, the RNAi is assumed to work well enough and depletion levels are rarely fully validated.

Here we first conduct an unbiased screen for shRNAs that specifically deplete RNA binding proteins in human 293T cells. We are able to effectively deplete 273 previously examined as well as previously uncharacterized RNA binding proteins (RBPs) by at least two fold at the mRNA level. For each depletion experiment, we generated stranded RNA-sequencing libraries and confirmed depletion by RPKM analysis. Finally, using this data we assessed the extent of gene expression changes as well as RNA splicing changes that were dependent on the level of each RBP. Integration of the data will likely reveal interesting insights into the functions of these RBPs.

5.2 Results

5.2.1 Identification of shRNAs that specifically deplete ~300 RNA binding proteins

In chapter 5 a primary list of 483 RNA binding proteins was compiled, containing RBPs with 54 different RNA binding domains (Figure 5.1A). This list was intersected with The RNAi Conortium (TRC) shRNA library version 1.0. For the 403 RBPs that were available in this TRC version, 3 shRNAs each were selected for screening and organized into 96-well plates (Figure 5.1B). Each screening plate also contained 3 control wells, including a positive control shRNA (shRNA TRCN0000016038 against *TARDBP*, previously shown by our group to work in 293T cells under these conditions), an empty pLKO.1 vector as an shRNA control, and a PLKO.1 vector expressing GFP as a transfection control. Clones were picked, transformed in bacterial cells and subjected to DNA isolation, prior to transient transfection into human 293T cells using an

automated liquid handling system. Cells containing expressed shRNAs survived puromycin selection, were harvested and RNA was isolated and quantified.

Primers for each of the 409 RBPs contained in our screen were computationally designed and optimized for qPCR using the *Primer3* software. Primers were designed to span introns on two consecutive constitutive exons to exclude any DNA contamination in our qPCR experiments. In some cases, where optimal intron spanning primers were not possible, primers were designed within the 3'UTR. Primers were validated by qPCR in 293T cells, selecting primers with single-peak melt curves and reasonable Ct values.

For each of the shRNA treated plates, cDNA was generated by RT-PCR and qPCRs were performed for each well using the Fluidigm BioMark HD System (Figure 5.1B). This system allowed us to assay the expression of 96 genes simultaneously across 96 different samples in a single run. Fluidigm 96.96 Dynamic Array IFCs were run two at a time with the same set of 96 assays for each pair, including primers measuring the 31 RBPs examined on the first plate, the 31 RBPs examined on the second plate, 10 reference genes (*GAPDH*, *ACTB*, *SDHA*, *RPS13*, *CYCA*, *RPL13A*, *TBP*, *HMBS*, *YWHA2*, *RPL27*), and a core set of major RBPs (hnRNP proteins and SR proteins, see Methods Section).

The resulting data was processed using Fluidigm's Real-Time PCR Analysis software. After normalization by the average of all 10 reference genes, delta Ct values were exported for further processing. For a robust and complete calculation of fold change for each gene, we calculated two different delta delta Ct (ddCt) values. First a *control* ddCt value was computed versus the normalized gene expression in the single empty-PLKO.1 treated control well. Occasionally all gene measurements were not available in the single control well, so we also computed a *median* ddCt value versus the median normalized expression in every other sample of the qPCR run (excluding the

wells treated with shRNAs targeting the same gene). Using these ddCt values, two fold changes were computed with $2^{-\text{median ddCt}}$ and $2^{-\text{Control ddCt}}$. shRNAs were considered validated when the RBP targeted had a *median* or *control* fold change that was less than 0.5 (more than 2-fold downregulated) in the shRNA-treated sample. Using these metrics, 482 shRNAs (comprising 245 RBPs) were validated using the *control* fold change, and 462 shRNAs (comprising 255 RBPs) were validated using the *median* fold change. The two fold change metrics produced two sets of depleted RBPs that were largely overlapping, so we took the union of the results, which yielded 544 validated shRNAs that specifically depleted 273 RBPs (Figure 5.1C). Encouragingly, this represented an unbiased subset of RBPs, containing equivalent proportions of the known RNA binding domains as the original list of ~500 RBPs (Figure 5.1D). This is the largest screen to actually validate shRNAs for RBPs, identifying 544 shRNAs that depleted 273 RBPs and provides an excellent resource for the scientific community (See Table 5.1 for all validated shRNAs).

5.2.2 RNA-seq Analysis of depleted samples shows depletion at RPKM level

To assess the genome-wide changes dependent on our depleted RBPs we performed strand-specific RNA-sequencing experiments (RNA-seq). Of the 273 significantly depleted RBPs, 21 samples were excluded from our RNA-seq analysis, since the expression of the RBP in 293T cells was not strongly detected in previous 293T RNA-seq experiments (RPKM <1) (Huelga et al., 2012). For the 252 remaining RBPs, if more than one shRNA was validated, we chose the shRNA sample yielding the greatest depletion of the RBP for RNA-seq. For comparison, 11 control samples and a set of 13 shRNA replicates were also selected for RNA-seq. RNA-seq libraries were prepared in three 96-well plates using the Tru-seq stranded mRNA HT kit (Illumina), and sequenced at varying depths on the Illumina HiSeq 2000 yielding 500 thousand to 44

million merged reads per sample (Figure 5.2A). Reads were mapped to the human hg19 genome, RPKMs (reads per kilobase per millions of reads) were calculated for all genes, and $\log_2$ fold changes were computed comparing the RPKM in each sample to the average RPKM in the controls from the same RNA-seq plate to control for library batch effects (Figure 5.2B).

The computed RPKM-based fold changes for the RBPs specifically depleted in each sample were, on average, also more than 2 fold down (corresponding to a –1 $\log_2$(Fold Change)) in the RNA-seq datasets as expected from our qPCR cutoffs (Figure 5.2C). While 70% of the datasets showed more than a 2-fold depletion by RPKM, the remaining datasets also show depletion, although not at a stringent 2-fold level (red shading, Figure 5.2C). Eight of the 263 (3%) RBP depletion experiments appear to have the RBP upregulated, however these datasets are poor library preparations and were excluded as outlier datasets (>0, Figure 5.2C). The outlier datasets include *PARN* depletion, a poorly sequenced library, and RBM15 depletion, a poorly generated library, which is clear upon visual inspection of the read distributions across genes (Figure 5.2D).

To directly correlate the RPKM-based fold changes obtained from RNA-seq with the delta Ct fold changes obtained from qPCR, we took the log2 of the qPCR fold changes and plotted them against each other (Figure 5.2E). Considering that these values are obtained using entirely different methodologies, in general, the fold changes computed from RNA-seq correlate quite well to the fold changes computed from qPCR with a pearson R=0.41. This correlation is improved to R=0.47 when considering only those RBPs that meet the 2-fold cutoff in the qPCR and RNA-seq experiments.

Reassuringly, when we inspect all RNA-seq experiments for the expression levels of all 252 RBPs we see that the RBP targeted is specifically depleted in each

dataset, and is of among the most downregulated RBP in the experiment (blue diagonal, Figure 5.2F). Again, we observe that the same experiments previously predicted as outliers, again appear to be clear outliers (including *PARN* and *RBM15B* depletions, Figure 5.2D), having expression levels that seem very high, however these datasets are poor libraries and can be excluded (darker columns, Figure 5.2F). In addition to the downregulation of the targeted RBP, close to 500 other genes on average are also up or downregulated in each of the RBP depletion data sets (Figure 5.2G), suggesting that the regulation of these genes is (directly or indirectly) dependent on the depleted RBP.

5.2.3 Alternative splicing analysis

Next we used the RNA-seq data to examine the effects on splicing for each RBP depletion. To compute the percent spliced in (PSI, Ψ) for all previously known alternatively spliced cassette exons, we applied the OldSplice algorithm (Lovci et al., 2013). This algorithm enumerates the exon-junction spanning reads for the inclusion event (reads spanning exon1-exon2 and exon2-exon3 for inclusion of exon2), and the exclusion event (reads spanning exon1-exon3 for exclusion of exon2) and computes a ratio of included reads to total junction reads for PSI. PSI values are further filtered for splicing events with at least 1 inclusion read, at least 1 exclusion read, and at least 30 total junction reads (included + excluded reads). For each RBP we obtain filtered PSI values for thousands of cassette exons, displaying bimodal distributions with many PSI values skewed towards 0 (complete exon exclusion) or 1 (complete exon inclusion). As an example, the PSI distribution for two RBFOX2 replicates and control are shown in Figure 5.3A. To identify differential splicing regulation we computed a delta PSI value (ΔΨ) for each exon by taking the difference in PSI between the RBP experiment and the mean of the batch control experiments ($\Delta\Psi = \Psi_{RBP} - \Psi_{Control}$). To determine significantly changed exons (*p*-value <0.05) we compute a Z-score, first taking the mean (μ) and

standard deviation (σ) of each RBP delta PSI distribution (RBFOX2 ΔΨ distribution shown in Figure 5.3B). An exon was significantly included if the ΔΨ score was greater than the μ+2σ and the exon was significantly excluded if the ΔΨ score was less than the μ–2σ. For each RBP we identified an average of 300 splicing changes (Figure 5.3C).

It has been shown for 6 members of the hnRNP family of RBPs that they can act in a coordinated and redundant manner, affecting exon inclusion together (Huelga et al., 2012). To analyze this within our 252 RBP dataset, we counted for each exon the number of RBP datasets where the exon was significantly regulated. Surprisingly, we observed a large extent of coordinated regulation, identifying many exons that are differentially included in more than 60 and up to 130 of the RBP datasets (Figure 5.3D), suggesting that the coordination of hundreds of proteins can be important for the inclusion of a particular exon in a direct or indirect manner.

To examine the extent of direct regulation of differentially included exons, we performed a motif analysis, in particular, for the known RBFOX2 binding motif GCAUG. For a stringent set of GCAUG motifs that are highly conserved across mammals (Lovci et al., 2013), we overlapped the motifs within 400 bases up and downstream of all differentially included exons in all datasets and calculated the fraction of included and excluded exons containing the motif. Promisingly, we see that the RBFOX2 (TRCN0000074544) dataset is among the most enriched datasets for this motif near its significantly included and excluded exons. However, the replicate RBFOX2 (TRCN0000074543) experiment using a different shRNA does not show a similar enrichment (Figure 5.3E). We attribute this difference in enrichment to a difference in the level of depletion for each of these experiments. The RBFOX2 dataset without the GCAUG enrichment is only depleted 2.5 fold, while the RBFOX2 dataset with the enrichment of GCAUG is depleted 4.6 fold.

This analysis suggests that there is a dose-dependent effect on splicing, requiring a certain level of depletion to identify direct differential splicing regulation. This will have to be taken into careful consideration as we analyze the remaining RBP datasets. Motifs for all of the RBPs have yet to be identified, so the analysis was limited to RBFOX2. In 2013, Hughes and colleagues published a compendium of RNA binding motifs, some identified using a method called RNACompete for 129 unique RBPs, as well as previously identified motifs using other methods for 75 additional RBPs (Ray et al., 2013). The identified motifs overlap with 83 more of our 252 depleted RBPs, and it will be interesting to also carry out the same motif enrichment experiment for these other RBPs with known motifs.

5.2.4 Replicate shRNA comparison for identification of robust RBP dependent events

With only one knockdown performed per shRNA, biological or technical variability could introduce unwanted variance. Additionally, as shRNA achieves knockdown by targeting the RNAi machinery to a desired mRNA via a 21 nucleotide complementary sequence, off-target knockdown can occur if this targeted sequence (or partial matches to it) is not unique in the genome. To address the degree of such effects in our dataset, we sequenced shRNA replicates for 12 RBPs using two different shRNAs, and one using the same shRNA (TARDBP). In our analysis of splicing above, we observed that varying levels of depletion cause differences between two experiments. As with the RBFOX2 replicate differences, we observed poor correlation between all pairs of shRNA replicates at the gene expression (RPKM, Figure 5.4A) as well as splicing (PSI, Figure 5.4B) level. We are still exploring whether these incongruities indicate true differences between shRNAs, or simply reflect different levels of shRNA knockdown efficiency. We note that the most correlated pair of shRNAs (TRCN0000001298 and

TRCN0000001299, targeting HNRNPU) have depletion levels that are quite similar (3.3 fold and 3.0 fold), suggesting that inefficient knockdown may drive the observed lack of correlation. To further refute the potential for shRNA off target affects, we checked the 3'UTRs of all downregulated genes in all datasets for the occurrence of sequences complimentary to the shRNA used, and identify no significant enrichment (data not shown), suggesting that the shRNAs do not have any strong off target affects.

To more closely explore differences in the level of depletion, we have repeated depletions for RBFOX2 in triplicate, using three different validated shRNAs (TRCN0000074544 (sh_1), TRCN0000074546 (sh_2), TRCN0000074543 (sh_3)). We confirmed RNA depletion in all 9 samples by qPCR (Figure 5.4C) and we also ran a western blot to confirm protein level depletion compared to control shRNA treated cells (Figure 5.4D). We were able to achieve depletions of RBFOX2 at various levels. While sh_1 and sh_3 showed a strong depletion of RBFOX2 at the RNA and protein level, sh_2 showed a weaker depletion.

For all 9 samples and 3 controls we subjected the RNA to splicing sensitive microarray analysis to measure splicing changes. Merging the values obtained from all 9 RBFOX2 experiments compared to the merge of the 3 controls we identified 250 exons that were differentially regulated by separation score (SepScore) analysis (Sugnet et al., 2006). Taking these 250 RBFOX putative regulated exons, we made a heatmap of the splicing measurement for each of the individual 9 replicates and 3 controls. Surprisingly, we observed robust and highly correlated splicing changes using both sh_1 and sh_3, which yielded similarly strong depletion of RBFOX2 protein (Figure 5.4E). Interestingly, the SepScores observed upon sh_2 knockdown were highly correlated but with a depressed slope, suggesting that the lesser degree of RBFOX2 protein knockdown lead to a correspondingly smaller splicing change (Figure 5.4E).

This replicate RBFOX2 experiment confirms that our RBP analyses may have muted effects when the depletion is not strong enough, and thus many false negatives. Further RT-PCR and qPCR validations will be required to validate that the splicing changes and gene regulation changes that we have measured are indeed true positives regulated by the RBP. This experiment also brings up another very important factor that should be considered in future analysis. Our depletion validations were done at the mRNA level, however for very stable proteins, the protein levels may not be sufficiently depleted. This will also affect our results, and western blots of depletion samples are also currently underway to examine the extent of this within our RBP datasets.

5.3 Discussion

Here we have identified shRNAs that deplete close to 300 human RNA binding proteins in a human cell-line. Within these RBP depletion conditions, we have also generated RNA-seq datasets to assess the transcriptomic changes dependent on each RBP. After confirming depletions in the RNA-seq datasets correlate with qPCR measurements, we have been able to identify differentially affected genes and exons. We have conducted analyses to explore shRNA off-target affects and have noted that varying levels of RBP depletion has the potential to result in many false negatives. Continued analysis will be required to uncover and validate the true positives.

In addition to many experimental validations, this data is ripe for further integrative analysis. Simple clustering analyses have already been completed to identify functional groups of RBPs that regulate clusters of other RBPs similarly (black boxes, Figure 5.5A). This gene expression clustering was independent of batch effects or level of depletion (right bars, Figure 5.5A). We have also begun to integrate the splicing data to identify splicing networks, or genes that affect sets of exons similarly (Figure 5.5B). RBP members of these identified functional groups will be further analyzed with follow-

up experiments including RNA immunoprecipitation followed by sequencing (RIP-seq) to identify directly regulated gene targets and splicing sensitive microarrays to confirm splicing changes.

All the RBPs in this analysis have not previously been implicated in splicing, so this analysis also has the potential to identify novel splicing factors. *In vitro* splicing tethering assays are already underway to show the direct regulation of splicing by RNA binding proteins that seems to regulate a large set of exons and have not previously been associated with splicing. Additionally this large dataset can be used in conjunction with machine learning algorithms to predict splicing levels based on RBP expression in cells. Thus, given a list of gene expression values, we have the potential to predict the exon inclusion levels for all exons given a properly trained machine-learning algorithm. Taken together, this dataset has already revealed some interesting findings and provides many avenues for further exploration with the potential to reveal many interesting insights in RNA biology.

5.4 Methods

5.4.1 RBP-specific shRNA preparation, transfection, and RNA extraction

For 403 RNA binding proteins 3 shRNA clones from the RNAi Consortium (TRC) were picked and pipetted in 96-well formats. Clone picking was performed using a Beckman Coulter BioMek FXP automated liquid handling system contained within a class 100 biosafety enclosure. The clones were grown using 1.2mL 2X LB containing 100ug/ml carbenicillin in 2mL deep-well culture blocks, while shaking at 37 degrees Celsius, 350rpm for 24 hours. The resultant cultures were pelleted at 1800 x g for 10 minutes, and the broth were decanted from the pellets, which were then frozen at -80C. Plasmid DNA was harvested from the frozen bacterial pellets using the Sigma

GenElute HP 96 Well Plasmid Miniprep Kit (NA9604). Plasmid DNA was quantified using Invitrogen Quant-iT PicoGreen dsDNA reagent (P11496) and reading on a Perkin-Elmer Envision microplate reader. The DNA was then normalized to 30ng/ul in H2O using the BioMek FXP.

$2x10^{-4}$ human 293T cells in 200ul of CGM were plated into each well of thirteen 96-well plates. 24 hours later 150ng of the prepped shRNA plasmid was mixed with 0.2ul Polyplus jetPRIME transfection reagent diluted into JetPrime Buffer using the BioMek FXP and incubated at RT for 10 minutes. The DNA:JetPrime complexes were then transferred to the 293T cells using the BioMek FXP. Each 96-well plate contained a positive control shRNA in well F12 (shRNA TRCN0000016038 against *TARDBP*, shown previously to work in 293T cells under these conditions) and an empty pLKO.1 vector in well G12 as an shRNA control. Puromycin was introduced at 24hrs by aspirating 100ul of CGM from each well followed by overlay of 100ul CGM/puromycin (7ug/ml). At 48hrs post puromycin selection 100ul of CGM per well was aspirated followed by overlay of 100ul CGM/puromycin (3.5ug/ml). 24 hours later (72hrs post puromycin selection) the growth media was aspirated from the transfected cells using a handheld multichannel manifold connected to a vacuum line. The cells were then lysed and RNA isolated using Sigma GenElute 96 Well Total RNA Purification Kit (RTN9604). The RNA was quantified using Invitrogen's Quant-iT RNA Assay kit (Q33140) and normalized to 10ng/ul in H2O using the BioMek FXP.

5.4.2 RBP-specific primer design and testing

Primers were designed for each gene encoding an RBP using in-house scripts and *Primer3*. Primers of ~20bp were designed such that they spanned introns on two consecutive constitutive exons and resulted in a target amplicon size of approximately 100bp. In some cases, where intron-spanning primers were not possible, primers were

designed on the 3'UTR with the same length and amplicon requirements. Primer validation by qPCR melt curve analysis was carried out on the 7900HT Fast Real-time PCR System (Applied Biosystems) using 10ng of untreated human 293T cDNA with Fast SYBR-Green Master mix (Applied Biosystems). Primers with single-peak melt curves and Ct values <~30 were considered acceptable validated primers.

5.4.3 RT-PCR and Fluidigm Biomark HD qPCR analysis

For each of the thirteen 96-well shRNA treated plates, cDNA was generated by reverse transcribing 100ng of the obtained total RNA extracted using oligo (dT) primer and Superscript III reverse transcriptase (RT; Invitrogen) with half of the manufacturer's suggested amount of RT enzyme. Fast Gene Expression Analysis Using EvaGreen on the BioMark HD System (Fluidigm) was performed according to the Advanced Development Protocol 37 for the 96.96 Dynamic Array IFC. Specifically, we conducted specific target amplification (STA) reactions for 14 cycles, Exonuclease I treatment, and diluted the products 10-fold with TE Buffer (TEKnova). 96.96 Dynamic Array IFCs were run two at a time with the same set of 96 assays for each pair, including primers measuring the 31 RBPs examined on the first plate, the 31 RBPs examined on the second plate, 10 reference genes (*GAPDH*, *ACTB*, *SDHA*, *RPS13*, *CYCA*, *RPL13A*, *TBP*, *HMBS*, *YWHA2*, *RPL27*), and a core set of major RBPs (*HNRNPA0, HNRNPA1, HNRNPA2B1, HNRNPAB, HNRNPA3, HNRNPC, HNRNPCL1, HNRNPD, HNRNPDL, HNRNPF, HNRNPH1, HNRNPH2, HNRNPH3, HNRNPK, HNRNPL, HNRNPM, HNRNPR, HNRNPU, SRSF2, SRSF3, SRSF4, SRSF5, TAF15, UPF1, TARDBP*).

The resulting data was processed using Fluidigm's Real-Time PCR Analysis software, using "Auto Global" threshold settings and "Linear Derivative" baseline correction. For each run, all 10 reference genes were named "Ref" to allow for all measurements to be normalized by their average expression. After normalization, delta

Ct values were exported. In-house scripts were used to calculate two different delta delta Ct (ddCt) values. A “Control ddCt” value was computed versus the normalized expression for the gene in the empty-PLKO.1 treated well (G12). We also calculated a “median ddCt” value versus the median normalized expression for the gene in every other sample of the run. Using these ddCt values, two fold changes were computed with $2^{-\text{median ddCt}}$ and $2^{-\text{Control ddCt}}$. An shRNA was considered validated when the shRNA targeted RBP had a median or Control fold change for <0.5 (2-fold).

5.4.4 RNA-seq library preparations

Three 96-well plates of RNA-seq libraries were prepared with 200ng RNA from the significantly depleted RBP samples using the Tru-seq stranded mRNA HT kit (Illumina). Final libraries were quantified using the KAPA Library Quantification Kit (KAPA Biosystems) and the TapeStation (Agilent). For the first 96 RNA-seq libraries (plate 1), the libraries were normalized to 20nM based on the KAPA quantification results, pooled equally as 12-plex, and diluted to 2nM. 9pM was loaded onto the HiSeq 2000 (Illumina) for single end (SE), 100bp, dual-indexed reads. Next, the same samples (plate 1) were re-pooled as 6-plex based on the ratio of reads from the first sequencing run, diluted, and re-sequenced on the Illumina HiSeq 2000 (Illumina) as SE, 100bp, dual-indexed reads. The second set of RNA-seq libraries (plate 2) were pooled equally as 6-plex and 12-plex pools and each sequenced on the HiSeq 2000. The third set of RNA-seq libraries (plate 3) were equally pooled as 96-plex, diluted to 2nM, and 6pM was loaded onto the MiSeq (Illumina) for paired-end sequencing. Libraries were re-pooled as 6-plex according to ratios of reads from MiSeq and sequenced on the HiSeq 2000.

5.4.5 RNA-seq data processing and gene expression analysis

Datasets were demultiplexed with the assigned barcodes and mapped using the RNA-STAR algorithm (Wu and Nacu, 2010), with parameters to discard multi-mapping reads and allowing for known exon-junction reads. Resequenced technical replicate experiment reads were merged and RPKM (reads per kilobase per millions of reads) expression values were calculated for all Gencode v17 genes. Fold changes were computed comparing the RPKM in the dataset to the mean RPKM for the batch-specific control experiments (control experiments on the same RNA-seq plate 1, 2 or 3).

5.4.6 RNA-seq splicing analysis

The OldSplice algorithm, previously published in (Lovci et al., 2013), was applied to the processed RNA-seq data to obtain percent-spliced-in (PSI values; Ψ). Exon Ψ values were filtered for exons with at least 30 total reads mapping to the junctions, and at least one read on the included and excluded junction. Ψ values were compared to the mean Ψ value for all PLKO control experiments in the same batch to obtain a delta Ψ (ΔΨ) value for all exons. For each dataset the mean and standard deviation of the delta Ψ distribution was computed. Significantly enriched changing exon targets were identified with a Z-score cutoff (p-value<0.05), or those targets that had a delta PSI less than $\mu-2\sigma$ (excluded) or greater than $\mu+2\sigma$ (included).

5.4.7 Splicing-sensitive microarray preparation

Stable knockdown of RNA binding proteins was performed by transfection of 293XT cells with the desired pLKO shRNA plasmid as well as lentiviral packaging plasmids (pMD26 and psPAX2). Media was changed after 24 hours, and virus-containing media was harvested 48 hours post-transfection and 0.45 um filtered. 293XT cells were infected with filtered virus for one day, and selected with 1 ug/ml Puromycin

beginning 48 hours post-infection. After 10 days, cells were harvested for RNA and protein. RNA was isolated using Trizol (Life Technologies) and DNAse treated using Turbo DNAse (Ambion). Samples for microarray analysis were prepared using the WT Expression Kit (Life Technologies), starting with 400 ng of RNA for each sample, and labeled with the GeneChip WT Terminal Labeling Kit (Affymetrix). Samples were then assayed using Affymetrix Human Transcriptome Array 2.0 microarrays, performed at the VA/VMRF Microarray & NGS Core (San Diego, CA). Each shRNA was assayed in biological triplicate, and was performed in the same batch as a paired non-target shRNA control (also in biological triplicate).

5.4.8 Splicing-sensitive microarray analysis

Initial normalization was performed using the RMA algorithm in the Expression Console Software (Affymetrix), and outlier microarrays were discarded from further analysis. To identify splicing events, annotation files for probes present on the Human Transcriptome Array 2.0 microarray were processed to identify all skipped exon, retained intron, and alternative 5' and 3' splice site events that had probesets covering both inclusion and exclusion present on the microarray. For each batch of experiments, probes were discarded if they were not identified as "present" (*p-value* ≤ 0.05) in at least half of the microarrays processed in each batch, and events in genes with less than 5 total probesets were not considered.

To identify significantly altered alternative splicing events, triplicate microarrays for an shRNA were compared against triplicate controls. Each alternative splicing event *E* was then queried using the following approach:

1. For each microarray *M*, a robust regression line was calculated for all probesets covering the entirety of gene *G* containing event *E*, regressed against the mean signal for that probeset across all 6 microarrays using the

robustfit algorithm (Matlab). This regression line was used to normalize for altered gene expression.

2. To test for altered event inclusion, residuals were calculated for those probesets querying inclusion (either covering the SE, retained intron, or 3' or 5' extended region itself, or those covering the exon:exon splice junctions created upon inclusion) by subtracting the observed probe signal with the expected signal based on the robust regression line. Significance was determined by ANOVA test on the residuals from control compared to shRNA samples.
3. Altered event exclusion was similarly assayed using probesets querying exclusion (those covering the exon:exon splice junctions created upon event exclusion).
4. To summarize the degree of altered splicing of each event in one value, a SepScore was calculated by subtracting the mean inclusion residual from the mean exclusion residual, as previously described (Sugnet et al., 2006).

5.5 RBP publication acknowledgement

Chapter 5 is in preparation to be submitted for publication in *Nature*. This dissertation author was the primary researcher/author who designed and directed the experiments, analyzed the data, and wrote the text. Co-authors who have also contributed to this work include Eric Van Nostrand (Yeo Lab, UC San Diego), Patrick Liu (Yeo Lab, UC San Diego), and Shawn Hoon (A*STAR, Singapore).

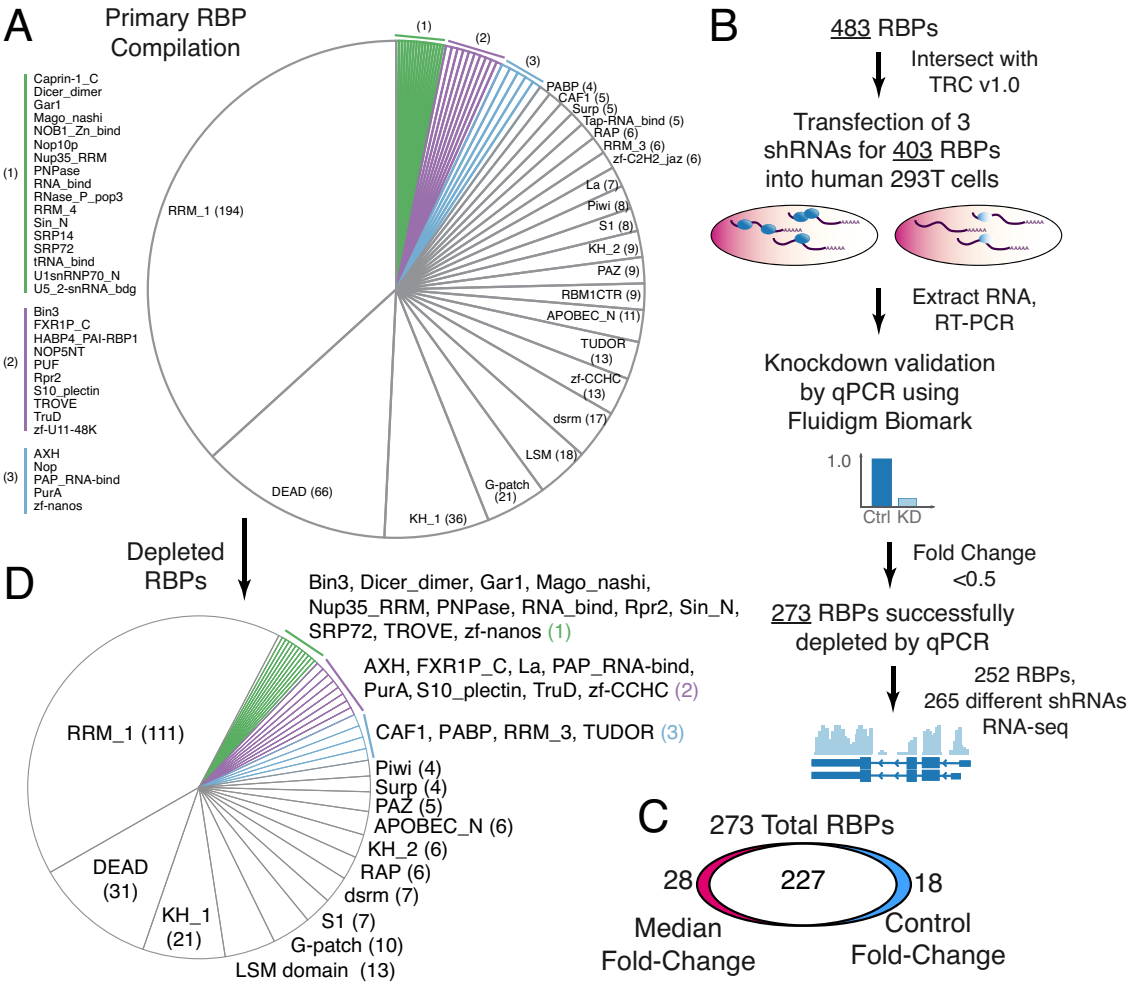


Figure 5.1: Systematic RNA binding protein depletion via shRNAs

A. The pie chart shows the distribution of RNA binding domains represented in our list of 483 RNA binging proteins. Domains that are represented 3 times, 2 times and 1 time are colored green, purple, and blue and are named on the left.
B. Experimental flow chart beginning with a primary list of 483 RBPs and ending with RNA-seq libraries for 252 different RBPs and 265 different shRNAs.
C. Overlap of RBPs that were successfully depleted based on the empty-vector *control* computed fold change and the *median* computed fold change.
D. The pie chart shows the distribution of RNA binding domains represented in our list of 273 successfully depleted RNA binging proteins. Domains that are represented 3 times, 2 times and 1 time are colored green, purple, and blue and are named on the left.

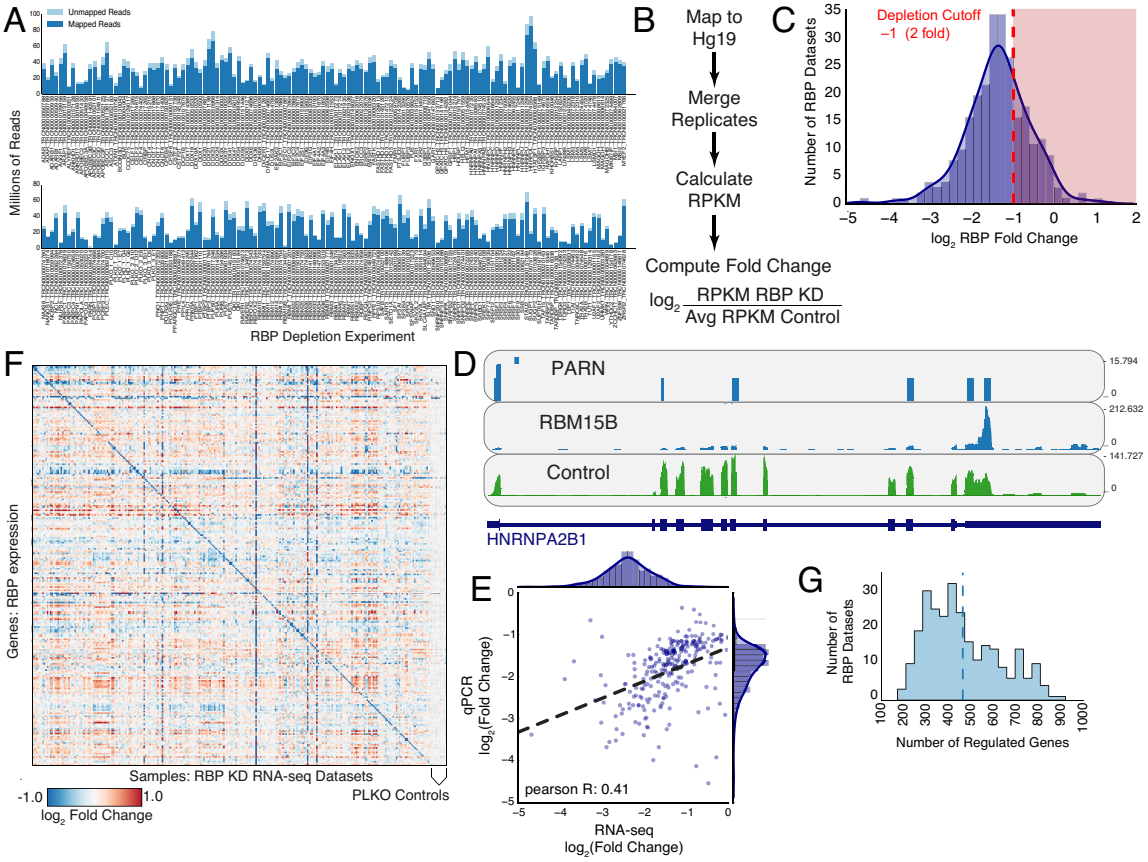


Figure 5.2: RNA-seq data confirms specific RBP depletion

A. Number of reads sequenced and mapped for each sample combining data from 1 to 3 HiSeq runs.
B. Flow chart of RNA-seq data analysis. For each dataset, we map the reads to the hg19 genome, merge the replicate sequencing experiments, calculate RPKM (reads per kilobase per million mapped reads) gene expression values, and compute fold changes compared to an average of the controls in each batch.
C. Histogram of $\log_2$ fold-changes for the depleted RBP in the corresponding RBP depletion sample. Red dashed line marks the 2-fold cutoff used to identify significantly depleted RBPs. RBPs falling in the red shading are mostly downregulated, although they do not meet our stringent cutoff.
D. Display of reads mapped to the HNRNPA2B1 gene transcript for two outlier datasets PARN and RBM15B, as well as a control experiment (green).
E. Scatter plot of $\log_2$ fold changes by delta Ct analysis qPCR for the *median* computed fold change compared to RNA-seq RPKM-based fold changes (Pearson R correlation = 0.41).
F. Heatmap of $\log_2$ fold-changes for the RBP gene expression (y axis) in the corresponding RBP depletion RNAseq dataset (x axis). Each axis is sorted in the same order, creating an obvious diagonal of the depleted RBPs. Blue is down regulated, red is upregulated.
G. Histogram of the number of genes that are significantly up or down regulated (Z-score < 1.96, *p*-value < 0.05) in each RBP depletion experiment.

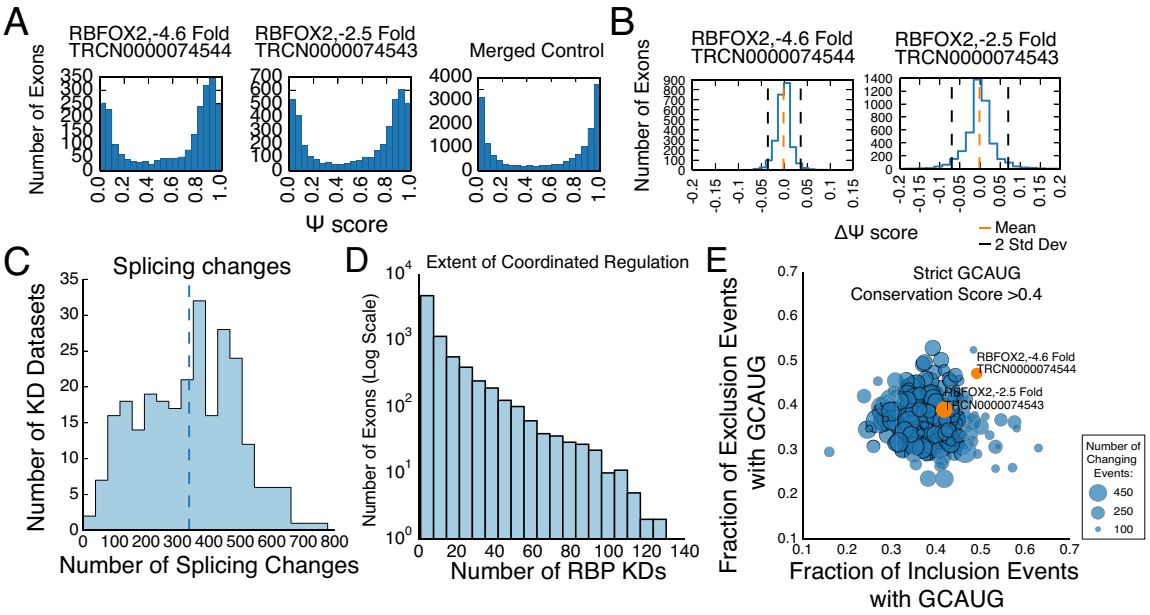


Figure 5.3: Splicing analysis identifies differentially regulated exons for each RBP depletion

A. PSI score histograms for all exons that pass stringent PSI filtering (at least 30 total junction reads, and at least one read on the included and excluded junction) in two RBFOX2 replicates and the merged control experiment.
B. Delta PSI distributions for RBFOX2 replicates compared to the merged control. Orange line marks the mean and the black lines mark +/– 2 standard deviations.
C. Histogram of number of splicing changes (included or excluded) in each RBP depletion experiment. The blue line marks the average number of changes.
D. Log scaled bar plot of the number of exons that are regulated in 0 to 140 experiments, displaying the extent of coordinated regulation. No exons are regulated in more than 140 RBP depletion experiments.
E. Scatter plot of the fraction of included vs excluded exons with a conserved GCAUG within 400 bases up and downstream. The size of each point reflects the number of regulated exons. Orange circles highlight the two RBFOX2 datasets that should show enrichment for the known GCAUG motif.

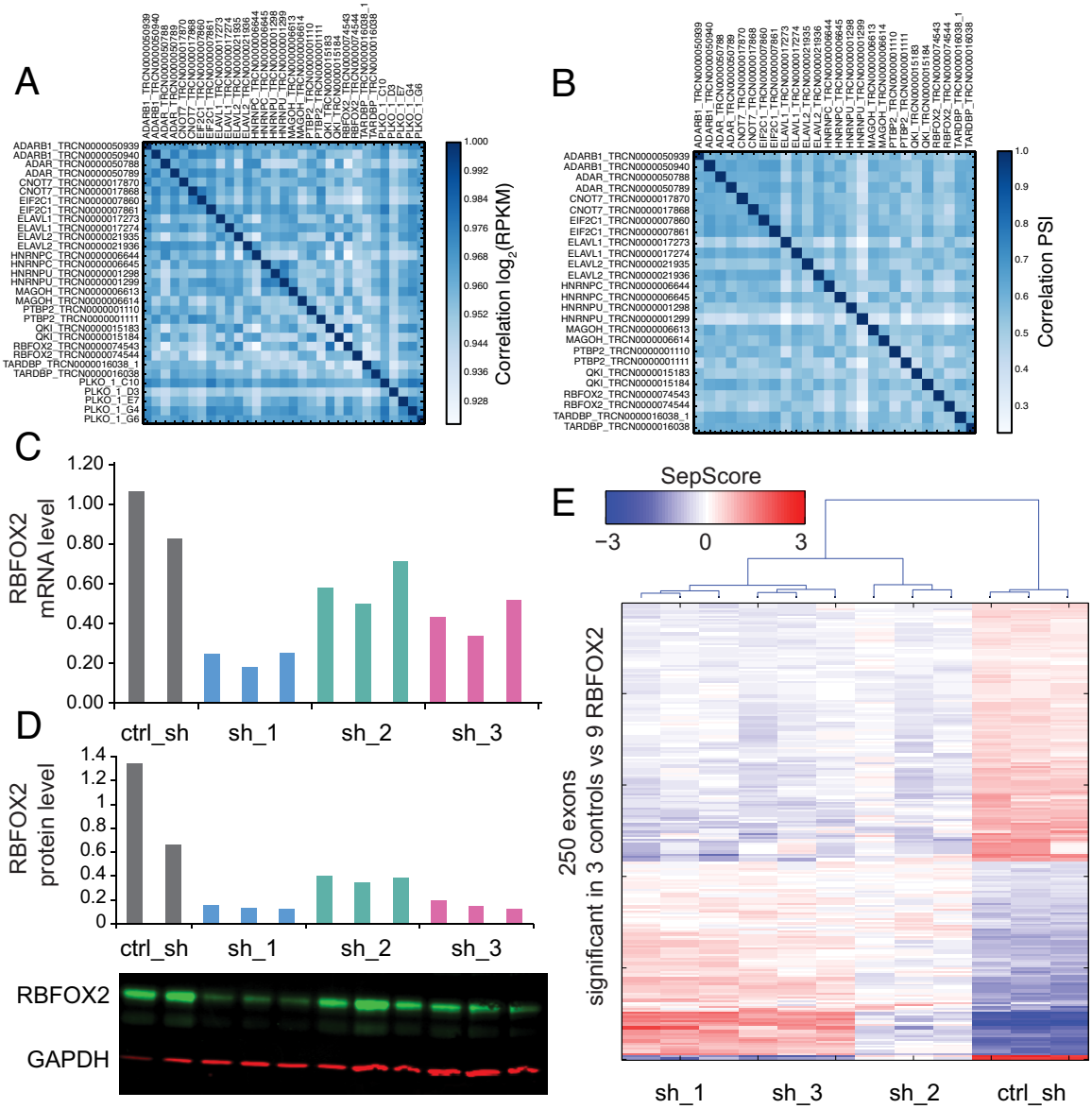


Figure 5.4: Concordance in gene expression and splicing changes between multiple shRNAs that target the same gene

Heatmap of correlation values computed for (**A**) the RPKM vectors, and (**B**) the PSI vectors for each RBP shRNA replicate experiment. In both cases, independent shRNAs targeting the same gene were not notably more correlated than shRNAs targeting different genes.

C. Relative mRNA expression levels measured by qPCR for RBFOX2 triplicate shRNA experiments and controls.

D. Western blot and corresponding quantification of protein levels for RBFOX2 triplicate shRNA experiments and controls. GAPDH was used as a loading control (red).

E. Heatmap of exon separation scores (SepScores) determined by microarray analysis for all 9 RBFOX2 triplicate shRNA experiments and controls. 250 exons are shown, which were determined to be significantly changing in a comparison of the 9 merged RBFOX2 samples versus the 3 controls.

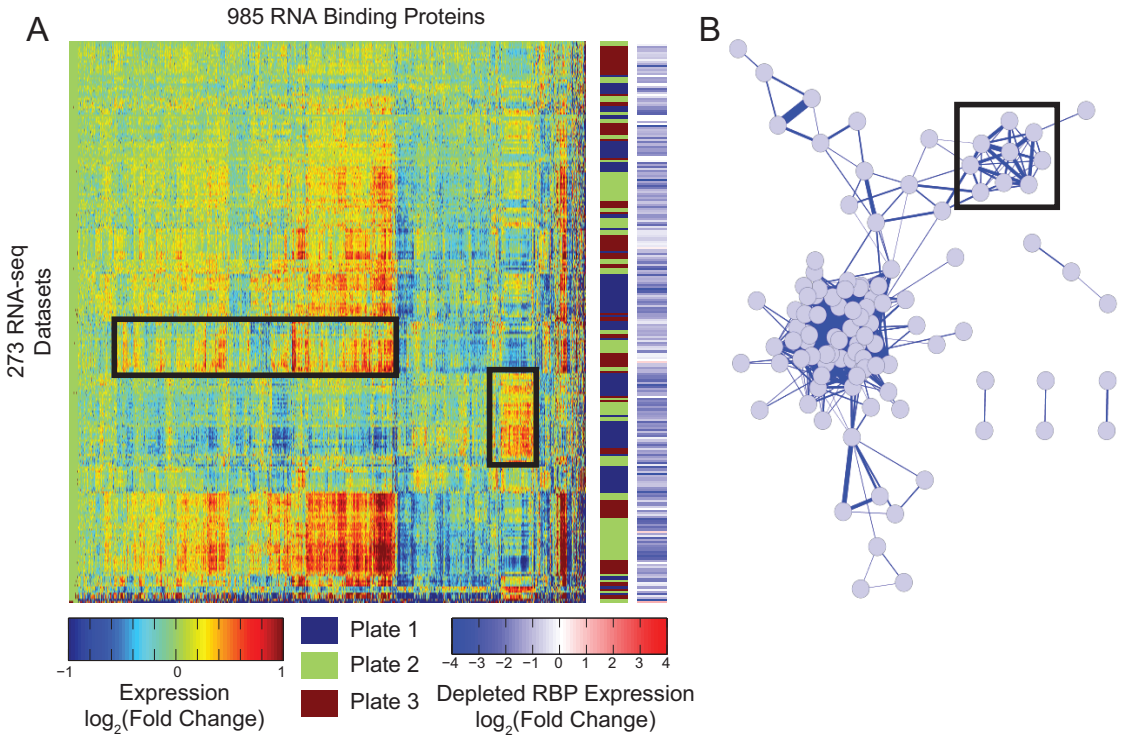


Figure 5.5: Integrative analysis of RBP RNA-seq data can reveal functional groups of RBPs

A. Clustered heatmap of expression values for 985 RBPs in all 273 RBP depletion RNA-seq datasets. Single heatmaps on the right display that the clustering is not dependent on batch affects (RNA-seq library plate) or level of RBP depletion. Black boxes highlight groups of genes that are regulated similarly across multiple datasets, or multiple datasets that show similar gene expression patterns.
B. Network diagram that represents the similarity or correlation of splicing measurements (PSI scores) in the different RBP datasets (nodes). The thickness of the line is representative of a stronger correlation. The black box highlights a group of RBP RNA-seq datasets that have very similar splicing measurements.

Table 5.1: Human 293T validated shRNAs for 273 RNA binding proteins

The RNAi Consortium (TRC) ID is given for all 544 validated shRNAs, along with the corresponding gene symbol for the targeted RBP. For many RBPs multiple shRNAs are listed. The final columns list the median and control fold changes measured by qPCR.

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
ADAD1	TRCN0000153428	0.442	1
ADAR	TRCN0000050788	0.156	0.067
ADAR	TRCN0000050789	0.468	0.203
ADAR	TRCN0000050790	0.834	0.361
ADARB1	TRCN0000050938	0.48	0.352
ADARB1	TRCN0000050939	0.267	0.195
ADARB1	TRCN0000050940	0.3	0.219
AGGF1	TRCN0000128297	0.444	0.373
AGGF1	TRCN0000128486	0.228	0.192
AGGF1	TRCN0000129956	0.183	0.154
AKAP1	TRCN0000013159	0.221	0.16
AKAP1	TRCN0000013160	0.676	0.49
ANKHD1	TRCN0000151413	0.532	0.47
ANKHD1	TRCN0000157128	0.369	0.326
ANKRD17	TRCN0000016948	0.45	0.457
ANKRD17	TRCN0000016950	0.404	0.411
APOBEC2	TRCN0000051433	0.485	0.462
APOBEC3A	TRCN0000049958	0.52	0.224
APOBEC3B	TRCN0000139463	0.19	0.227
APOBEC3B	TRCN0000140591	0.155	0.186
APOBEC3B	TRCN0000140731	0.209	0.25
APOBEC3C	TRCN0000052100	0.472	0.304
APOBEC3C	TRCN0000052101	0.163	0.105
APOBEC3D	TRCN0000154811	0.634	0.19
APOBEC3D	TRCN0000155111	0.713	0.213
APOBEC3D	TRCN0000156985	0.967	0.289
APOBEC3F	TRCN0000052123	0.419	0.132
APOBEC3F	TRCN0000052124	0.485	0.152
APOBEC3F	TRCN0000052125	0.113	0.036
APOBEC3G	TRCN0000052188	0.468	0.439
APOBEC3G	TRCN0000052190	0.324	0.303
APOBEC3H	TRCN0000051798	0.202	0.332
APOBEC3H	TRCN0000051799	0.037	0.06
ASCC1	TRCN0000136016	0.197	0.188
ASCC1	TRCN0000136137	0.143	0.137

Table 5.1 (continued): Human 293T validated shRNAs for 273 RNA binding proteins

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
ASCC1	TRCN0000136268	0.221	0.212
ASCC3	TRCN0000052183	0.355	0.313
ASCC3	TRCN0000052184	0.552	0.486
ASCC3	TRCN0000052185	0.357	0.315
ATXN1	TRCN0000003730	0.611	0.36
ATXN1	TRCN0000003731	0.642	0.379
BCDIN3D	TRCN0000129696	0.497	0.518
BCDIN3D	TRCN0000130043	0.41	0.427
BLM	TRCN0000004904	0.345	0.321
BLM	TRCN0000004905	0.505	0.469
CCDC75	TRCN0000127583	0.256	0.255
CCDC75	TRCN0000129017	0.204	0.203
CCDC75	TRCN0000130518	0.284	0.283
CELF1	TRCN0000017283	0.506	0.421
CELF1	TRCN0000017285	0.324	0.269
CELF2	TRCN0000074613	0.475	0.316
CELF2	TRCN0000074614	0.26	0.173
CELF2	TRCN0000074615	0.256	0.17
CELF3	TRCN0000005729	0.352	11.8
CELF4	TRCN0000074405	0.47	0.361
CIRBP	TRCN0000017263	0.419	0.432
CIRBP	TRCN0000017265	0.2	0.206
CNOT4	TRCN0000015215	0.182	0.181
CNOT7	TRCN0000017868	0.23	0.253
CNOT7	TRCN0000017869	0.336	0.37
CNOT7	TRCN0000017870	0.105	0.115
CNOT8	TRCN0000013408	0.208	0.245
CNOT8	TRCN0000013409	0.436	0.512
CNOT8	TRCN0000013410	0.417	0.49
CPEB4	TRCN0000150825	0.109	0.138
CPSF6	TRCN0000000151	0.484	0.516
CSTF2	TRCN0000153435	0.302	0.362
CSTF2	TRCN0000153738	0.156	0.187
CSTF2	TRCN0000157143	0.348	0.418
CSTF2T	TRCN0000141996	0.428	0.32
CSTF2T	TRCN0000145028	0.561	0.419
CSTF2T	TRCN0000145349	0.221	0.165

Table 5.1 (continued): Human 293T validated shRNAs for 273 RNA binding proteins

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
DCP1B	TRCN0000107675	0.337	0.368
DCP1B	TRCN0000107676	0.214	0.234
DDX1	TRCN0000050498	0.363	0.459
DDX1	TRCN0000050499	0.39	0.494
DDX1	TRCN0000050500	0.148	0.187
DDX17	TRCN0000074388	0.37	0.598
DDX17	TRCN0000074389	0.18	0.29
DDX18	TRCN0000050533	0.087	0.076
DDX18	TRCN0000050534	0.21	0.184
DDX18	TRCN0000050535	0.447	0.391
DDX19A	TRCN0000050218	0.493	0.491
DDX19A	TRCN0000050220	0.262	0.261
DDX20	TRCN0000007887	0.321	0.413
DDX21	TRCN0000051198	0.16	0.155
DDX21	TRCN0000051199	0.244	0.235
DDX21	TRCN0000051200	0.203	0.195
DDX23	TRCN0000050968	0.32	0.28
DDX24	TRCN0000050313	0.365	0.503
DDX24	TRCN0000050314	0.171	0.236
DDX25	TRCN0000155017	0.102	0.172
DDX25	TRCN0000157728	0.182	0.306
DDX3X	TRCN0000000002	0.478	0.401
DDX3X	TRCN0000000003	0.45	0.378
DDX3Y	TRCN0000000020	0.475	0.518
DDX3Y	TRCN0000000022	0.324	0.354
DDX4	TRCN0000051153	1.505	0.295
DDX4	TRCN0000051154	0.803	0.157
DDX4	TRCN0000051155	0.848	0.166
DDX41	TRCN0000001268	0.343	0.325
DDX47	TRCN0000005618	0.426	0.404
DDX5	TRCN0000001127	0.421	1
DDX53	TRCN0000050334	0.769	0.236
DDX53	TRCN0000050335	1.285	0.394
DDX54	TRCN0000049954	0.34	0.343
DDX58	TRCN0000151446	0.757	0.277
DDX58	TRCN0000152922	0.789	0.289
DDX58	TRCN0000153712	1.235	0.452

Table 5.1 (continued): Human 293T validated shRNAs for 273 RNA binding proteins

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
DDX6	TRCN0000074695	0.491	0.705
DDX6	TRCN0000074696	0.211	0.303
DDX60	TRCN0000050185	0.595	0.498
DHX36	TRCN0000050703	0.481	0.492
DHX8	TRCN0000000026	0.443	0.39
DICER1	TRCN0000051261	0.5	0.441
DNAJC17	TRCN0000074663	0.214	0.159
DNAJC17	TRCN0000074664	0.476	0.352
DNAJC17	TRCN0000074665	0.353	0.261
DUS2L	TRCN0000064789	0.283	0.245
EIF2AK2	TRCN0000001380	0.227	0.204
EIF2AK2	TRCN0000001381	0.311	0.28
EIF2C1	TRCN0000007860	0.369	0.296
EIF2C1	TRCN0000007861	0.418	0.335
EIF2C2	TRCN0000007864	0.41	0.344
EIF2C3	TRCN0000007869	0.187	0.167
EIF2C3	TRCN0000007870	0.245	0.218
EIF2C4	TRCN0000007873	0.273	0.243
EIF2S1	TRCN0000159909	0.164	0.151
EIF3B	TRCN0000154572	0.199	0.191
EIF3B	TRCN0000155603	0.153	0.147
EIF3B	TRCN0000155723	0.45	0.432
EIF3G	TRCN0000062033	0.414	0.717
EIF3G	TRCN0000062035	0.266	0.46
EIF4A1	TRCN0000052193	0.254	0.237
EIF4A1	TRCN0000052194	0.214	0.199
EIF4A2	TRCN0000051869	0.397	0.435
EIF4A2	TRCN0000051870	0.23	0.252
EIF4A3	TRCN0000061854	0.374	0.374
EIF4A3	TRCN0000061855	0.328	0.328
EIF4B	TRCN0000062598	0.169	0.2
EIF4B	TRCN0000062599	0.335	0.398
EIF4H	TRCN0000151553	0.251	0.229
EIF4H	TRCN0000153576	0.304	0.277
EIF4H	TRCN0000153719	0.412	0.376
ELAVL1	TRCN0000017273	0.332	0.28
ELAVL1	TRCN0000017274	0.307	0.259

Table 5.1 (continued): Human 293T validated shRNAs for 273 RNA binding proteins

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
ELAVL1	TRCN0000017275	0.492	0.415
ELAVL2	TRCN0000021934	0.504	0.499
ELAVL2	TRCN0000021935	0.414	0.41
ELAVL2	TRCN0000021936	0.321	0.318
ENOX1	TRCN0000130565	0.379	0.337
ENOX1	TRCN0000130610	0.332	0.295
ENOX1	TRCN0000148027	0.096	0.086
ENOX2	TRCN0000062018	0.274	0.235
ENOX2	TRCN0000062020	0.335	0.288
ERCC3	TRCN0000022079	0.378	0.394
ERCC3	TRCN0000022081	0.482	0.503
ESRP2	TRCN0000127635	0.207	0.168
ESRP2	TRCN0000130676	0.146	0.119
EWSR1	TRCN0000000034	0.379	0.338
EWSR1	TRCN0000000035	0.483	0.43
EWSR1	TRCN0000000036	0.247	0.22
FASTK	TRCN0000006320	0.216	0.2
FASTKD1	TRCN0000128611	0.536	0.448
FASTKD1	TRCN0000128653	0.419	0.35
FASTKD1	TRCN0000129813	0.383	0.32
FASTKD2	TRCN0000140593	0.321	0.31
FASTKD2	TRCN0000140752	0.372	0.359
FASTKD2	TRCN0000145126	0.015	0.015
FASTKD3	TRCN0000147212	0.439	0.438
FASTKD3	TRCN0000149056	0.495	0.493
FASTKD5	TRCN0000123224	0.176	0.179
FASTKD5	TRCN0000123225	0.23	0.235
FASTKD5	TRCN0000123226	0.204	0.208
FMR1	TRCN0000059758	0.353	0.392
FTSJD2	TRCN0000141341	0.497	0.494
FTSJD2	TRCN0000142265	0.447	0.444
FUBP1	TRCN0000013295	0.461	0.514
FUBP3	TRCN0000020574	2.561	0.343
FUBP3	TRCN0000020575	1.257	0.168
FUBP3	TRCN0000020576	0.41	0.055
FUS	TRCN0000010450	0.722	0.406
FUS	TRCN0000039824	0.333	0.187

Table 5.1 (continued): Human 293T validated shRNAs for 273 RNA binding proteins

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
FUS	TRCN0000039825	0.67	0.377
FXR1	TRCN0000158932	1.153	0.399
FXR1	TRCN0000159153	1.04	0.359
FXR2	TRCN0000013453	0.302	0.06
FXR2	TRCN0000013454	0.259	0.051
FXR2	TRCN0000013455	0.209	0.041
G3BP1	TRCN0000008719	0.443	0.159
G3BP1	TRCN0000008720	0.564	0.202
G3BP1	TRCN0000008721	0.981	0.352
G3BP2	TRCN0000047548	0.685	0.26
G3BP2	TRCN0000047549	1	0.38
G3BP2	TRCN0000047550	0.264	0.1
GAR1	TRCN0000147162	0.407	0.382
GAR1	TRCN0000147663	0.095	0.089
GAR1	TRCN0000147983	0.202	0.19
GPATCH2	TRCN0000133713	0.471	0.799
GPATCH3	TRCN0000122671	0.426	0.477
GPATCH3	TRCN0000141198	0.414	0.463
GPATCH4	TRCN0000134282	0.31	0.966
GPKOW	TRCN0000130596	0.326	0.942
GPKOW	TRCN0000148427	0.498	1.442
GRSF1	TRCN0000075044	0.358	0.802
HBP1	TRCN0000015273	0.377	0.78
HBP1	TRCN0000015274	0.22	0.455
HBP1	TRCN0000015275	0.451	0.935
HDLBP	TRCN0000150941	0.423	0.252
HDLBP	TRCN0000151220	0.408	0.243
HDLBP	TRCN0000153797	0.127	0.076
HELQ	TRCN0000051558	0.645	0.203
HELQ	TRCN0000051559	0.325	0.102
HFM1	TRCN0000154456	0.13	1
HNRNPA0	TRCN0000017233	0.244	0.024
HNRNPA0	TRCN0000017234	0.606	0.06
HNRNPA0	TRCN0000017235	0.222	0.022
HNRNPA1	TRCN0000006582	0.897	0.441
HNRNPA1	TRCN0000006583	0.42	0.206
HNRNPA1	TRCN0000006584	0.508	0.249

Table 5.1 (continued): Human 293T validated shRNAs for 273 RNA binding proteins

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
HNRNPA3	TRCN0000074508	0.32	0.451
HNRNPA3	TRCN0000074509	0.428	0.604
HNRNPA3	TRCN0000074510	0.176	0.248
HNRNPAB	TRCN0000074520	0.293	0.36
HNRNPC	TRCN0000006644	0.063	0.022
HNRNPC	TRCN0000006645	0.217	0.076
HNRNPC	TRCN0000006646	0.357	0.125
HNRNPD	TRCN0000001293	0.294	0.407
HNRNPF	TRCN0000001234	0.395	0.259
HNRNPH2	TRCN0000075093	0.326	0.298
HNRNPH2	TRCN0000075094	0.155	0.142
HNRNPH2	TRCN0000075095	0.173	0.158
HNRNPH3	TRCN0000074618	0.572	0.223
HNRNPH3	TRCN0000074619	1.083	0.421
HNRNPH3	TRCN0000074620	0.156	0.061
HNRNPK	TRCN0000062453	0.781	0.396
HNRNPK	TRCN0000062454	0.607	0.308
HNRNPK	TRCN0000062455	0.448	0.227
HNRNPL	TRCN0000017244	1.706	0.434
HNRNPL	TRCN0000017245	0.78	0.198
HNRNPL	TRCN0000017246	1.019	0.259
HNRNPU	TRCN0000001298	0.297	0.326
HNRNPU	TRCN0000001299	0.396	0.434
HNRPDL	TRCN0000074554	0.417	0.516
HTATSF1	TRCN0000006587	0.388	0.323
HTATSF1	TRCN0000006588	0.338	0.281
HTATSF1	TRCN0000006589	0.557	0.464
IFIH1	TRCN0000050849	0.388	0.331
IGF2BP1	TRCN0000075149	0.463	0.396
IGF2BP3	TRCN0000074675	0.358	0.52
IMPDH1	TRCN0000026530	0.316	0.373
IMPDH1	TRCN0000026570	0.429	0.506
KHDRBS1	TRCN0000000045	0.512	0.402
KHDRBS1	TRCN0000000046	0.311	0.244
KHSRP	TRCN0000013253	0.364	0.263
KHSRP	TRCN0000013254	0.246	0.178
KHSRP	TRCN0000013255	0.44	0.318

Table 5.1 (continued): Human 293T validated shRNAs for 273 RNA binding proteins

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
LARP1	TRCN0000152624	0.3	0.425
LARP1	TRCN0000152891	0.437	0.617
LIN28B	TRCN0000122599	0.385	0.454
LSM1	TRCN0000040228	0.702	0.256
LSM1	TRCN0000040229	0.998	0.364
LSM1	TRCN0000040230	0.696	0.254
LSM3	TRCN0000072318	0.263	0.211
LSM3	TRCN0000072320	0.36	0.289
LSM4	TRCN0000072309	0.478	0.417
LSM4	TRCN0000072310	0.382	0.334
LSM5	TRCN0000074453	0.296	0.269
LSM5	TRCN0000074454	0.043	0.039
LSM5	TRCN0000074455	0.396	0.36
LSM6	TRCN0000074719	0.382	0.434
LSM6	TRCN0000074720	0.307	0.348
LSM7	TRCN0000074743	0.552	0.475
LSM7	TRCN0000074746	0.489	0.421
LSMD1	TRCN0000074478	0.28	0.258
LSMD1	TRCN0000074480	0.397	0.366
MAGOH	TRCN0000006613	0.18	0.247
MAGOH	TRCN0000006614	0.078	0.108
MAGOHB	TRCN0000072324	0.389	0.42
MAGOHB	TRCN0000072325	0.314	0.339
MKI67IP	TRCN0000074499	0.379	0.367
MKI67IP	TRCN0000074500	0.512	0.495
MSI1	TRCN0000064005	0.422	1
MSI1	TRCN0000064007	0.379	1
MSI2	TRCN0000062808	0.315	0.256
MSI2	TRCN0000062809	0.144	0.117
MSI2	TRCN0000062811	0.212	0.172
MYEF2	TRCN0000021764	0.421	0.592
MYEF2	TRCN0000021766	0.272	0.382
NAA38	TRCN0000074760	0.334	0.347
NANOS1	TRCN0000118074	0.483	0.513
NCBP2	TRCN0000059993	0.467	0.469
NCBP2	TRCN0000059994	0.448	0.451
NCL	TRCN0000062283	0.398	0.372

Table 5.1 (continued): Human 293T validated shRNAs for 273 RNA binding proteins

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
NCL	TRCN0000062284	0.242	0.226
NUP35	TRCN0000072378	0.434	0.513
NUP35	TRCN0000072379	0.495	0.585
NXF2	TRCN0000059294	0.497	0.704
NXF3	TRCN0000060324	0.182	0.184
NXF3	TRCN0000060325	0.096	0.097
PABPC1	TRCN0000074639	0.462	0.398
PABPC1	TRCN0000074640	0.491	0.424
PABPC1L2A	TRCN0000146321	0.315	0.256
PABPC1L2A	TRCN0000148970	0.605	0.492
PABPC3	TRCN0000159750	0.405	0.374
PABPC4	TRCN0000074658	0.43	0.393
PABPC5	TRCN0000147141	0.049	0.127
PABPN1	TRCN0000000120	0.375	0.387
PABPN1	TRCN0000000122	0.423	0.436
PAPOLA	TRCN0000053020	0.39	0.397
PAPOLG	TRCN0000053220	0.452	0.571
PARN	TRCN0000049743	0.288	0.28
PARN	TRCN0000049744	0.255	0.248
PCBP1	TRCN0000074668	0.315	0.272
PCBP1	TRCN0000074669	0.354	0.306
PCBP1	TRCN0000074670	0.381	0.329
PCBP2	TRCN0000074684	0.615	0.5
PCBP3	TRCN0000074698	0.479	0.61
PCBP3	TRCN0000074700	0.085	0.109
PIWIL2	TRCN0000007881	0.147	0.132
PIWIL2	TRCN0000007883	0.521	0.467
PIWIL4	TRCN0000140185	0.404	0.476
PIWIL4	TRCN0000143890	0.45	0.529
PLEC	TRCN0000082813	0.365	0.25
PLEC	TRCN0000082815	0.16	0.109
PNLDC1	TRCN0000049769	0.339	1.047
PNLDC1	TRCN0000049770	0.306	0.947
PNO1	TRCN0000072353	0.304	0.291
PNO1	TRCN0000072354	0.448	0.429
PNPT1	TRCN0000035905	0.376	0.34
PNPT1	TRCN0000035906	0.548	0.496

Table 5.1 (continued): Human 293T validated shRNAs for 273 RNA binding proteins

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
POLR2G	TRCN0000021854	0.434	0.37
POLR2G	TRCN0000021855	0.419	0.357
POLR3E	TRCN0000053188	0.391	0.352
PPARGC1B	TRCN0000008597	0.419	0.41
PPIE	TRCN0000049371	0.467	0.515
PPIL4	TRCN0000049413	0.383	0.344
PPIL4	TRCN0000049415	0.51	0.459
PPRC1	TRCN0000063294	0.481	0.48
PPRC1	TRCN0000063295	0.438	0.436
PRKRA	TRCN0000052644	0.367	0.313
PRKRA	TRCN0000052646	0.463	0.395
PTBP1	TRCN0000001062	0.451	0.335
PTBP1	TRCN0000010584	0.573	0.426
PTBP2	TRCN0000001110	0.139	0.143
PTBP2	TRCN0000001111	0.183	0.188
PUF60	TRCN0000017248	0.309	0.408
PUF60	TRCN0000017249	0.324	0.428
PURA	TRCN0000014534	0.371	0.499
PURB	TRCN0000154286	0.066	0.084
PUS7	TRCN0000061933	0.321	0.299
PUS7	TRCN0000061934	0.272	0.253
PUS7L	TRCN0000145228	0.314	0.303
QKI	TRCN0000015183	0.378	0.337
QKI	TRCN0000015184	0.28	0.249
RAVER1	TRCN0000074623	0.205	0.156
RAVER1	TRCN0000074624	0.243	0.184
RAVER1	TRCN0000074625	0.306	0.232
RAVER2	TRCN0000134973	0.26	0.22
RAVER2	TRCN0000136631	0.35	0.296
RAVER2	TRCN0000136677	0.363	0.307
RBFOX2	TRCN0000074543	0.302	0.325
RBFOX2	TRCN0000074544	0.118	0.126
RBM10	TRCN0000074733	0.195	0.243
RBM11	TRCN0000159793	0.375	0.229
RBM11	TRCN0000159794	0.222	0.136
RBM11	TRCN0000160396	0.247	0.151
RBM12	TRCN0000074890	0.377	0.354

Table 5.1 (continued): Human 293T validated shRNAs for 273 RNA binding proteins

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
RBM14	TRCN0000072695	0.241	0.324
RBM15	TRCN0000074704	0.148	0.181
RBM15	TRCN0000074705	0.337	0.413
RBM15B	TRCN0000134657	0.24	0.385
RBM15B	TRCN0000135036	0.395	0.634
RBM15B	TRCN0000138870	0.449	0.722
RBM17	TRCN0000001200	0.195	0.154
RBM17	TRCN0000001201	0.156	0.123
RBM17	TRCN0000001202	0.404	0.318
RBM22	TRCN0000008650	0.148	0.172
RBM23	TRCN0000148528	0.439	0.489
RBM23	TRCN0000149951	0.435	0.484
RBM23	TRCN0000150168	0.3	0.334
RBM24	TRCN0000116198	0.092	0.082
RBM25	TRCN0000074750	0.078	0.086
RBM26	TRCN0000074523	0.486	0.53
RBM28	TRCN0000140479	0.419	0.468
RBM28	TRCN0000141802	0.194	0.216
RBM28	TRCN0000142682	0.464	0.518
RBM3	TRCN0000001248	0.115	0.135
RBM3	TRCN0000001249	0.252	0.297
RBM3	TRCN0000001250	0.426	0.502
RBM33	TRCN0000149980	0.463	0.536
RBM34	TRCN0000147509	0.456	0.543
RBM38	TRCN0000150789	0.242	0.097
RBM38	TRCN0000152908	0.153	0.061
RBM38	TRCN0000153519	0.057	0.023
RBM39	TRCN0000021771	0.133	0.142
RBM41	TRCN0000133976	0.888	0.477
RBM41	TRCN0000134374	0.505	0.271
RBM45	TRCN0000118592	0.45	0.395
RBM46	TRCN0000135051	0.153	0.148
RBM47	TRCN0000074563	0.322	0.324
RBM7	TRCN0000074529	0.299	0.271
RBM7	TRCN0000074530	0.257	0.232
RBMS1	TRCN0000074933	0.436	0.486
RBMS3	TRCN0000151205	0.379	0.277

Table 5.1 (continued): Human 293T validated shRNAs for 273 RNA binding proteins

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
RBMS3	TRCN0000152525	0.361	0.265
RBMS3	TRCN0000153312	0.568	0.416
RBMX2	TRCN0000074584	0.448	0.388
RBMX2	TRCN0000074586	0.547	0.473
RDBP	TRCN0000074958	0.393	0.265
RDBP	TRCN0000074959	0.661	0.446
RDBP	TRCN0000074960	0.564	0.38
RECQL	TRCN0000052000	0.481	0.534
RECQL	TRCN0000052001	0.475	0.528
RECQL4	TRCN0000051169	0.392	0.413
RNPC3	TRCN0000148266	0.206	0.194
ROD1	TRCN0000061784	0.497	0.533
RPP21	TRCN0000049719	0.381	0.498
RPP21	TRCN0000049721	0.265	0.346
RPS10	TRCN0000074808	0.382	0.39
RPS10	TRCN0000074810	0.238	0.243
RPS3	TRCN0000005165	0.078	0.082
SAFB	TRCN0000022099	0.187	0.259
SAFB	TRCN0000022100	0.151	0.208
SAFB	TRCN0000022101	0.278	0.384
SART3	TRCN0000128080	0.377	0.506
SART3	TRCN0000148744	0.299	0.401
SART3	TRCN0000148806	0.291	0.39
SETD1A	TRCN0000150815	0.441	0.532
SETD1A	TRCN0000152242	0.335	0.404
SF1	TRCN0000001069	0.25	0.248
SF1	TRCN0000001070	0.255	0.253
SF3A1	TRCN0000006597	0.221	0.253
SF3B4	TRCN0000000039	0.125	0.192
SF3B4	TRCN0000000040	0.235	0.363
SF3B4	TRCN0000000041	0.338	0.52
SFPQ	TRCN0000001083	0.286	0.407
SFSWAP	TRCN0000017223	0.449	0.557
SKIV2L	TRCN0000051813	0.554	0.492
SKIV2L	TRCN0000051814	0.314	0.279
SKIV2L	TRCN0000051815	0.317	0.282
SKIV2L2	TRCN0000051973	0.396	0.354

Table 5.1 (continued): Human 293T validated shRNAs for 273 RNA binding proteins

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
SKIV2L2	TRCN0000051975	0.504	0.449
SLC4A1AP	TRCN0000044093	0.409	0.553
SLC4A1AP	TRCN0000044094	0.175	0.237
SLC4A1AP	TRCN0000044095	0.251	0.34
SLTM	TRCN0000135106	0.288	0.32
SLTM	TRCN0000135690	0.487	0.541
SLTM	TRCN0000136959	0.315	0.35
SND1	TRCN0000049653	0.316	0.302
SND1	TRCN0000049654	0.514	0.491
SND1	TRCN0000049655	0.356	0.34
SNRNP200	TRCN0000051830	0.438	0.459
SNRNP35	TRCN0000157288	0.389	0.607
SNRPA	TRCN0000006609	0.332	0.409
SNRPB	TRCN0000072474	0.499	0.847
SNRPB	TRCN0000072475	0.396	0.672
SNRPB2	TRCN0000017253	0.294	0.352
SNRPB2	TRCN0000017254	0.334	0.399
SNRPB2	TRCN0000017255	0.097	0.116
SNRPD1	TRCN0000005381	0.13	0.135
SNRPD1	TRCN0000005382	0.382	0.398
SNRPD2	TRCN0000074398	0.294	0.301
SNRPD2	TRCN0000074400	0.396	0.405
SNRPD2	TRCN0000074401	0.217	0.222
SNRPE	TRCN0000072269	0.136	0.122
SNRPE	TRCN0000072270	0.31	0.277
SNRPF	TRCN0000074353	0.485	0.414
SNRPF	TRCN0000074354	0.364	0.312
SNRPF	TRCN0000074355	0.357	0.305
SNRPG	TRCN0000074339	0.291	0.295
SNRPG	TRCN0000074340	0.161	0.163
SRBD1	TRCN0000048763	0.437	0.494
SREK1	TRCN0000001308	0.489	0.466
SRP72	TRCN0000150839	0.398	0.324
SRSF1	TRCN0000001093	0.421	0.398
SRSF1	TRCN0000001095	0.283	0.268
SRSF12	TRCN0000151704	0.117	0.13
SRSF12	TRCN0000156842	0.16	0.177

Table 5.1 (continued): Human 293T validated shRNAs for 273 RNA binding proteins

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
SRSF3	TRCN0000001224	0.361	0.404
SRSF4	TRCN0000006627	0.39	0.443
SRSF6	TRCN0000006622	0.321	0.309
SRSF7	TRCN0000001140	0.416	0.537
SRSF7	TRCN0000001141	0.226	0.291
SRSF7	TRCN0000001142	0.108	0.139
SRSF9	TRCN0000006630	0.453	0.466
SSB	TRCN0000062193	0.235	0.135
SSB	TRCN0000062194	0.105	0.061
SSB	TRCN0000062195	0.368	0.212
STAU2	TRCN0000153783	0.442	0.353
STAU2	TRCN0000157149	0.18	0.144
STAU2	TRCN0000178809	0.593	0.474
SUGP1	TRCN0000074773	0.604	0.444
SUGP2	TRCN0000153203	0.631	0.452
SUGP2	TRCN0000153826	0.315	0.226
SUPT6H	TRCN0000019730	0.309	0.268
TAF15	TRCN0000020141	0.474	0.487
TAF15	TRCN0000020143	0.335	0.344
TARBP2	TRCN0000019339	0.309	0.483
TARBP2	TRCN0000019340	0.269	0.421
TARDBP	TRCN0000016038	0.362	0.317
TARDBP	TRCN0000016039	0.548	0.48
TARDBP	TRCN0000016040	0.378	0.331
TBRG4	TRCN0000154962	0.448	0.434
TBRG4	TRCN0000155887	0.485	0.47
TDRD3	TRCN0000073293	0.226	0.198
TEP1	TRCN0000039830	0.337	0.454
TEP1	TRCN0000039831	0.306	0.413
THOC4	TRCN0000000351	0.526	0.404
TIA1	TRCN0000074463	0.441	0.348
TIA1	TRCN0000074464	0.6	0.473
TNRC6C	TRCN0000004491	0.406	0.458
TNRC6C	TRCN0000004492	0.289	0.326
TRA2A	TRCN0000074548	0.213	0.155
TRA2A	TRCN0000074549	0.412	0.3
TRA2B	TRCN0000000115	0.627	0.49

Table 5.1 (continued): Human 293T validated shRNAs for 273 RNA binding proteins

Gene	shRNA TRC ID	Median Fold Change	Control Fold Change
TRA2B	TRCN0000000116	0.505	0.395
TRA2B	TRCN0000000117	0.552	0.432
TUT1	TRCN0000129314	0.356	0.334
TUT1	TRCN0000129603	0.469	0.44
TUT1	TRCN0000150242	0.473	0.443
U2AF1	TRCN0000001155	0.273	0.238
U2AF1	TRCN0000001156	0.47	0.409
U2AF2	TRCN0000001160	0.44	0.362
U2AF2	TRCN0000001162	0.418	0.344
UHMK1	TRCN0000003281	0.487	0.694
WRN	TRCN0000004899	0.579	0.448
WRN	TRCN0000004900	0.633	0.489
YTHDC2	TRCN0000154178	0.379	0.453
YTHDC2	TRCN0000157638	0.491	0.588
ZCCHC17	TRCN0000062683	0.54	0.476
ZNF638	TRCN0000146611	0.458	0.431
ZRSR2	TRCN0000074628	0.394	0.497
ZRSR2	TRCN0000074629	0.492	0.622

5.6 References

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., Jr., and Yeo, G.W. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Rep *1*, 167-178.

Lagier-Tourenne, C., Polymenidou, M., and Cleveland, D.W. (2010). TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. Hum Mol Genet *19*, R46-64.

Lovci, M.T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T.Y., Stark, T.J., Gehman, L.T., Hoon, S., Massirer, K.B., Pratt, G.A., Black, D.L., Gray, J.W., Conboy, J.G., and Yeo, G.W. (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat Struct Mol Biol *20*, 1434-1442.

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., and Hughes, T.R. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature *499*, 172-177.

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., and Ares, M., Jr. (2006). Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLoS Comput Biol *2*, e4.

Venables, J.P., Lapasset, L., Gadea, G., Fort, P., Klinck, R., Irimia, M., Vignal, E., Thibault, P., Prinos, P., Chabot, B., Abou Elela, S., Roux, P., Lemaitre, J.M., and Tazi, J. (2013). MBNL1 and RBFOX2 cooperate to establish a splicing programme involved in pluripotent stem cell differentiation. Nat Commun *4*, 2480.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873-881.

6 Summary and Perspectives

The work presented takes several leaps toward understanding the complex and combinatorial regulation of RNA processing by RNA binding proteins (RBPs) through multiple avenues. First, I've taken a careful look at some key developmental (LIN28), and disease related (TDP-43) RBPs, providing insight into their unique individual roles in the cell and in disease. Next I've studied a family of abundant and ubiquitously expressed RBPs, the hnRNPs, gaining a new understanding of their coordinated and redundant regulation of alternative splicing. Then, I expanded my studies to identify the protein interactors for hnRNP proteins, and in doing so, have identified some putative novel RBPs. Finally, I've broadened my focus to a much larger set of RNA binding proteins, including those with known and unknown functions, to understand individual roles and also begin to build a more complete picture of the complex regulation of RNA processing by many RBPs.

The data I have generated is still rich in additional information that will be mined. As an example, hnRNP proteins have been implicated in the regulation of a handful of microRNAs (miRNA) (Guil and Caceres, 2007) and also large intergenic noncoding RNAs (lincRNAs) (Huarte et al., 2010). It is worth extending the analysis of hnRNP binding sites and regulation to these non-coding RNAs. The study of non-coding RNAs, specifically lincRNAs, is a burgeoning area of RNA research, since many still have unknown functions. Identifications of RBPs that regulate the processing of ncRNAs and affect their function will aid in our understanding of these curious RNA molecules. Additionally, more detailed information to the actual coordination of regulation by multiple RNA binding proteins should be explored. For example, an interesting approach has been made in this direction for two SR proteins, SRSF1 and SRSF2 (Pandit et al., 2013),

where CLIP-seq experiments were done for each of the SR proteins under depletion conditions of the other SR protein. This pilot experiment gave rise to some interesting information about the redundancy of these particular RBPs, but more similar experiments like this should be explored with other RNA binding proteins, such as those important in disease like TDP43 and FUS/TLS, or between hnRNPs that hold similar binding profiles. Further, more information about whether roles are indeed coordinated, binding the same RNAs together, or binding the same RNA in different cellular compartments or at different times, should be analyzed. Experiments such as “CLIP-CLIP”, where one protein and its RNA targets are immunoprecipitated and then the eluted RNAs are further immunoprecipitaed with a different protein to identify RNA targets that are bound by both proteins together. Further optimization of these experiments will be necessary, since current protocols do not yield enough RNA product for a double IP. Recent advances in low-input sequencing may make some of these directions possible.

6.1 A wealth of “Big Data”

In the coming years, more data will be generated to elucidate more RBP functions. In addition to individual lab efforts to generate RNA targets and data related to RNA binding proteins, the NIH funded ENCODE (Encyclopedia of DNA elements) project has begun a new RNA element effort, which plans to identify RNA targets via CLIP-seq for 250 RBPs in two human cell lines, K562 and MCF7. In both of these lines, the same RBPs will also be depleted and analyzed by RNA-sequencing, a study accelerated by my identification of shRNAs that specifically deplete close to 300 RBPs (Chapter 5). This large-scale generation of data will introduce many new computational challenges. Efficient and accurate methods for quality control, automated standard analysis pipelines, and visualization of this large amount of data will be crucial for the

success of these projects. Additionally, storage of the data and computational frameworks for managing the data will be incredibly important in making this data useful and accessible. Most importantly, novel ways of data integration to answer questions within RNA biology will need to be explored.

As we enter into this new, “Big Data” era of biological research, more computational scientists will be required for the development of novel algorithms and analysis methods, which will be key to making sense of the wealth of biological data. The vast amount of data generated in the years to come can subsequently be used as input to machine learning algorithms to predict regulatory changes given direct RBP binding and contextual genomic information. The ability to model such combinatorial regulation and identify the features that are most predictive of the changes will help establish a better understanding of an underlying *RNP code*.

Additionally, as more and more factors that influence splicing are revealed by microarray and sequencing technologies, the *splicing code* continues to unravel. The splicing field is now nearing the end of a cataloging phase, as researchers are now attempting to pull together the individual regulatory elements involved in constitutive and alternative exon splicing into a cohesive code. Various computational approaches are being employed to sort and combine identified splicing regulatory elements as part of a layered splicing network. Sequence elements, traits near alternatively spliced exons, and proximal direct RBP binding provide a set of features useful for training machine learning classifiers to predict exon inclusion or exclusion.

The initial successes of algorithms like ExonScan, which uses splice site models and known ESEs and ESSs to predict the splicing patterns of transcripts (Wang et al., 2004), prove that this sort of predictive analysis of alternative splicing is possible. Other algorithms that rely upon basic splicing elements such as sequence conservation and

motifs for their predictions have been successful in predicting unannotated, alternative exons. Sorek et al. were the first to make alternative splicing predictions without the use of ESTs and were able to determine many novel splice variants using only the surrounding genomic sequence data (Sorek et al., 2004). Another algorithm, ACESCAN, combined multiple sequence features into a regularized least-squares classifier to identify exons subject to evolutionarily conserved exon skipping (Yeo et al., 2005). This method accurately predicts about 2,000 alternative conserved exons in the human and mouse genomes. In 2005, Dror et al. used a support vector machine (SVM) classifier with seven features including, length of the exon, exon divisibility by three, conservation to mouse, and various flanking sequence features to predict alternatively spliced exons (Dror et al., 2005). While these predictive methods combine multiple splicing features, many more features exist, and can be combined for improved predictive algorithms.

Brendan Frey's group developed the first combinatorial method for predicting tissue-specific exon inclusion, that combines over 1000 established *cis* and *trans* SREs into a SVM classifier (Barash et al., 2010). This approach delineates tissue-specific splice patterns and can correctly predict splicing for 80% of cassette exons across 4 mouse tissues. This approach supported a giant leap in our ability to predict the occurrence of the tissue-specific alternative splicing.

Post-transcriptional gene expression regulation, specifically through alternative splicing, is crucial for most cell processes. Novel tissue-specific transcript variants are quickly being discovered through microarray and sequencing technologies, and techniques such as CLIP-seq are rapidly contributing to our understanding splicing code's defining factors. As more data is generated, we are beginning to rely more heavily upon computational tools that draw increasingly significant results from this abundance of data. Eventually, a complete understanding of alternative splicing may

guide new ways to control stem cell fate, allow finer distinction among cell subtypes, and provide a molecular basis for human disease model development.

6.2 A new perspective: the single cell

In the twilight of this work, a whole new area of research has sprouted, single cell analysis. This area of research has reshaped the way that we think about many problems in biology and, in particular, transcriptomics. When we study a population of cells, the measure of RNA transcript abundance as well as RNA modifications such as editing and splicing are averages of many individual cells within a bulk population. Recent discoveries from single cells have piqued the interest of many scientists around the world revealing that the average measure of the population loses all information related to the actual heterogeneity of individual cells.

One example in particular is that of alternative splicing. From our RNA-sequencing data of bulk cells we can get measures of the percent spliced in a particular exon is for a given gene. However, the flaw here lies in the fact that we take this measurement from a large population of cells. This begs the question then, is our percentage reflective of the percent of isoforms of the gene within a cell that have the exon spliced in, or the percent of cells that has the gene with the exon spliced in? Thus, what is the true percent spliced in for an individual cell? Scientists have begun to make headway in answering these questions and other related questions using single cell RNA-sequencing. Barbara Wold and colleagues have found that, in fact, individual cells seem to have percent spliced in values that are more binary, either the exon is in or the exon is out, compared to bulk cells where values distribute between 0 and 1, suggesting that only one isoform of that gene exits in a particular cell (Marinov et al., 2014).

Single cell sequencing can also come in handy when we start to think about depleting an RBP within a population of cells. If we get a 50% KD, is it because 50% of

cells have 100% KD, or is it down 50% across all cells, or somewhere in between? Additionally, when single-cell methods become more mature, performing experiments like CLIP-seq in a set of single cells will give some interesting insight into the actual binding preferences of RBPs. We can begin to ask if “weakly” bound sites identified from bulk cell CLIP-seq are actually transient interactions of an RBP with an RNA molecule in few cells of a population? And, for “strongly” bound sites identified from bulk cell CLIP-seq, is this RNA bound in every cell of a population? Single cell analysis and sequencing can begin to tease apart differences across a population of cells and reveal novel findings about the cellular homeostasis.

This field of research has been growing quickly, in part, due to the development of kits and technologies that aid in the isolation and experimental analysis of single cells. Kits specific to single cell nucleic acid extraction and amplification are becoming a targeted product for development at biotech companies and are evolving quickly. Additionally, improved capturing techniques for single cells such as Fluidigm’s microfluidic C1 system are making single cell data generation more accessible and automated as opposed to traditional FACs sorting techniques. Coupled with improved library prep methods for small quantity samples (Clontech Smart Seq) and the rapid decrease in sequencing cost, we can now sequence a population of hundreds of cells individually at the single cell level.

Similar to the RNA ENCODE effort, along with this new technology comes new computational data analysis challenges. We do not quite know how to handle single cell data yet, and we are still developing interesting ways to use this data in ways benefiting to single cells compared to bulk cell data. Novel visualization techniques and cutting-edge analysis and integration techniques will be key to fully understand the wealth of information coming from single cell data.

Taken together, the future of RNA biology, and specifically our understanding of RNA binding proteins in their post-transcriptional control will rely on the development of novel experiments and systems such as single cells, as well as the advances in accurate data generation, and advanced computational analyses to aid in a deeper understanding of “the knot” of RNA regulation by RNA binding proteins (Figure 6.1).

6.3 Summary publication acknowledgment

Chapter 6 is, in part, published as a chapter entitled “Alternative Splicing in Stem Cells” in the electronic book *Computational Biology of Embryonic Stem Cells* (Huelga and Yeo, 2012). This dissertation author was the primary author who reviewed the literature and wrote text.

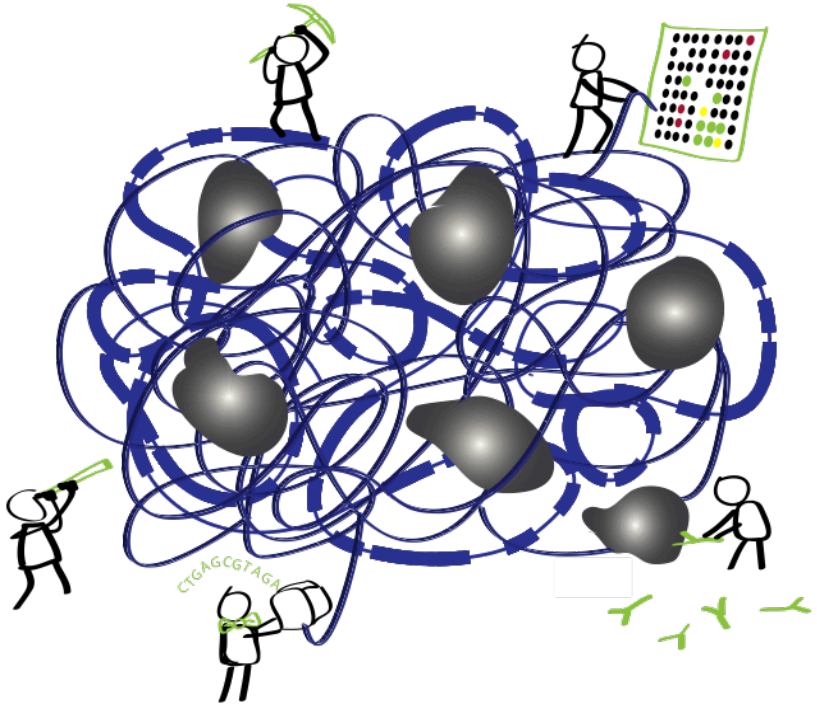


Figure 6.1: The untangling of RNA-protein knots

A metaphorical depiction of a knot of RNA bound by RNA binding proteins. Scientists, and the Yeo Lab in particular, use several focused and genome-wide approaches (Microarrays, RNA-sequencing, and CLIP-sequencing) to reveal basic RNA processing events and begin to untangle this knot of RNA regulation.

6.4 References

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53-59.

Dror, G., Sorek, R., and Shamir, R. (2005). Accurate identification of alternatively spliced exons using support vector machine. Bioinformatics *21*, 897-901.

Guil, S., and Caceres, J.F. (2007). The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. Nat Struct Mol Biol *14*, 591-596.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M., Attardi, L.D., Regev, A., Lander, E.S., Jacks, T., and Rinn, J.L. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell *142*, 409-419.

Huelga, S.C., and Yeo, G.W. (2012). Alternative Splicing in Stem Cells. In Computational Biology of Embryonic Stem Cells, M. Zhan, ed. (Bentham Science Publishers), pp. 147-160.

Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., and Wold, B.J. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res *24*, 496-510.

Pandit, S., Zhou, Y., Shiue, L., Coutinho-Mansfield, G., Li, H., Qiu, J., Huang, J., Yeo, G.W., Ares, M., Jr., and Fu, X.D. (2013). Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. Mol Cell *50*, 223-235.

Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G., and Shamir, R. (2004). A non-EST-based method for exon-skipping prediction. Genome Res *14*, 1617-1623.

Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell *119*, 831-845.

Yeo, G.W., Van Nostrand, E., Holste, D., Poggio, T., and Burge, C.B. (2005). Identification and analysis of alternative splicing events conserved in human and mouse. Proc Natl Acad Sci U S A *102*, 2850-2855.