UC Santa Cruz
UC Santa Cruz Electronic Theses and Dissertations

Title

Investigation Of In Vivo RNA-Protein Interactions Using Individual Nucleotide Resolution Cross-Linking Immunoprecipitation (ICLIP)

Permalink

https://escholarship.org/uc/item/1483v1tz

Author

Howard, Jonathan Michael

Publication Date

2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

INVESTIGATION OF *IN VIVO* RNA-PROTEIN INTERACTIONS USING
INDIVIDUAL NUCLEOTIDE RESOLUTION CROSS-LINKING
IMMUNOPRECIPITATION (ICLIP)

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

DOCTOR OF PHILOSOPHY

In

MOLECULAR, CELL, AND DEVELOPMENTAL BIOLOGY

by

Jonathan M. Howard

December 2015

The Dissertation of Jonathan M. Howard is
approved:

_____
Assoc. Professor Jeremy Sanford, chair

_____
Distinguished Professor Manuel Ares, Jr.

_____
Department Chair, Professor Alan Zahler

_____
Vice Provost and Dean of Graduate Studies

Table of Contents

Thesis Abstract

Eukaryotic gene expression involves a complex system of checkpoints that regulate RNA biogenesis, maturation, and localization. Along the way, these RNA will encounter a host of trans-acting factors, collectively known as RNA binding proteins (RBPs), which will govern the fate of RNA at each stage of gene expression. These RNA-protein interactions are the principal regulators of post-transcriptional control and are critical to the accurate expression of human genes. Misregulation of these interactions has been seen to be highly associated with several neurological disorders, cancer and inflammatory diseases. However, the regulatory mechanisms by which RBP function in the context of their RNA associations remains less understood. For my dissertation, I studied the RNA-protein interactions of a small set of proteins involved in the global regulation of transcription, processing and stability of the cellular transcriptome.

In order to interrogate the global RNA interactions of the proteins of interest, I performed a technique called individual nucleotide resolution crosslinking immunoprecipitation, or iCLIP. This technique takes advantage of photoreactive amino acids and nucleic acids in order to capture the in situ RNA-protein interactions of endogenous factors from both cell culture and tissue samples. Partial RNase digestion, stringent purification conditions, and advantageous use of reverse transcriptase difficulty reading through protein adducts allow for direct identification both the genomic origin of the CLIP RNA and the exact crosslinking site of the RNA-binding protein of interest at high specificity. These data form a functional RNA-protein interaction map to elucidate

putative molecular roles for any RBP. I have used this technique on three separate

projects in order to elucidate the mechanisms by which these RNA binding proteins

regulate their targets and interact and modulate the binding of other RNA-binding

proteins. First, I produced a literature review of a family of splicing factors known as SR

proteins and their multifaceted roles in human gene expression regultation. Secondly, I

studied the effects of modulating the cellular levels of a splicing repressor, hnRNP A1, on

the RNA recognition and binding of two splicing enhancer, U2AF65 and SRSF1, and the

subsequent changes in pre-mRNA splicing. Thirdly, in two concurrent collaborations, I

elucidated the endogenous RNA binding targets of IGF2BP3, an oncofetal RBP whose

up-regulation is associated with aggressive pancreatic cancer and B-cell leukemia.

Finally, I elucidated the effects of human disease-associated genetic mutations on RNA-

protein interactions in the context of pre-mRNA splicing. These studies are discussed

herein.

Acknowledgments

I cannot express enough thanks to my committee for their continued support and encouragement: Dr. Jeremy Sanford, my advisor; Dr. Manuel Ares, Jr.; and Dr. Alan Zahler. I offer my sincere appreciation for the learning opportunities provided by my committee. Thanks to my parents, Dr. and Mrs. Michael and Sharon Howard, my brother David, my brothers-from-other-mothers Ben Edquist and Matthew Pelz, and family, both biological and in Myrtle Beach. Your encouragement and support when the times got rough are much appreciated and duly noted. Finally, thank you to my friends and lab mates. If it were not for you, I would have finished my thesis two years earlier.

Dedication

To those who inspired it and will not read it.

Chapter 1

Introduction

Eukaryotic gene expression involves multiple steps that control the use of genetic information (DNA) to produce a functional molecular product (RNA or protein). And while DNA and protein function are regulated on multiple levels, whether it be transcription factors or histone modifications modulating DNA availability or post-translational modifications and ligands modifying proteins, the production and use of RNAs is arguably the most heavily regulated step of the central dogma of molecular biology. From their production through transcription, to their maturations through processes such as capping, splicing and poly-adenylation, to their final intended roles in translation or downstream gene expression, RNAs pass through a series of regulatory steps intended to ensure proper gene expression, but also to quickly respond to the needs of the cell.

The many steps that are required to regulate RNA biogenesis and downstream use are dominated by interactions with a superfamily of factors called RNA binding proteins (RBPs). RBPs often bind to RNA in order to create ribonucleic protein particles (RNPs), which regulate RNA maturation, localization, utilization, or degradation/turnover (Figure 1). These proteins have the capacity to interact with single-stranded RNA, usually, but not always, in a sequence-specific manner, or through double-stranded structures, such as stem-loops, depending on the specificity and composition of their respective RNA-binding domains (RBD) (Lunde et al., 2007) .The three most common RBDs encompass roughly 50 percent of all RBPs, compared to the more than 80 percent of DNA-binding proteins that fall under the 3 most common DNA-binding domains (Gerstberger et al., 2014b). This reflects the

structural and functional diversity of RNA as compared to DNA, but also the layered

complexity of post-transcriptional gene expression that has evolved to require a

varied ensemble of *trans*-acting regulating factors.

RNA-binding proteins are also associated with a host of human diseases

(Castello et al., 2013; Gerstberger et al., 2014a; Lukong et al., 2008). Hundreds of

disease-associated mutations map to RBPs, the vast majority of which are mRNA

binding proteins, suggesting mutation can not only affect protein structure and

function but also downstream regulation of gene expression. Also, there is evidence

that certain RNA-binding proteins are strongly dysregulated in several cancers,

suggesting downstream effects on expression regulation that may help to support

oncogenesis and cancer progression in a possible "non-oncogene" context (Luo et al.,

2009).

Messenger RNA (mRNA) binding proteins comprise the largest and most

functionally diverse group of RBPs, regulating all major steps of eukaryotic gene

expression (Gerstberger et al., 2014b).   These proteins are known to interact with

mRNA during any step between biogenesis and decay. While domain studies of

putative RNA-binding and RNA-processing proteins had once put the number of total

RBPs at approximately 700, recent *in vivo* RNA-protein crosslinking experiments in

conjunction with mass spectrometry now put this number over 800 (Castello et al.,

2012). And while the majority of all RBPs have only one RNA-binding domain,

mRNA binding proteins tend to have multiple RNA binding domains, possibly

allowing for flexibility in RNA recognition or increased functionality with regards to

RNA binding (i.e. protein-protein interactions, non-specific RNA interactions) (Gerstberger et al., 2014b).

Herein, I will discuss at length a canonical example of an RNA binding protein family, the SR family of splicing factors, and their multifaceted roles in mammalian gene expression. I will also discuss the roles of hnRNP A1 and U2AF65 in post-transcriptional gene regulation. Futhermore, I will discuss the known interplay between hnRNP A1, and U2AF65 and SRSF1 which is integral to defining 3' and 5' splice sites of exons during pre-mRNA splicing. I will then give a brief introduction to an additional RNA binding protein, IGF2BP3, and its roles in development and cancer progression. Finally, I will give a brief overview of the history of techniques used to interrogate the RNA targets of RNA binding proteins, focusing on the main experimental technique I used during my research, individual nucleotide resolution crosslinking immunoprecipitation, or iCLIP.

hnRNP A1 in Post-Transcriptonal Gene expression

Heterogeneous nuclear ribonucleoproteins (hnRNPs) represent one of the largest and most diverse groups of RBPs within the cell. These proteins are known to function in all aspects of eukaryotic gene expression from transcription and processing of pre-mRNAs to export and regulation of expression outside the nucleus (Jean-Philippe et al., 2013). This family of proteins was first identified as forming "histone-like" RNP particles on nascent transcripts produced from RNA polymerase

II, much like "beads on a string" structures of DNA nucleosomes (Beyer et al., 1977). These proteins vary structurally, however the majority contains an RNA recognition motif (RRM), hnRNP K homology domain (KH) and/or arginine-glycine-glycine repeat domains (RGG), all of which are known to confer RNA-binding potential (Han et al., 2010).

By far the most highly and ubiquitously expressed of the hnRNP factors is hnRNP A1. Structurally, hnRNP A1 consists of two N-terminal RRM domains followed by a C-terminal region contains an RGG box RNA binding domain and a nuclear localization sequence (He and Smith, 2009). The most studied function of hnRNP A1 is its regulation of constitutive and alternative splicing. With respect to constitutive splicing, purification of spliceosomal complexes finds the presence of hnRNP A1, indicating that this protein participates in spliceosome assembly (Jurica et al., 2002; Zhou et al., 2002). hnRNP A1 has been shown to "proofread" 3' splice site AG dinucleotides and regulate U2AF binding, thus hnRNP A1 can act as an enhancer of U2AF binding when the appropriate 3' splice site signals are present, but a repressor if the site is suboptimal (Tavanez et al., 2012). hnRNP A1 has also been shown to regulate alternative splicing of cassette exons, which will be discussed below. However the functions of hnRNP A1 occur primarily through association with exonic and intronic splicing silencers (ESSs and ISSs) which have been previously identified (Burd and Dreyfuss, 1994; Hamilton et al., 1993; Huelga et al., 2012; Pastor and Pagani, 2011).

hnRNP A1 also has functions outside of splicing with respect to gene regulation. For example, hnRNP A1 has been shown to bind to the promoter regions of a subset of gene and act as both a transcriptional activator and repressor (Campillos et al., 2003; Xia, 2005). hnRNP A1 also associates with 7SK RNA and causes TEF-b dissociation and a correlated activation of RNA Pol II (Barrandon et al., 2007). hnRNP A1 has also been shown to bind to repeat sequences in telomeres and regulate their biogenesis and potentially protect against their degradation (Zhang et al., 2006). A subset of hnRNP proteins, including hnRNP A1 are seen to shuttle between the nucleus and cytoplasm suggesting hnRNP A1 may regulate RNA targets outside of the nucleus (Mili et al., 2001). For example, hnRNP A1 has been shown to associate with AU-rich elements in its RNA targets, which are known to regulate mRNA stablilty and translation (Hamilton et al., 1997; Hamilton et al., 1993). hnRNP A1 has also been observed to associate with IRES and enhance cap-independent translation for a subset of transcripts (Jo et al., 2008). Most recently, hnRNP A1 has been seen to regulate miRNA biogenesis and processing of miR- 18a and let-7a (albeit with opposite outcomes) through interactions in the loop region (Guil and Caceres, 2007; Michlewski and Caceres, 2010; Michlewski et al., 2011). Overall, hnRNP A1 regulates several aspects of mRNA and ncRNA biogenesis and metabolism.

While hnRNP A1 appears to play multifaceted roles in numerous areas of gene expression, it also appears to play roles in disease-associated pathways. For example dysregulation of hnRNP A1 expression has been show to correlate with development of neurodegenerative disorders, such as Alzheimer's disease and

amyotrophic lateral sclerosis (ALS) (Bekenstein and Soreq, 2013). This may be due to, in part, the presence of a glycine-rich "prion-like" carboxy-terminal domain (Shorter and Taylor, 2013). hnRNP A1 overexpression has also been associated with oncogenesis and shown in a subset of cancer types, such as colorectal, breast, liver, and lung (David et al., 2010; Pino et al., 2003; Ushigome et al., 2005; Zhou et al., 2013). Thus, understanding the RNA-protein interaction of hnRNP A1, as well as the additional factors in which it interacts with on a functional level, will afford better understanding of hnRNP A1 roles in of gene regulation and molecular mechanisms of disease.

U2AF65 as regulator of post-transcriptional control

U2 snRNP Auxiliary factor 65 (U2AF65) is part of a heterodimer complex (which includes U2AF35) that is considered one of the most critical factors important for 3' splice site definition during pre-mRNA splicing. U2AF65 was initially found to act as an enhancer of U2 snRNP binding with the intronic branch point sequence during splicing (Ruskin et al., 1988). Further evidence in *Drosophila* found U2AF65 could regulate Sex lethal (Sxl) splicing via association with a poly-pyrimidine tract found upstream of most exonic 3' splice sites (Valcarcel et al., 1993). Additionally, work in fission yeast suggest that U2AF65/35 can associate in context of U2AF35 binding to AG dinucleotides, suggesting cooperativity in defining exonic 3' splice sites (Webb and Wise, 2004). U2AF65/35 appears to be required for *in vivo* splicing

and survival in mammals, *Drosophila*, and *Xenopus* (Wu and Fu, 2015) . The exception, *Saccharomyces cerevisiae*, does have a homolog for U2AF65 called Mud2, however this protein has been shown to be non-essential (Kistler and Guthrie, 2001).

Structurally, U2AF65 contains and N-terminal arginine/serine-rich (RS) motif, two internal RNA recognition motifs (RRM), as well as a C-terminal U2AF homology motif (UHM) (Kielkopf et al., 2004; Wu and Fu, 2015). The RS domain is thought to interact non-specifically with the negative phosphate backbone near the branch point sequence to quench electrostatic charges and facilitate U2 snRNA interactions (Valcarcel et al., 1996). The UHM has been shown to dimerize with the UHM of Splicing factor 1 to enhance interactions near the branch point adenosine (Rain et al., 1998). The RRMs are thought to both facilitate binding to polypyrimidine tracts, acting together as a putative molecular "rheostat" that can detect inherent sequence variation in the polypyrimidine tracts in order to modulate splicing via recruitment of U2 snRNP (Mackereth et al., 2011). Interestingly, structure studies suggest a small proline-rich region that exists between the RS and N-terminal RRM of U2AF65 can mediate binding to U2AF35 UHM (Kielkopf et al., 2001), potentially leading to RNA binding via its RRMs (Kellenberger et al., 2002).

Binding of U2AF65 near 3' splice sites is also regulated by a host of splicing factors that influence early spliceosome assembly at 3' and 5' splice sites. Recent CLIP data for U2AF65 shows its occupancy at roughly 88% of all human 3' splice sites, reinforcing its acknowledged function as a splicing factor (Shao et al., 2014).

However, this data reveals not only that U2AF65 does not bind at all functional 3'
splice sites, but also has the capacity to bind within and downstream of exons,
suggesting that U2AF65 may not be required for all splicing events and may have
function in RNA processing outside of the 3' splice site This suggests that U2AF65
RNA binding can be influenced by the assembly of RBPs at or near a given splice
site. In fact, recent research suggests that U2AF65 also has the capacity to repress
splicing through interaction with sites outside the 3' splice site (Cho et al., 2015).
These types of interactions may also influence U2AF65-RNA association at RNA
outside of the 3' splice site context, as hnRNP C antagonizes U2AF65 *Alu* element
binding to regulate cryptic exon splicing (Zarnack et al., 2013). Therefore, U2AF65-
RNA associations, while essential for the majority of splicing events, occurs and is
regulated outside of 3' splice sites, suggesting potential functions beyond splice site
definition.

hnRNP A1 as antagonist to splicing enhancer factors

Currently, two models exist that illustrate the interplay amongst RBPs during
exon definition: the recruitment model and the inhibition model. (To simplify, I will
discuss these models in the context of a known splicing enhancer SRSF1 and splicing
repressor hnRNP A1). The 'recruitment' model states that splicing factors, like
SRSF1, bind to exonic splicing sequences and facilitate loading of spliceosomal
factors, which are associated with the E complex of spliceosome assembly (Mabon

and Misteli, 2005). These types of splice site definitions are thought to occur through protein-protein interactions between enhancer-bound SRSF1 and recruited splicesosomal factors. The "inhibitor" model states that splicing enhancers compete with splicing repressive factors in the context of alternative exons. Indeed, pre-mRNA splicing research is filled with examples of splicing competition between SR proteins and its molecular archrival, the hnRNP family, and interaction between these two protein families has become a mainstay rule in the world of splicing factors (Cartegni and Krainer, 2002; Dirksen et al., 2000; Kashima and Manley, 2003; Rooke et al., 2003; Venables et al., 2005; Zahler et al., 2004).

The antagonistic relationship of between SRSF1 and hnRNP A1 is often representative of "inhibitor" roles of hnRNP A1 suggested in the literature. Multiple models have been described in the splicing field as to how hnRNP A1 plays roles as an "inhibitor" to exon inclusion during pre-mRNA splicing. For example, there is "competition" model (Figure 1, first panel), in which hnRNP A1 will bind to splicing regulatory sequences called "exonic splicing silencers" (ESSs) found proximal to "exonic splicing enhancers" (ESEs) which would be normally occupied by splicing enhancer proteins such as SRSF1 (Eperon et al., 2000; Zahler et al., 2004). These proteins can ultimately compete for binding in these contexts and regulate exon inclusion/skipping. Another model is the "spreading" model (Figure 1, middle panel), in which hnRNP A1 occupying an ESS can promote multimerization of hnRNP A1 that can spread across a regulated exon, thus antagonizing binding of splicing enhancers and splicesomal factors, such as U2AF2 (Okunola and Krainer, 2009; Zhu

10

et al., 2001). Finally, a third model proposed is the "looping" model (Figure 1, last panel), in which hnRNP A1 binding to splicing regulatory elements called "intronic splicing silencers" (ISSs) on either side of an alternative exon can dimerize, causing "looping out" of an exon from the pre-mRNA and promote its exclusion from the final mature transcript (Blanchette and Chabot, 1999). While it has not been suggested that these models are mutually exclusive, they do suggest that hnRNP A1 can regulate exonic skipping through multiple mechanisms. However, these models have only been observed using *in vitro* biochemical means, and have not been studied on a global scale.

Techniques to elucidate RNA-protein interactions and specificity

The ultimate fate of any RNA in a cell is mainly regulated by two distinct, but not mutually exclusive, elements: the cis-elements that give a RNA a particular functionality and inherent destination and the trans-acting factors that shape its birth, use, and ultimate demise. Recent research has focused on understanding the RNA sequences and RNA binding proteins work together to regulate gene expression (Li et al., 2014a; Ray et al., 2009). However, the molecular functions of many RNA binding proteins have been studied in a low-throughput manner, whether it be *in vitro* processing of splicing constructs, gel shift assays to assay protein specificity for an RNA target. What was needed was a more systematic way to investigate RNA-

protein interactions to elucidate binding specificity and give insights into general

mechanism by which these RBPs function.

The development of Systematic evolution of ligands by exponential

enrichment, or SELEX, allowed for a more analytical methodology to explore the

RNA specificity of a range of RNA binding proteins (Ellington and Szostak, 1990).

This technique relies on three processes: selection of ligand from a randomized pool

of RNA sequences that bind to a target; washing away bound RNA ligands from non-

bound RNA ligands via affinity methods; and amplification of the bound RNA

ligands via RT-PCR. The outcome, in principle, is isolation of a highly specific pool

of RNA oligos containing motifs with high affinity to the ligand of choice. This

technique was initially used to find RNA aptamers that preferentially bound to small

molecules in hopes of isolating sequences with ribozyme potential (Huizenga and

Szostak, 1995). It was subsequently appropriated in use of finding enzyme-binding

specificity of S-adenosyl methionine (SAM), a common methyl donor for methyl

transfer reactions (Burke and Gold, 1997). These experiments were often

complemented by electromobility shift assays (EMSAs) in order to validate SELEX

results and determine dissociation constants of these ligand-aptamer complexes.

SELEX was also performed in the context RNA immunoprecipitation (RNA-IP or

RIP), in which proteins purified either through an affinity tag or antibodies were

mixed with pools of randomized RNA aptamers (Reid et al., 2009). Bound RNAs are

enriched through multiple rounds of selection and ultimately sequenced to determine

those sequences that had high affinity for a given RBP. Collectively these data were

12

useful in the context of understanding the specificity of a ligand for RNA, they lack information with regards to the cellular and molecular context with which these interactions would occur.

Functional or *in vivo* SELEX (Berkhout and Klaver, 1993) established an in-cell selection system to determine the functionality of RNA sequences in the context of splicing reporter minigene transiently transfected into cell culture (Fairbrother and Chasin, 2000). Depending on the extent to which a reporter exon was included or skipped in the final product from the reporter, the incorporated sequence would be classified as an exonic splicing enhancer (ESE) or exonic splicing silencer (ESS), respectively. These data could be combined with prior classic SELEX experiments to suggest possible mechanisms by which a specific splicing factor could associate with RNA in order to elicit a splicing outcome. However, these data only suggested a correlation between the RNA-protein interaction affecting RNA processing in the context of a single exogenous splicing reporter. What these data lacked was an understanding of the interactions between a protein and endogenous RNA on a transcriptome-wide scale in order to determine the endogenous function of these RNA-protein interactions.

In 2003, Ule et al. provided a technique called CLIP (crosslinking immunoprecipitation) , a method that allowed for *in situ* RNA-protein associations to be maintained and identified by sequencing on a protein-by-protein basis (Ule et al., 2003). This technique combined UV-crosslinking of mouse brain cells in combination with immunoprecipitation of neuronal-specific splicing factors neuro-oncological

ventral antigen -1 and -2 (NOVA1 and NOVA2). Briefly speaking, CLIP relies on the photoreactivity of polar amino acids and the carbon rings of nucleic acids (specifically pyrimidines) to form irreversible covalent crosslinks between protein and RNA. These crosslinks allow for RNA-protein interactions to be maintained, immunoprecipitation and subsequent partial trimming of the RNA by nuclease digestion. This forms a small "CLIP tag", often 30-50 nucleotides in length bound to the protein of interest. The RNA CLIP tags are then modified by 5' and 3' dephosphorylation and subsequent 5' re-phosphorylation with a radioactive phosphate group using gamma-ATP. Immunoprecipitation RNA-protein complexes are then subjected to denaturing protein gel electrophoresis and then transferred to nitrocellulose. RNA-protein complexes running 10-30 kDa larger than their RNA-free forms will be extracted from the nitrocellulose and subjected to Proteinase K digestion to free crosslinked RNA. These RNAs are subjected to reverse transcription followed by PCR using linkers added after immunoprecipitation and subsequently cloned into a vector for Sanger sequencing. This generates a genomic map of the *in situ* interactions of a protein with RNA in-cell, but also offer a view of the RNA regulon of a protein of interest.

Biochemical and high-throughput modifications have been made to this original CLIP methodology in order to optimize sensitivity, resolution, and reduce inherent bias. The development of affordable high-throughput sequencing massively increased the amount of RNA that could be sequenced from a CLIP experiment, a technique coined HITS-CLIP (high-throughput sequencing of RNA from crosslinking

immunoprecipitation) (Licatalosi et al., 2008). This allowed for a more sensitive genome-wide look at the biochemical footprint of an RBP and its RNA interactions. PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) utilizes photoreactive ribonucleoside analogs, such as 4-thiouridine (4-SU) and 6-thioguanosine (6-SG), added to cell culture that is then incorporated into nascent RNA transcripts in-cell (Hafner et al., 2010). Crosslinking with UV radiation at 364 nm creates crosslinks between these photoreactive nucleosides and bound RBPs. Individual nucleotide resolution CLIP (iCLIP) takes advantage of reverse transcriptase (RT) falling off at a protein adduct left over from the protease digestion a majority of the time (Konig et al., 2010) . Circularization of the ssDNA products and relinearization at an incorporated restriction enzyme site allows for ligation of the 5' linker sequence following ssDNA synthesis. Sequencing libraries are then produced, sequenced and mapped at the nucleotide level nucleotide, allowing for RNA-protein interaction sites to be identified at high resolution. An illustrated flowchart of this process can be found in Figure 2.

As CLIP and related techniques has become more widely implemented over the last decade, over 100 proteins from multiple model systems have been assayed. These proteins can be found in the Appendix. Furthermore, research utilizing CLIP is now focusing on interplay between multiple proteins, such as splicing factors, to understand how levels of one protein perturb the RNA binding of another. For example, Pandit et al. used CLIP-seq to interrogate the effects of modulating the levels of SRSF1 and SRSF2 on the RNA binding of both these proteins, ultimately

elucidating multiple exons in which they contend for binding, and other in which they cooperate to affect alternative splicing (Pandit et al., 2013). Additionaly, Zarnack et al. focused on the antagonization of U2AF65 by hnRNP C with respect to *Alu* exonization. Knockdown of hnRNP C showed increased binding of U2AF65 to antisense *Alu* element RNA within pre-mRNA and use of internal 3' cryptic splice sites, leading to exonization of these repetitive sequences (Zarnack et al., 2013). Studies such as these will ultimately lead to understanding not only the protein-RNA interactions facilitate post-transcriptional gene regulation, but the competition/cooperation between trans-acting factors.

The oncofetal RNA binding protein IGF2BP3 and cancer

The Insulin Growth Factor 2 mRNA Binding Proteins (IGF2BP1, -2 and -3) are a family of structurally and functionally related RBPs with tissue-specific and developmentally regulated expression patterns (Hansen et al., 2004; Yaniv and Yisraeli, 2002). This subfamily of proteins originally identified as trans-acting factors that associate with IGF-II mRNA (Liao et al., 2005; Nielsen et al., 1999). These factors belong to the larger VICKZ family of RNA binding proteins, which are generally involved with RNA localization, RNA stability, and translation (Yisraeli, 2005). This family of factors possesses a unique combination of RNA-binding domains: an N-terminal containing two RNA recognition motifs (RRMs) and a C-terminal region with four hnRNP K-homology (KH) domains (Bell et al., 2013).

16

While the binding capacity of the RRMs has yet to be determined, the KH domains are thought to be essential in mediating RNA binding and mRNP formation (Wachter et al., 2013).

IGF2BP3 specifically is orthologous to the Xenopus Vegetal 1 RNA-binding protein (Vgl-RBP/Vera), which is important for Vgl mRNA subcellular localization and maintaining cell polarity in *Xenopus* oocytes (Nielsen et al., 2001). Developmentally, IGF2BP proteins are expressed during embryonic development and are responsible for nerve cell migration in *Xenopus* and *Drosophilia* (Boylan et al., 2008; Yaniv et al., 2003). Before birth in mammals, expression diminishes to negligible levels (occluding IGFBP2 in select tissues), suggesting that these proteins are mainly "oncofetal" (Hansen et al., 2004). Improper expression of IGF2BP proteins is also associated with a variety of cancers (Bell et al., 2013).

On the molecular level, what is known is that IGF2BP3 appears to regulate the mRNA metabolism outside of the nucleus. IGF2BP3 binds to the 5' UTR of IGF-II mRNA and appears to enhance its translation (Liao et al., 2005), while also binding to the 3'UTR of CD44, inhibiting its degradation (Vikesaa et al., 2006). IGF2BP3 also appears to bind to the 3' UTR of HMGA2 and repress Argonaute 2 binding and subsequent miRNA-mediated mRNA decay, creating what is described as a RNA "safe house" within the cytoplasm (Jonson et al., 2014). PAR-CLIP analysis of the IGF2BP protein family has offered insight as to thousands of potential mRNA targets of IGF2BP3, suggesting it and other IGF2BP proteins overlap in their respective targets and that binding appears to occur throughout the mRNAs with which they

associate (Hafner et al., 2010). However these targets have yet to be validated on any level. Furthermore, these experiments were performed in HEK 293 cells, which may not give a clear picture as to how IGF2BP3 functions in a disease context. Overall, these data suggest that IGF2BP3 plays an important role in post-transcriptional gene regulation in regulating localization, stability, and translation of target RNAs.

Conclusion of Introduction

In the following chapters, I will present work from two projects I have worked on during my graduate career of my dissertation will focus on exon identity, a process in with the 3' and 5' splice sites of exonic unit are defined for use by the spliceosome. I will focus determining the interplay of three classic splicing factors, hnRNP A1, SRSF1, and U2AF65 and with regards to 3' splice site definition on a transcriptome-wide scale. In the fourth and fifth chapters, I will focus on determining the RNA-binding specificity of and orphan RNA binding protein IGF2BP3 in pancreatic ductal adenocarcinoma and B-cell acute lymphocytic leukemia, as well as elucidating the possible roles of IGF2BP3 in RISC- (RNA-induced silencing) mediated gene regulation. Chapter 6 will be discuss the effects of human genetic mutations on exon identity. And my final chapter will discuss my findings on a broader scale as well as focus on future directions of these two projects.

Figure 1: Models of hnRNP A1-mediated splicing repression. Potential models by which hnRNP A1 is thought to regulate binding of additional splicing factors to promote exon skipping. These models universally suggest hnRNP A1 act primarily as an antagonizer of splicing enhancers through different, but not necessarily mutually exclusive, mechanisms.
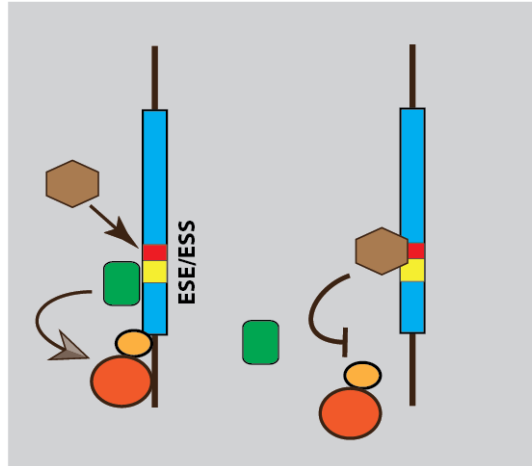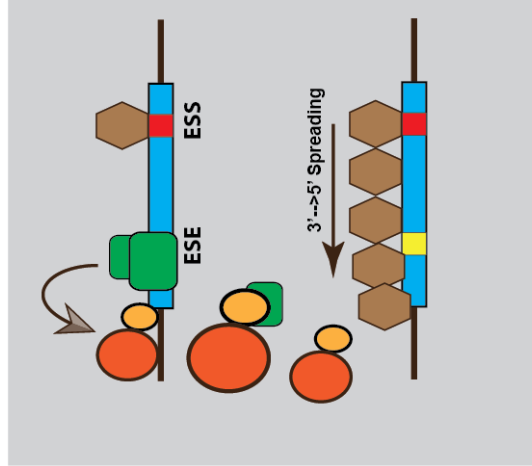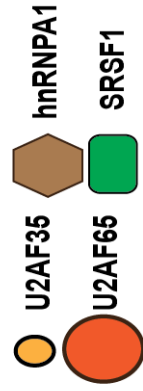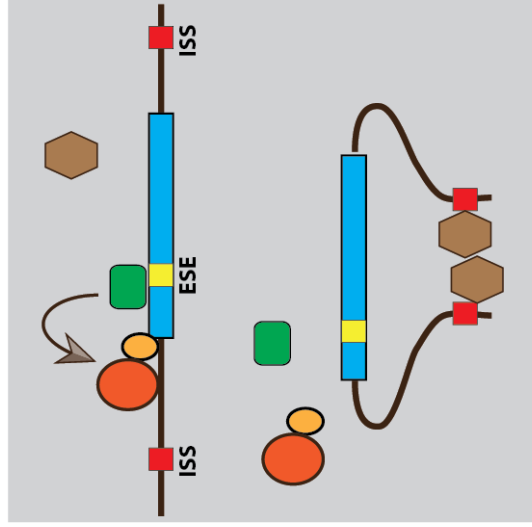
Figure 2: The iCLIP protocol. iCLIP begins with treatment of culctured cells with UV irradiation, creating irreversible covalent crosslinks at sites of protein–RNA interactions, preserving *in vivo* interactions (1). The cells are then lysed and samples are treated with nuclease to partially digested RNA and obtain RNA fragments in an optimal size (2). After immunoprecipitation of the protein–RNA complexes, the RNA is dephosphorylated using phosphatase, an adapter is ligated to the 3′ end of the RNA and the 5′ end is radioactively labeled T4 polynucleotide kinase (3-5). Samples are then run using MOPS-based SDS–PAGE to separate by molecular weight and then transferred to nitrocellulose membrane to purify crosslinked protein–RNA complexes filter out free RNA (6). The RNA is recovered from the nitrocellulose membrane by digesting the protein with proteinase K, (7) and then iCLIP RNA is reverse transcribed into cDNA, which most often truncates at the polypeptide remaining at the crosslink site (8). Urea-based PAGE is used for size selection of the cDNA rand is followed by cDNA circularization, which attaches the second adapter to the 3′ end of cDNA (9). Restriction enzyme digestion linearizes the cDNA before PCR amplification using an oligo that complements an internal BamHI site. Linearized cDNA is ready for PCR amplification and amplicon libraries are subjected to High-throughput sequencing (10). Adapted from (Huppertz et al., 2014)
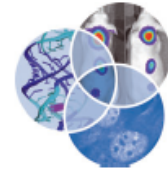
Individual Nucleotide resolution
Crosslinking Immunoprecipitation (iCLIP)
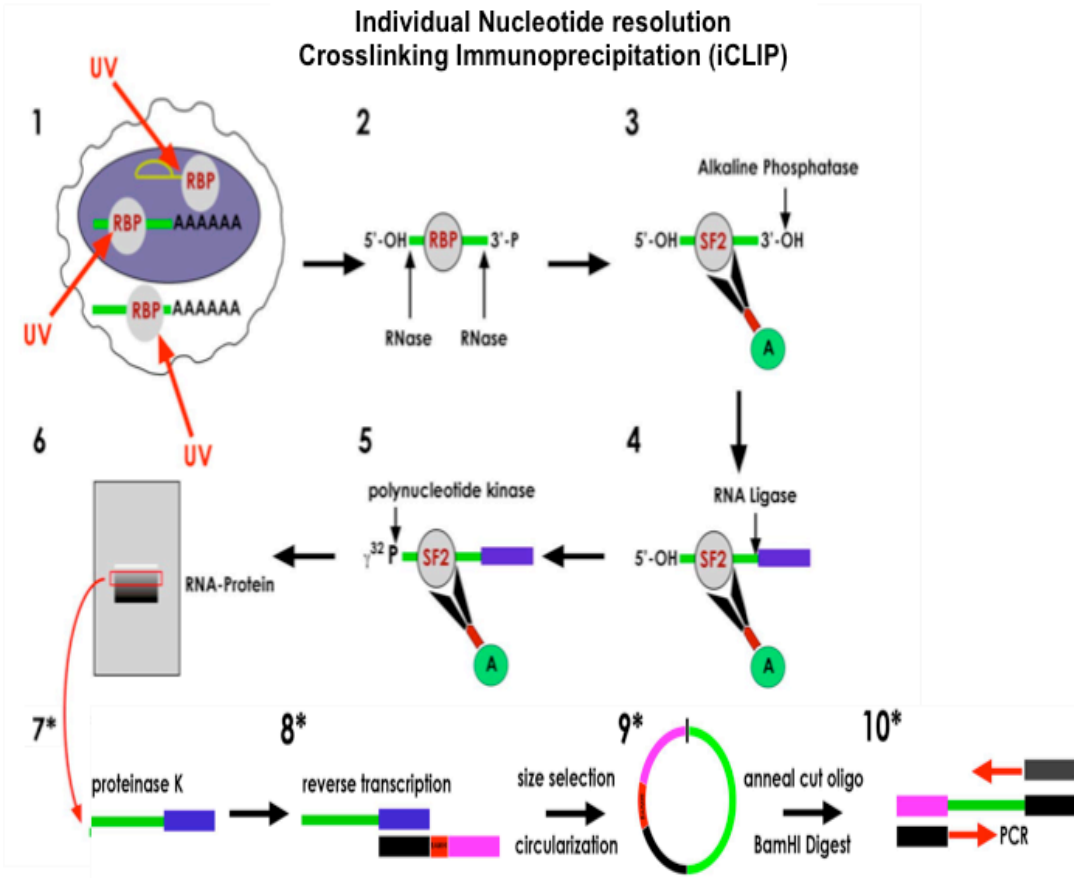
Chapter 2

The RNAissance family: SR proteins as multifaceted regulators

of gene expression

The RNAissance family: SR proteins as multifaceted regulators of gene expression

Jonathan M. Howard and Jeremy R. Sanford*

Serine and arginine-rich (SR) proteins play multiple roles in the eukaryotic gene expression pathway. Initially described as constitutive and alternative splicing factors, now it is clear that SR proteins are key determinants of exon identity and function as molecular adaptors, linking the pre-messenger RNA (pre-mRNA) to the splicing machinery. In addition, now SR proteins are implicated in many aspects of mRNA and noncoding RNA (ncRNA) processing well beyond splicing. These unexpected roles, including RNA transcription, export, translation, and decay, may prove to be the rule rather than the exception. To simply define, this family of RNA-binding proteins as splicing factors belies the broader roles of SR proteins in post-transcriptional gene expression. © 2014 John Wiley & Sons, Ltd.

INTRODUCTION

Post-transcriptional regulation is critical to the accurate expression of human genes. This process is overseen in part by a large superfamily of RS-domain-containing proteins found throughout metazoans,[1] which contain SR and 'SR-related' proteins (reviewed in Ref 2). Although it is often precarious to separate the two groups on a functional level, for the purposes of this review we will focus on the subfamily of 'classic' SR proteins. SR proteins are structurally defined as a family of RNA-binding proteins with a modular domain structure consisting of one to two amino-terminal RNA recognition motifs (RRMs) and a carboxyl-terminal domain (CTD) rich in serine and arginine dipeptide repeats.[3] There are 12 canonical members of the SR protein family that share this characteristic domain structure (see Table 1). SR proteins are intimately involved in the gene expression pathway, influencing both nuclear pre-messenger RNA (pre-mRNA) processing as well as the cytoplasmic fate of the mature RNA (mRNA)

* Correspondence to: jsanfor2@ucsc.edu

Department of Molecular, Cellular and Developmental Biology, University of California Santa Cruz, Santa Cruz, CA, USA

message. Prominently known for their requirement in spliceosome assembly and regulation of alternative splicing decisions, important distinctions in SR protein biology have emerged over the years. For example, some SR proteins have a life in the cytoplasm, whereas others remain confined in the nucleus. Here, we will provide a summary of the nuclear roles of SR proteins as well as their emergent postsplicing functions in gene expression.

From the beginning, the functional characterization of SR proteins alluded to both diverse and redundant activities. The founding member of the SR protein family, SRSF1, was identified and characterized by concurrent studies using biochemical complementation assays. Not only was SRSF1 seen to preferentially enhance usage of the proximal authentic 5′ splice site (ss) of a β-globin splicing reporter,[67] but it also altered splicing ratios of SV40 pre-mRNA, enhancing small T mRNA isoform production.[68] These data implicated the first 'classic' SR protein as a regulator of both constitutive and alternative splicing. A second SR protein, SFRS6, was shown to complement splicing-deficient extracts to promote β-globin splicing, as well as alternative 5′ ss usage in β-thalassemic pre-mRNA.[69,70] At roughly the same time, a third SR protein, SRSF2, was shown to influence splice site selection. Using RNase T1 protection and immunoprecipitation assays, SRSF2 was seen to

TABLE 1 | The SR Protein Family

Gene Symbol	Domain Structure	Protein Aliases	Shuttling	Molecular Functions	Biological Processes	References
SRSF1	RRM RRMH RS	SF2, ASF, SRp30A	Yes	Pre-mRNA splicing; mRNA export; translation; mRNA biogenesis; mRNA stability; NMD; transcriptional elongation	Apoptosis; cell cycle; senescence; cell growth proliferation; SUMOylation; genomic stability; cytoskeleton organization; embryogenesis; retinal development; cardiac development; cancer	4–14
SRSF2	RRM RS	SC35, SRp30B	No	Pre-mRNA splicing; genomic stability; transcriptional elongation	Cell survival; cell cycle; cancer; metastasis; senescence; apoptosis; development; neural plasticity; metabolism	15–23
SRSF3	RRM RS	SRp20	Yes	Pre-mRNA splicing; mRNA export; (viral) mRNA translation; transcriptional elongation	Cell adhesion and migration; cell cycle; cell proliferation; cellular senescence; aerobic glycolysis; neuronal survival and growth; apoptosis; glucose and lipid metabolism; cholesterol homeostasis; Long-term Memory formation; development; neurological disorders; cancer	24–34
SRSF4	RRM RRMH RS	SRp75	Yes	Pre-mRNA splicing	Neural differentiation	34
SRSF5	RRM RRMH RS	SRp40	No	Pre-mRNA splicing; (viral) mRNA translation	Insulin signaling; lipid transport; cell cycle; apoptosis; cancer; bipolar disorder	35–38
SRSF6	RRM RRMH RS	SRp55	Yes	Pre-mRNA splicing; (viral) mRNA translation	*Drosophila* development; cardiac development; eye development; apoptosis; wound healing; cell cycle; cytoskeleton organization; genomic integrity; angiogenesis; lipid transport; muscle development; calcium metabolism	37,39–52
SRSF7	RRM Zn RS	9G8	Yes	Pre-mRNA splicing; mRNA export; (viral) mRNA processing	Microtubules stabilization; viral infection	47,48

25

TABLE 1 | continued

Gene Symbol	Domain Structure	Protein Aliases	Shuttling	Molecular Functions	Biological Processes	References
SRSF8	RRM–RS	SRp46	ND	pre-mRNA splicing	N/A	53
SRSF9	RRM–RRMH–RS	SRp30c	ND	Pre-mRNA splicing mRNA translation	Glucocorticoid signaling; apoptosis; cell adhesion	49,54–56
SRSF10	RRM–RS	SRp38, SRrp40	Yes	Pre-mRNA splicing; mRNA translation	Stress response; neuronal differentiation; cholesterol biosynthesis; cell cycle	57–63
SRSF11	RRM–RS	p54, NET2	ND	Pre-mRNA splicing; genomic stability	Genomic integrity; ATP synthesis	64,65
SRSF12	RRM–RS	SRrp35	ND	Pre-mRNA splicing	Cell cycle	66

RRM, RNA recognition motif; RRMH, RNA recognition motif homology; RS, arginine/serine-rich motif; Zn, zinc-binding domain; mRNA, messenger RNA; miRNA, microRNA; NMD, nonsense-mediated decay.
Includes domain configuration of protein members, protein aliases, shuttling activities, reported molecular functions, and biological processes.

interact with both the 5′ ss and 3′ ss independently.[71] Furthermore, interactions between U1 and U2 small nuclear ribonucleoproteins (snRNPs) bound to 5′ ss and 3′ ss were shown to occur in an SRSF2-dependent manner, implicating SR proteins in early spliceosome architecture. SRSF2 was also shown to have similar effects on splice site selection as SRSF1, alluding to functional redundancy during spliceosome assembly.

The SR protein family was rapidly expanded through clever biochemical fractionation by Zahler et al., who co-purified a group of proteins (including SRSF1, SRSF2, and SRSF6) by ammonium sulfate fractionation and precipitation with magnesium chloride.[72] This approach revealed five proteins of various molecular weights, which were selectively purified from both HeLa cell extract and calf thymus. These proteins presented reactivity to mAb104, an antibody previously shown to recognize phosphorylated SR proteins.[69,73,74] Furthermore, four of these proteins were shown to rescue splicing of β-globin and *ftz* splicing reporters in splicing-deficient extracts, providing evidence that, like SRSF1 and SRSF2, these proteins were splicing factors.[72] Finally, microsequencing of these proteins showed highly similar amino acid compositions, as well as an abundance of serine/arginine dipeptides, on which their family name is based.[72]

REGULATION OF SR PROTEINS BY POST-TRANSLATIONAL MODIFICATION

Post-translational modification plays critical roles in regulation of SR protein activity and localization. Phosphorylation of SR proteins is regulated by the SR-specific protein kinase (SRPK) family and other CMGC kinase family members, such as Clk/Sty (cdc2-like kinase/serine, threonine, and tyrosine kinase).[75,76] These kinases share similar abilities to phosphorylate serine residues throughout the RS domain but differ in their specificity and mechanism of phosphorylation.[77–80] Dynamic phosphorylation of SR proteins is vital to the initiation and progression of spliceosome assembly to catalysis.[81–85] Mechanistically, it is thought that phosphorylation of the RS domain increases RNA-binding specificity[86] and is also important for specific protein–protein interactions within the prespliceosome.[87] Structurally, phosphorylation results in entropic reduction of the intrinsically disordered RS domain to promote more ordered side chains for molecular recognition.[88] Together, these data advocate that phosphorylation states of SR proteins act as 'molecular switches' during spliceosome assembly.

RS domain phosphorylation also influences the dynamics of SR protein localization in the cell. Release of SR proteins from nuclear speckles requires phosphorylation by the RS domain kinase Clk/Sty.[76] Following spliceosome assembly, SR proteins encounter one of two potential paths: for a subset of SR proteins, rephosphorylation by Clk/Sty and by nuclear SRPKs will target the protein for nuclear recycling for further rounds of pre-mRNA splicing, whereas other SR proteins remain dephosphorylated and associated with spliced mRNAs.[72,89−91] These studies suggest that dephosphorylated SR proteins retained on spliced mRNA may be a signal that an mRNA is ready for nuclear export. Following mRNA export and translation, SR proteins can then be rephosphorylated by cytoplasmic SRPKs, which facilitates interactions with transportin-SR and their import back into the nucleus.[92,93]

SR protein phosphorylation is further modulated in response to a variety of different cellular conditions and signals. Changes in phosphorylation and subcellular distribution of SR proteins accompany the global regulation of RNA metabolism during early development,[94,95] viral infection,[96] and cell cycle progression.[75] One recent example demonstrates SR protein phosphorylation as a direct result of epidermal growth factor (EGF) signaling. EGF signaling is shown to increase AKT activation, which in turn activates SRPK and subsequent upregulation of SR protein phosphorylation.[97] These data implicate SR proteins as integral players in propagating EGF signaling, which is linked to numerous human cancers.

SR proteins are modified by a variety of other marks including methylation and acetylation. The consequences of these modifications are less well understood than phosphorylation but they appear to be functionally relevant. For example, acetylation of SRSF2 occurs in response to genotoxic stress. Acetylation within the RRM domain correlates with pre-mRNA alternative splicing regulation of caspase-8, a factor involved in apoptosis.[98] Arginine methylation also influences SR protein localization and activity.[15] Blocking methylation affects alternative splicing, translation, and mRNA decay, most likely owing to misregulation of SR protein localization.[15,99,100]

THE COMPLEX ROLES OF SR PROTEINS IN PRE-mRNA SPLICING

The spliceosome is assembled *de novo* on each and every intronic substrate. This dynamic process involves the sequential recruitment and rearrangement of uracil-rich snRNP particles (U snRNPs).

SR proteins contribute to spliceosome assembly primarily through the recognition of exonic splicing enhancers (ESEs).[101−103] These interactions are particularly important during formation and stabilization of the early (E) complex[71,104−107] (Figure 1). E complex is defined by association of both the U1 snRNP and the heterodimeric splicing factor U2 snRNP auxiliary factor (U2AF) with the 5′ ss and 3′ ss, respectively. This step is mediated by phosphorylation-dependent interactions between the RS and RRM domains of ESE-bound SR proteins and the U1-70K at the 5′ ss[84,108] and the small subunit of U2AF (U2AF35) at the 3′ ss.[109−112] E complex is assumed to form on either end of an intron; however, the same interaction network can occur across exons in a process called exon definition (see below).

E complex is converted to A complex by the addition of the U2 snRNP. During A complex formation, SR proteins are thought to promote interactions of U2 snRNP with the branchpoint sequence through nonspecific interactions of the RS domain with the phosphodiester backbone, possibly neutralizing its negative charge and enhancing base pairing.[113] Additionally, SR proteins are implicated in recruitment of the U4/U6.U5 tri-snRNP,[114] forming a cross-exon 'B-like' complex, which can ultimately rearrange into cross-intron B complexes.[115] The RS domain of SR proteins (presumably not associated with ESEs) is also hypothesized to associate with the phosphodiester backbone near the 5′ ss to promote U6 binding.[116] Finally, extensive remodeling and rearrangement of RNA–RNA and RNA–protein interactions, coupled with dephosphorylation of SR proteins, results in formation of the catalytically active C complex.[81,117,118] In summary, SR proteins promote recruitment of multiple factors throughout spliceosome assembly, and are critical in formation of the final catalytic core.

SR proteins also play important roles in establishing exon–intron boundaries in large metazoan genes. The process of 'exon definition' is hypothesized to solve a significant problem related to finding relatively short exons within the context of long intronic sequences.[119] Exon definition occurs through a complex interaction network that links the 3′ ss at the 5′ end of the exon with the 5′ ss at the 3′ end of the exon (reviewed in Ref 120). In metazoans exon definition precedes intron definition in which 5′ ss and 3′ ss are paired during spliceosome assembly.[115] SR proteins also contribute to intron definition through a series of protein–protein interactions mediated by the RS domain linking U1 snRNP at the 5′ ss to U2AF35 at the 3′ ss.[107] Intron bridging has been alluded to through protein–protein interaction studies,
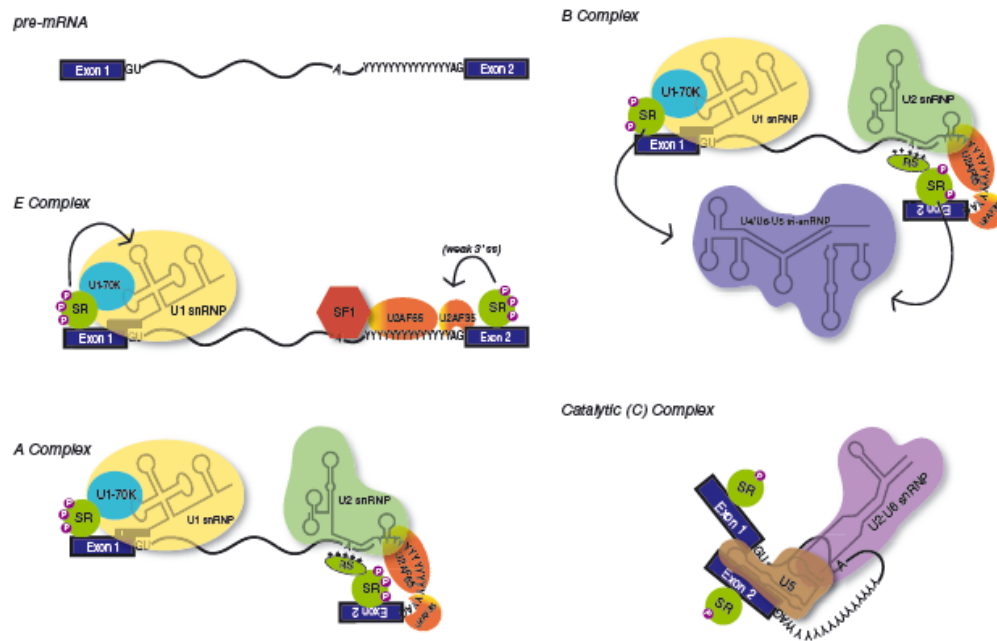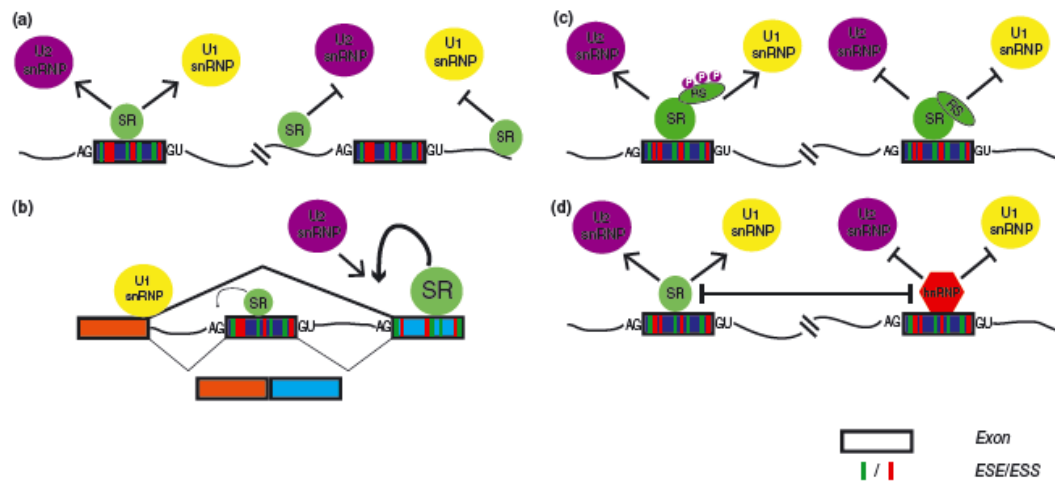
FIGURE 1 | Serine and arginine-rich (SR) proteins regulate spliceosome assembly. Spliceosome assembly onto the pre-messenger RNA (pre-mRNA) occurs in a coordinated, stepwise manner. In E complex, SR proteins regulate U1 snRNP recruitment to the 5′ splice site GU, and U2AF35/65 bound to the pyrimidine tract and 3′ splice site AG. In the A complex, SR proteins may facilitate U2 snRNP binding at the branchpoint by neutralizing the negative phosphodiester backbone charge. SR proteins can also recruit U4/U6-U5 tri-snRNP during B complex. Molecular rearrangements and dephosphorylation of SR proteins occur to form the catalytically active C complex, in which U2 and U6 interact, and U6 replaces U1 snRNP, and U5 coordinates exons prior to splicing and ligation. SF1, splicing factor 1; snRNP, small nuclear ribonucleoprotein; SR, SR protein; RS, arginine/serine motif; 5′ and 3′ splice sites are indicated by GU and AG dinucleotides, respectively; (Y)n, polypyrimidine tract; P, phosphate moiety].

and observed on splicing substrates using electron microscopy,[121] but the precise role of SR proteins is not well understood.

The mechanisms described above not only contribute to the roles of SR proteins in constitutive splicing but similarly in alternative splicing. The distinction between the two processes is simply the context in which SR proteins engage the pre-mRNA[112] (see Figure 2(a)). A general theme emerging from both *in vitro* and *in vivo* assays is that SR proteins act as enhancers of splicing when associated with exonic sequences, but function as silencers while binding to intronic sequences downstream of the 5′ ss.[101,122] However, this simplistic perspective belies the complex *cis*-regulatory landscape of most regulated exons. Several features distinguish alternative exons from constitutive exons, including their shorter length and weaker 5′ ss. Exonic splicing regulatory sequences (ESRs) are also more strongly conserved in the context of alternative exons when compared with their counterparts in constitutive exons,[123] reflecting requirements for ESRs in definition of suboptimal exons.[112] Remarkably, the

same ESR sequence can have opposing affects on splicing when placed into distinct positions within the same alternative exon.[123] These studies suggest that the regulatory roles of SR proteins in alternative splicing are most likely highly position and context dependent.

SR proteins can also have long-range effects on regulation of alternative exons. Several studies described a new mode for SR protein-mediated splicing regulation that occurs through their association with constitutive exons that are adjacent to alternative exons[124–126] (see Figure 2(b)). For example, SRSF1 has been shown to promote skipping of exon 16 in CamKIIδ through its association with downstream constitutive exon 17.[125] A similar mechanism influences splicing of the receptor tyrosine kinase MET, a key driver of malignant breast cancer.[127] In this context, elevated SRSF1 levels lead to increased skipping of exon 11, an effect mediated by ESEs located within exon 12.[128] These data suggest an intriguing model in which SR proteins may alter the competition between 3′ ss of adjacent exons with a common upstream 5′ ss.[123,124,129,130]

FIGURE 2 | Serine and arginine-rich (SR) proteins regulate alternative splicing. (a) SR proteins have been shown to promote or inhibit U1/U2 small nuclear ribonucleoprotein (snRNP) recruitment with respect to their orientation to 5' and 3' splice sites. (b) SR proteins bound to adjacent exons can compete for U2 snRNP recruitment to their respective 3' splice sites, likely depending on the 'strength' of the SR protein to recruit spliceosomal factors. (c) Phosphorylation states of the RS domain can influence SR protein-dependent recruitment of U1 and U2 snRNPs. (d) Antagonistic relationships of SR proteins and hnRNP proteins often influence recruitment of spliceosomal factors. ESE/ESS, exonic splicing enhancers/exonic splicing silencers.

Another important aspect of the mechanisms through which SR proteins influence alternative splicing involves their interplay with members of the heterogeneous nuclear ribonucleoprotein (hnRNP) family. The hnRNPs include several well-established splicing repressors, which mediate the repressive effects of exonic splicing silencers (ESSs). The functional antagonism of SR proteins and hnRNP proteins was first observed between SRSF1 and hnRNP A1 on several different alternative splicing modalities[131] (see Figure 2(d)). Not surprisingly, the underlying molecular mechanisms can be quite distinct. In the case of competing splice site donors, SRSF1 promoted selection of the proximal 5' ss (closest to the 3' ss), whereas hnRNP A1 promoted usage of more distal sites[132] by reducing binding of U1 snRNP at the proximal site. This functional antagonism also extends to alternative cassette exons. In an elegant series of experiments, Zhu et al. demonstrated that binding of an SR protein to an ESE inhibits the repressive effects of hnRNPs bound to adjacent silencers.[133] Because the relative expression levels of SR proteins and hnRNPs can vary dramatically across tissues and during tumorigenesis,[134,135] the complex functional interplay between hnRNPs and SR proteins is likely to play important roles in regulating patterns of alternate splicing across a wide array of conditions.

While often thought of as general splicing enhancers, there are also instances where SR proteins can inhibit splicing (see Figure 2(c)). For example, SRSF9 promotes skipping of exon 7B in the hnRNP A1 pre-mRNA.[136] This activity requires an intronic splicing silencer element located upstream of the exon 7B 3' ss. Similarly, the poorly characterized SRSF11 is reported to promote skipping of exon 10 of the Tau pre-mRNA by binding an ESS.[57] In contrast to these transcript-specific effects, SRSF10 functions as an inducible, global repressor of splicing.[137] SRSF10 activity is inhibited by phosphorylation-dependent interactions with 14-3-3 proteins. Conditions that promote activation of protein phosphatase 1, including heat shock and mitosis, lead to dephosphorylation of SRSF10, liberation from 14-3-3 proteins, and activation of splicing repressor activity.[59] Although the mechanisms of splicing inhibition are likely to be very different for each of these SR proteins, it is nonetheless intriguing that SR proteins are capable of having potentially opposite effects on splicing depending on their phosphorylation state[138] or the context in which they engage the pre-mRNA.

GLOBAL ANALYSIS OF SR PROTEIN RNA-BINDING SPECIFICITY

SR proteins are sequence-specific RNA-binding proteins. For most SR proteins, a putative consensus motif has been identified (reviewed in Ref 139), but the challenge now is to determine how these elements function

within different sequence contexts. Additionally, it is clear that there is significant functional redundancy in binding specificity.[140–142] These data imply that SR proteins may compete with each other for binding to closely related sites.

The first clues for understanding how this competition plays out on a global scale emerged from studies of SR proteins distribution on fixed insect polytene chromosomes and amphibian oocyte lampbrush chromosomes.[73,129] Imaging of nascent transcripts on *Chironomus tentans* polytene chromosomes revealed that SR proteins are distributed across the genome in a nonrandom pattern. Distinct SR protein staining patterns were observed at different loci, suggesting that different combinations of SR proteins are associated with nascent transcripts. High-resolution analysis of the Balbiani ring (BR) locus revealed that the BR mRNP is extensively remodeled as it is matured. Many of the SR proteins are replaced between the steps of mRNA export and translation, such that only SRSF1 is bound to polyribosome-associated mRNPs.[130] This observation is consistent with work from mammalian cells, which demonstrate that SR proteins are sorted on nascent transcripts through a phosphorylation-dependent mechanism.[143]

The high-throughput sequencing and crosslinking immunoprecipitation (HITS-CLIP) method allowed for global analysis of *in situ* protein–RNA interactions and provides key information such as consensus binding motifs, genome-wide binding-site distribution, and gene ontology of RNA targets.[144] HITS-CLIP analysis of SRSF1 revealed a diverse pool of RNA transcripts, including mRNA, microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), and intergenic transcripts of unknown function, advocating roles for SR proteins beyond pre-mRNA processing.[124,145] Subsequent studies confirmed many of these hypotheses, including interactions with long noncoding RNAs (ncRNAs) MALAT1 and *Xist*, as well as precursors of miRNA processing.[146–149] Furthermore, gene ontology analysis of SRSF1 mRNA targets showed an enrichment for RNA processing factors, suggesting a broad, highly integrated post-transcriptional network that governs splicing factor levels and ultimately global gene expression.[145] Similar to SRSF1, SRSF3 and SRSF4 engage a functionally diverse pool of RNA transcripts. However, their consensus binding sites and their CLIP tag distribution across transcripts are distinct.[150] Most intriguing was that CLIP tags for both SRSF3 and SRSF4 were enriched in ncRNAs, many of which have yet to be prescribed functions within the cell. Finally, SRSF3, but not SRSF4, was seen to regulate splicing of other splicing factors, further supporting the hypothesis of

regulatory cascades that may extend the roles of SR proteins beyond their known targets.[145,150,151]

Global studies of SR protein RNA target specificity demonstrate that most exons are bound by at least one SR protein.[126,130,145,150] A major challenge now is to understand how different combinations of SR proteins influence splicing of specific exons. Recent work from Pandit et al. suggests that this is likely to be a complex problem.[126] An initial comparison of SRSF1 and SRSF2 *in situ* binding sites in mouse embryo fibroblasts demonstrated that both proteins have considerable overlap in their binding specificity, suggesting that there may be competition for binding to related exon sequences. Interestingly, depletion of SRSF2 resulted in complex changes in SRSF1 binding-site occupancy. In some cases, SRSF1 binding increased in the absence of SRSF2, whereas the opposite pattern was observed at other locations. Together, these data suggest that SR proteins can play both redundant roles and cooperative roles in exon recognition.

EMERGING ROLES FOR SR PROTEINS IN GENE EXPRESSION

While it is generally accepted that the majority of splicing occurs in a co-transcriptional manner, only recently have these two processes been observed to directly regulate one another.[152,153] Live cell imaging initially showed dynamic recruitment of various splicing factors from nuclear speckles to sites of transcription activation.[154] Indeed, SR proteins colocalize with RNA polymerase II (Pol II) in nuclear speckles, an interaction mediated by the Pol II CTD[155] in a serine phosphorylation-dependent manner.[85,156] Truncation of the CTD prevents targeting of splicing factors to sites of transcription and markedly inhibits pre-mRNA splicing.[157] Also, selective mutations in the CTD cause diffusion of SR proteins away from nuclear speckles and accumulation of unspliced β-globin transcripts.[158] These data indicate that interactions between SR proteins and Pol II are involved in splicing regulation. Using a minigene spicing reporter, de la Mata and Kornblihtt showed that the CTD was required for SRSF3 recruitment and subsequent exon exclusion. Furthermore, the effect of the CTD on SRSF3-regulated alternative splicing was independent of transcription kinetics.[159] Together, these data imply that the CTD may play a direct role in spliceosome assembly through SR protein recruitment. Conversely, recent data imply that the association of SR proteins and Pol II may occur only after initiation of transcription. In the context of nascent FBJ murine osteogenic sarcoma RNA, association of various SR

proteins with Pol II was seen to be RNA dependent,[160] suggesting that recruitment of SR proteins to actively transcribed genes may not occur during initiation in all contexts. Additional experiments are needed to determine the mechanisms through which SR proteins regulate co-transcriptional alternative splicing. Regardless, these data provide a functional link between the processes of Pol II transcription and alternative splicing decisions mediated by SR proteins.

SR proteins may also directly regulate elongation rates of RNAPII. In general, depletion of either SRSF1 or SRSF2 has global effects on Pol II transcription in cells, and SRSF2 levels have been shown to affect the accumulation of Pol II at gene loci.[161] Mechanistically, SRSF2 is thought to enhance the release of transcriptional regulator TEFb from 7SK RNA owing to emergence of an SRSF2-recognized ESE following initial Pol II elongation. This may induce SRSF2 to switch from the 7SK RNA to nascent RNA, triggering the coordinated release of P-TEFb from the 7SK complex, and subsequent phosphorylation and unpausing of Pol II.[162] This suggests that some SR proteins may have direct effects in recruitment of Pol II factors to initiated Pol II complexes to facilitate elongation.

Compartmentalization of genetic material in the nucleus allows for separation of mRNA transcription from its fate as a template for protein synthesis. The discovery that a subset of SR proteins shuttle between the nucleus and cytoplasm (reviewed in Ref 163) immediately suggested that SR proteins might remain bound with their mRNA targets beyond pre-mRNA splicing (Figure 3). At least one non-canonical function for shuttling SR proteins appears to be in mRNA export pathways. Specific roles for the SR proteins SRSF3 and SRSF7 in intronless histone mRNA export provided the first direct evidence for SR protein moonlighting.[164] Shuttling SR proteins can interact with the canonical mRNA export factor nuclear RNA export factor 1 (NXF1; also known as TAP).[165,166] These data suggest that SR proteins may function broadly in the export of both spliced and unspliced mRNAs.[164,167,168] Interactions between NXF1 and SR proteins require dephosphorylation of the RS domain, suggesting an elegant mechanism for signaling the completion of an export-ready mRNA.[166] Surprisingly, depletion of specific SR proteins does not induce general defects in mRNA export,[143] suggesting that SR proteins may function redundantly or play roles in nuclear export of specific mRNAs.

The nucleocytoplasmic shuttling SR protein SRSF1 is readily detectable in the polyribosome fraction of cultured human cells suggesting that SR proteins may be involved in mRNA translation.[169]

This hypothesis was confirmed through three different functional assays. Perhaps most importantly, nonshuttling mutants of SRSF1 failed to enhance mRNA translation.[91,169] Subsequent experiments demonstrated that SRSF1 stimulates translation initiation through a novel mechanism involving the mechanistic target of rapamycin (mTOR) complex.[170] These data support an intriguing model whereby SRSF1 functions as an adaptor protein linking specific mRNA transcripts to translational control by mTOR (see Figure 4(a)). Recent work from the Caceres laboratory dramatically extends this model with the identification of >500 mRNAs that are likely to be translationally controlled by SRSF1 and mTOR.[171]

Moreover, other shuttling SR proteins including SRSF3 and SRSF7 are implicated in translational control. SRSF3 and SRSF7 mediate the effects of two distinct cis-regulatory elements including a viral internal ribosome entry site (IRES) and cellular constitutive transport elements (CTEs)[168,172] (see Figure 4(b)). Similarly, SRSF5 and SRSF6 enhance translation of gag protein from unspliced HIV-1 RNA, an activity that depends on their ability to shuttle from the nucleus to the cytoplasm.[173] SR proteins also have the potential to repress translation. During Xenopus development SRSF10 has been shown to interact with the peptidyltransferase center of 28S rRNA in undifferentiated neural cells.[174] Furthermore, this mechanism may help neuronal stem cells to maintain an undifferentiated state. These data paint a larger role for SR proteins in translation through regulating interactions with their respective RNA targets and translational machinery.

The roles of SRSF1 in mRNA translation suggest an intriguing hypothesis that the fates of mRNA isoforms generated by alternative splicing may be subject to differential translation.[58,169,175] This idea is supported by the recent observation that >30% of alternative mRNA isoforms exhibit differential polyribosome association.[176] Shuttling SR proteins, such as SRSF1, which remain associated with its targets throughout the gene expression pathway, are likely to contribute to this mechanism.[124,143] Overexpression of SRSF1 results in isoform-specific recruitment of mRNAs to polyribosomes, suggesting a direct role in coordinating the alternative fates of mRNA isoforms.[171]

In addition to splicing, export, and translation, SR proteins also influence mRNA stability.[177] This activity can occur through several different mechanisms. First, SR proteins regulate alternative splicing and in many cases, such as post-transcriptional control of splicing factor levels, alternative splicing generates isoforms that are inherently unstable.[151,178] Unstable
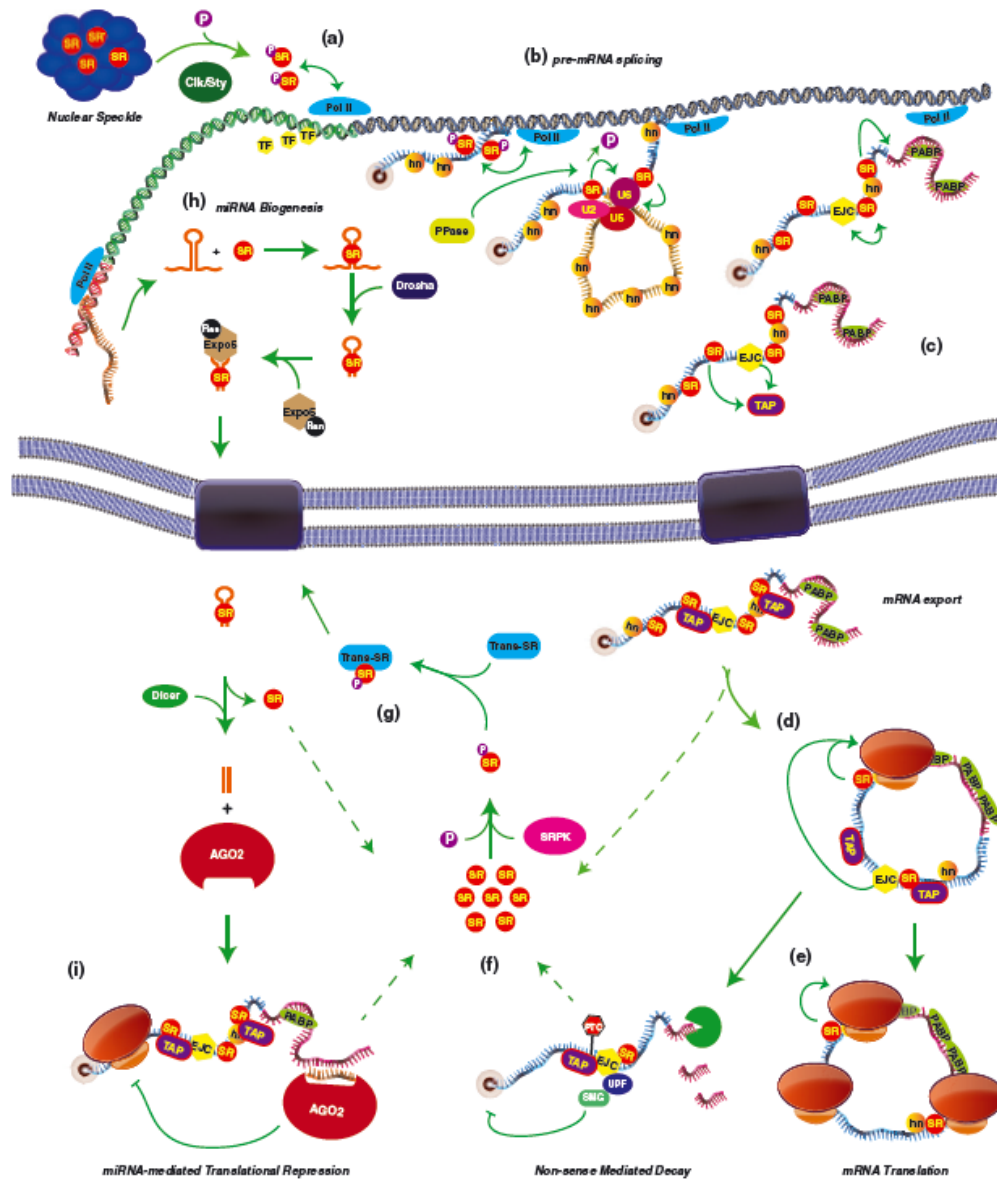
FIGURE 3 | The life cycle of a serine and arginine-rich (SR) protein. (a) SR proteins remain localized to nuclear speckles until they are phosphorylated by Clk/Sty. At this point they can be recruited to areas of active transcription, possibly in Pol II-dependent manner. (b) SR proteins can then bind to splicing enhancers in the nascent pre-messenger RNA (pre-mRNA) transcript to facilitate spliceosome assembly co-transcriptionally in phosphorylation-dependent manner. (c) Following maturation of the mRNA transcript, SR proteins, along with other factors (e.g., EJC proteins), can facilitate TAP binding to the mRNP and subsequent nuclear export. (d) After export, SR proteins can enhance the pioneering round of mRNA translation and send the transcript down one of two pathways: (e) the ribosome encounters no pretermination codons (PTCs) and continues with steady-state translation or (f) a PTC is encountered and nonsense-mediated decay proceeds. (g) Released SR proteins can then be phosphorylated by cytoplasmic SRPK, which triggers binding of transportin-SRs and nuclear import of SR proteins for storage or further rounds of splicing. (h) SR proteins may also play roles in miRNA biogenesis by facilitating export of pre-microRNAs to the cytoplasm for further processing and use in RNA-induced silencing (i). SR, SR protein; P, phosphate moiety; TF, transcription factor; Pol II, RNA polymerase II; hn, hnRNP proteins; U2, U5, U6, U snRNPs; PPase, protein phosphatase; EJC, exon junction complex; PABP, poly-A binding protein; TAP, TAP/nuclear export factor 1; Exo, exosome; SRPK, SR protein kinase; Trans-SR, transportin-SR; Expo5, exportin 5; AGO2, argonaute 2.
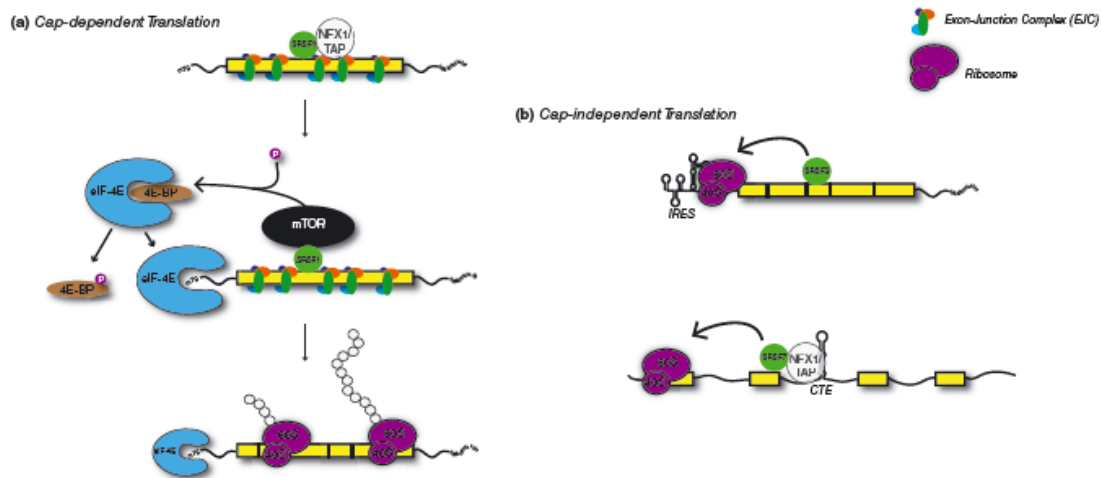
FIGURE 4 | Serine and arginine-rich (SR) proteins function in translation initiation. (a) SRSF1 bound to exported mRNAs can associate with mechanistic target of rapamycin (mTOR) kinase and recruit it to cytoplasmic mRNP complexes. This facilitates phosphorylation of 4E-BP, causing dissociation from eIF4E and increasing the efficiency of cap-dependent translation initiation. (b) SR proteins have also been shown to enhance cap-independent translation initiation of viral RNAs that contain internal ribosome entry site (IRES) elements and constitutive transport elements (CTEs).

isoforms contain premature termination codons that trigger the nonsense-mediated decay (NMD) RNA surveillance pathway. In contrast to this splicing-based mode of gene regulation, SR proteins have also been shown to directly enhance NMD.[179] Intriguingly, this study showed that the RS domain is required for augmentation of NMD, but not SR protein shuttling activity. These data suggest that SR proteins stimulate a rate-limiting step in the nucleus or that SR proteins may regulate the expression of NMD factors. Alternatively, SRSF1 may stimulate the pioneer round of mRNA translation leading to more efficient NMD.[180] Taken together, there is little doubt that SR proteins have complex effects on transcript stability.

In contrast to their roles in pre-mRNA splicing, relatively little is known concerning the molecular mechanisms through which SR proteins influence postsplicing steps of gene expression. One hypothesis is that shuttling SR proteins work in concert with the exon junction complex (EJC) to influence mRNA export,[165,181] stability,[177] and translation.[169,170,182] The EJC is deposited near exon–exon junctions as a result of pre-mRNA splicing and regulates post-transcriptional control of mature mRNAs. Proteomic analysis of EJC factors revealed numerous RNA-independent interactions with SR proteins. This observation is in good agreement with analysis of EJC RNA footprints, which revealed a myriad of noncanonical binding sites (i.e., those not centered 24 nucleotides upstream of an exon–exon junction).

A significant proportion of these footprints overlapped with ESEs, which are often occupied by SR proteins. EJC–SR protein interactions appear to be functionally significant as SRSF1 and SRSF3 exhibit reduced mRNA-binding activity when the EJC factor eIF4AIII is depleted from cells.[183] Together, these data reveal extensive, cooperative associations between SR proteins and the EJC in mRNP biogenesis and may explain their functional redundancy in regulation of mRNA export,[165,181] translation,[169,170,182] and decay,[180,181] as well as maintenance of genomic stability.[4,184]

The roles of SR proteins in gene regulation extend beyond mRNA processing. SR proteins have recently been attributed functional roles in miRNA biogenesis. Specifically, SRSF1 has been found to associate with primary-miR-7 transcript through a putative SRSF1-binding site in the stem loop. This interaction promotes cleavage of the pre-miR by the microprocessor complex protein Drosha.[149] The role for SRSF1 in miRNA biogenesis appears to be direct, as it is independent of its role in splicing. HITS-CLIP analysis of SRSF1, SRSF3, and SRSF4 suggests that most SR proteins interact with a small but distinct group of miRNAs, suggesting that shutting SR proteins are involved in miRNA biogenesis on a more general level.[145,150] Overall, these data highlight a potential coordination between splicing regulation and miRNA-mediated transcriptome regulation.

CONCLUSION

The characteristics of SR proteins mirror another general regulator of nucleic acid structure and function: histones. Similar to histones, SR proteins control the accessibility of their nucleic acid targets to the gene expression machinery. The similarities extend to their biochemical properties as well. Both are highly alkaline, associate with, and regulate the use of their respective bound nucleic acids, and can form homo- and hetero-oligomers to package DNA/RNA within the cell. Also, both histones and SR proteins are extensively post-translationally modified, which can control the functionality of the nucleic acid to which they are bound. Furthermore, both sets of proteins are used as the foundational basis for recruitment of additional factors to their respective nucleic acids to accomplish biochemical work, whether it be DNA-binding complexes that regulate and catalyze transcription or RNA-binding complexes that regulate pre-mRNA splicing, mRNA export, translation, and degradation. Similar comparisons have been made of hnRNP oligomers that have the ability to wrap up RNA species to form mRNPs

that share some resemblance to nucleosomes.[185,186] Clearly, this hypothesis warrants further investigation given the intimate association of SR proteins with virtually every aspect of post-transcriptional gene regulation.

The coming years will undoubtedly see an explosion in data utilizing high-throughput assays (e.g., HITS-CLIP and iCLIP) to determine the transcriptome-wide RNA interaction networks of SR proteins. These studies will provide a more general overview as to what RNAs SR proteins associate with and how they bind to them. This will certainly solidify the notion that SR proteins function in all aspects of RNA metabolism and gene expression rather than just splicing. The challenge for future work is to begin to determine how fluctuations in the levels of SR proteins influence the binding specificity of other SR proteins and splicing factors globally.[126,187–189] These types of experiments will elucidate the context-specific interactions that determine how exon identity is established in living cells as well as other RBP 'codes' as they function in downstream steps of gene expression. Overall, the near future holds a greater understanding of how SR proteins govern the RNA world.

REFERENCES

1. Boucher L, Ouzounis CA, Enright AJ, Blencowe BJ. A genome-wide survey of RS domain proteins. *RNA* 2001, 7:1693–1701.

2. Blencowe BJ, Bowman JA, McCracken S, Rosonina E. SR-related proteins and the processing of messenger RNA precursors. *Biochem Cell Biol* 1999, 77:277–291.

3. Manley JL, Krainer AR. A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). *Genes Dev* 2010, 24:1073–1074.

4. Li X, Manley JL. Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell* 2005, 122:365–378.

5. Pelisch F, Gerez J, Druker J, Schor IE, Munoz MJ, Risso G, Petrillo E, Westman BJ, Lamond AI, Arzt E, et al. The serine/arginine-rich protein SF2/ASF regulates protein sumoylation. *Proc Natl Acad Sci U S A* 2010, 37:16119–16124.

6. Shimoni-Sebag A, Lebenthal-Loinger I, Zender L, Karni R. RRM1 domain of the splicing oncoprotein

SRSF1 is required for MEK1-MAPK-ERK activation and cellular transformation. *Carcinogenesis* 2013, 11:2498–2504.

7. Fregoso OI, Das S, Akerman M, Krainer AR. Splicing-factor oncoprotein SRSF1 stabilizes p53 via RPL5 and induces cellular senescence. *Mol Cell* 2013, 1:56–66.

8. Das S, Anczukow O, Akerman M, Krainer AR. Oncogenic splicing factor SRSF1 is a critical transcriptional target of MYC. *Cell Rep* 2012, 2:110–117.

9. Karni R, de Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nat Struct Mol Biol* 2007, 3:185–193.

10. Kanadia RN, Clark VE, Punzo C, Trimarchi JM, Cepko CL. Temporal requirement of the alternative-splicing factor Sfrs1 for the survival of retinal neurons. *Development* 2008, 23:3923–3933.

11. Gabut M, Dejardin J, Tazi J, Soret J. The SR family proteins B52 and dASF/SF2 modulate development

of the Drosophila visual system by regulating specific RNA targets. *Mol Cell Biol* 2007, 28:3087–3097.

12. Li X, Wang J, Manley JL. Loss of splicing factor ASF/SF2 induces G2 cell cycle arrest and apoptosis, but inhibits internucleosomal DNA fragmentation. *Genes Dev* 2005, 22:2705–2714.

13. Xu X, Yang D, Ding JH, Wang W, Chu PH, Dalton ND, Wang HY, Bermingham JR Jr, Ye Z, Liu F, et al. ASF/SF2-regulated CaMKIIδ alternative splicing temporally reprograms excitation-contraction coupling in cardiac muscle. *Cell* 2005, 1:59–72.

14. Longman D, Johnstone IL, Caceres JF. Functional characterization of SR and SR-related genes in Caenorhabditis elegans. *EMBO J* 2000, 7:1625–1637.

15. Sinha R, Allemand E, Zhang Z, Karni R, Myers MP, Krainer AR. Arginine methylation controls the subcellular localization and functions of the oncoprotein splicing factor SF2/ASF. *Mol Cell Biol* 2010, 30:2762–2774.

16. Edmond V, Merdzhanova G, Gout S, Brambilla E, Gazzeri S, Eymin B. A new function of the splicing factor SRSF2 in the control of E2F1-mediated cell cycle progression in neuroendocrine lung tumors. *Cell Cycle* 2013, 12:1267–1278.

17. Apostolatos A, Song S, Acosta S, Peart M, Watson JE, Bickford P, Cooper DR, Patel NA. Insulin promotes neuronal survival via the alternatively spliced protein kinase CδII isoform. *J Biol Chem* 2012, 12:9299–9310.

18. Sharma S, Liao W, Zhou X, Wong DT, Lichtenstein A. Exon 11 skipping of E-cadherin RNA downregulates its expression in head and neck cancer cells. *Mol Cancer Ther* 2011, 9:1751–1759.

19. Edmond V, Brambilla C, Brambilla E, Gazzeri S, Eymin B. SRSF2 is required for sodium butyrate-mediated p21(WAF1) induction and premature senescence in human lung carcinoma cell lines. *Cell Cycle* 2011, 12:1968–1977.

20. Xiao R, Sun Y, Ding JH, Lin S, Rose DW, Rosenfeld MG, Fu XD, Li X. Splicing regulator SC35 is essential for genomic stability and cell proliferation during mammalian organogenesis. *Mol Cell Biol* 2007, 15:5393–5402.

21. Meshorer E, Bryk B, Toiber D, Cohen J, Podoly E, Dori A, Soreq H. SC35 promotes sustainable stress-induced alternative splicing of neuronal acetylcholinesterase mRNA. *Mol Psychiatry* 2005, 11:985–997.

22. Gabut M, Mine M, Marsac C, Brivet M, Tazi J, Soret J. The SR protein SC35 is responsible for aberrant splicing of the E1α pyruvate dehydrogenase mRNA in a case of mental retardation with lactic acidosis. *Mol Cell Biol* 2005, 8:3286–3294.

23. Crovato TE, Egebjerg J. ASF/SF2 and SC35 regulate the glutamate receptor subunit 2 alternative flip/flop splicing. *FEBS Lett* 2005, 19:4138–4144.

24. Kim J, Park RY, Chen JK, Kim J, Jeong S, Ohn T. Splicing factor SRSF3 represses the translation of programmed cell death 4 mRNA by associating with the 5′-UTR region. *Cell Death Differ* 2014, 3:481–490.

25. Kano S, Nishida K, Nishiyama C, Akaike Y, Kajita K, Kurokawa K, Masuda K, Kuwano Y, Tanahashi T, Rokutan K. Truncated serine/arginine-rich splicing factor 3 accelerates cell growth through up-regulating c-Jun expression. *J Med Invest* 2013, 3–4:228–235.

26. Sen S, Jumaa H, Webster NJ. Splicing factor SRSF3 is crucial for hepatocyte differentiation and metabolic function. *Nat Commun* 2013, 4:1336.

27. Wang Z, Chatterjee D, Jeon HY, Akerman M, Vander Heiden MG, Cantley LC, Krainer AR. Exon-centric regulation of pyruvate kinase M alternative splicing via mutually exclusive exons. *J Mol Cell Biol* 2012, 2:79–87.

28. Jia R, Li C, McCoy JP, Deng CX, Zheng ZM. SRp20 is a proto-oncogene critical for cell proliferation and tumor induction and maintenance. *Int J Biol Sci* 2010, 6:806–826.

29. Goncalves V, Matos P, Jordan P. Antagonistic SR proteins regulate alternative splicing of tumor-related Rac1b downstream of the PI3-kinase and Wnt pathways. *Hum Mol Genet* 2009, 18:3696–3707.

30. Watanuki T, Funato H, Uchida S, Matsubara T, Kobayashi A, Wakabayashi Y, Otsuki K, Nishida A, Watanabe Y. Increased expression of splicing factor SRp20 mRNA in bipolar disorder patients. *J Affect Disord* 2008, 1–2:62–69.

31. Jumaa H, Wei G, Nielsen PJ. Blastocyst formation is blocked in mouse embryos lacking the splicing factor SRp20. *Curr Biol* 1999, 16:899–902.

32. Antunes-Martins A, Mizuno K, Irvine EE, Lepicard EM, Giese KP. Sex-dependent up-regulation of two splicing factors, Psf and Srp20, during hippocampal memory formation. *Learn Mem* 2007, 10:693–702.

33. Corbo C, Orru S, Salvatore F. SRp20: an overview of its role in human diseases. *Biochem Biophys Res Commun* 2013, 1:1–5.

34. Anko ML, Morales L, Henry I, Beyer A, Neugebauer KM. Global analysis reveals SRp20- and SRp75-specific mRNPs in cycling and neural cells. *Nat Struct Mol Biol* 2010, 8:962–970.

35. Lu C, Li JY, Ge Z, Zhang L, Zhou GP. Par-4/THAP1 complex and Notch3 competitively regulated pre-mRNA splicing of CCAR1 and affected inversely the survival of T-cell acute lymphoblastic leukemia cells. *Oncogene* 2013, 50:5602–5613.

36. Akula N, Barb J, Jiang X, Wendland JR, Choi KH, Sen SK, Hou L, Chen DT, Laje G, Johnson K, et al. RNA-sequencing of the brain transcriptome implicates dysregulation of neuroplasticity, circadian rhythms and GTPase binding in bipolar disorder. *Mol Psychiatry* 2014. doi: 10.1038/mp.2013.170.

37. Dance GS, Sowden MP, Cartegni L, Cooper E, Krainer AR, Smith HC. Two proteins essential for apolipoprotein B mRNA editing are expressed from a single gene through alternative splicing. *J Biol Chem* 2002, 15:12703–12709.

38. Patel NA, Kaneko S, Apostolatos HS, Bae SS, Watson JE, Davidowitz K, Chappell DS, Birnbaum MJ, Cheng JQ, Cooper DR. Molecular and genetic studies imply Akt-mediated signaling promotes protein kinase CβII alternative splicing via phosphorylation of serine/arginine-rich splicing factor SRp40. *J Biol Chem* 2005, 14:14302–14309.

39. Jin W, Cote GJ. Enhancer-dependent splicing of FGFR1 α-exon is repressed by RNA interference-mediated down-regulation of SRp55. *Cancer Res* 2004, 24:8901–8905.

40. Juge F, Fernando C, Fic W, Tazi J. The SR protein B52/SRp55 is required for DNA topoisomerase I recruitment to chromatin, mRNA release and transcription shutdown. *PLoS Genet* 2010, 9:e1001124.

41. Lal S, Allan A, Markovic D, Walker R, Macartney J, Europe-Finner N, Tyson-Capper A, Grammatopoulos DK. Estrogen alters the splicing of type 1 corticotropin-releasing hormone receptor in breast cancer cells. *Sci Signal* 2013, 282:ra53.

42. Ring HZ, Lis JT. The SR protein B52/SRp55 is essential for Drosophila development. *Mol Cell Biol* 1994, 11:7499–7506.

43. Shukla S, Dirksen WP, Joyce KM, Le Guiner-Blanvillain C, Breathnach R, Fisher SA. TIA proteins are necessary but not sufficient for the tissue-specific splicing of the myosin phosphatase targeting subunit 1. *J Biol Chem* 2004, 14:13668–13676.

44. Tran Q, Coleman TP, Roesser JR. Human transformer 2β and SRp55 interact with a calcitonin-specific splice enhancer. *Biochim Biophys Acta* 2003, 2:141–152.

45. Tran Q, Roesser JR. SRp55 is a regulator of calcitonin/CGRP alternative RNA splicing. *Biochemistry* 2003, 4:951–957.

46. Ramchatesingh J, Zahler AM, Neugebauer KM, Roth MB, Cooper TA. A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer. *Mol Cell Biol* 1995, 9:4898–4907.

47. Lopez-Mejia IC, Vautrot V, De Toledo M, Behm-Ansmant I, Bourgeois CF, Navarro CL, Osorio FG, Freije JM, Stevenin J, De Sandre-Giovannoli A, et al. A conserved splicing mechanism of the LMNA gene controls premature aging. *Hum Mol Genet* 2011, 23:4540–4555.

48. Valente ST, Gilmartin GM, Venkatarama K, Arriagada G, Goff SP. HIV-1 mRNA 3′ end processing is distinctively regulated by eIF3f, CDK11, and splice factor 9G8. *Mol Cell* 2009, 2:279–289.

49. Jain A, Wordinger RJ, Yorio T, Clark AF. Spliceosome protein (SRp) regulation of glucocorticoid receptor isoforms and glucocorticoid response in human trabecular meshwork cells. *Invest Ophthalmol Vis Sci* 2012, 2:857–866.

50. Fic W, Juge F, Soret J, Tazi J. Eye development under the control of SRp55/B52-mediated alternative splicing of eyeless. *PLoS One* 2007, 2:e253.

51. Hara H, Takeda T, Yamamoto N, Furuya K, Hirose K, Kamiya T, Adachi T. Zinc-induced modulation of SRSF6 activity alters Bim splicing to promote generation of the most potent apoptotic isoform BimS. *FEBS J* 2013, 14:3313–3327.

52. Jensen MA, Wilkinson JE, Krainer AR. Splicing factor SRSF6 promotes hyperplasia of sensitized skin. *Nat Struct Mol Biol* 2014, 2:189–197.

53. Soret J, Gattoni R, Guyon C, Sureau A, Popielarz M, Le Rouzic E, Dumon S, Apiou F, Dutrillaux B, Voss H, et al. Characterization of SRp46, a novel human SR splicing factor encoded by a PR264/SC35 retropseudogene. *Mol Cell Biol* 1998, 8:4924–4934.

54. Xu Q, Leung DY, Kisich KO. Serine-arginine-rich protein p30 directs alternative splicing of glucocorticoid receptor pre-mRNA to glucocorticoid receptor β in neutrophils. *J Biol Chem* 2003, 29:27112–27118.

55. Fu Y, Huang B, Shi Z, Han J, Wang Y, Huangfu J, Wu W. SRSF1 and SRSF9 RNA binding proteins promote Wnt signalling-mediated tumorigenesis by enhancing β-catenin biosynthesis. *EMBO Mol Med* 2013, 5:737–750.

56. Cloutier P, Toutant J, Shkreta L, Goekjian S, Revil T, Chabot B. Antagonistic effects of the SRp30c protein and cryptic 5′ splice sites on the alternative splicing of the apoptotic regulator Bcl-x. *J Biol Chem* 2008, 31:21315–21324.

57. Wu JY, Kar A, Kuo D, Yu B, Havlioglu N. SRp54 (SFRS11), a regulator for tau exon 10 alternative splicing identified by an expression cloning strategy. *Mol Cell Biol* 2006, 26:6739–6747.

58. Blaustein M, Pelisch F, Tanos T, Munoz MJ, Wengier D, Quadrana L, Sanford JR, Muschietti JP, Kornblihtt AR, Caceres JF, et al. Concerted regulation of nuclear and cytoplasmic activities of SR proteins by AKT. *Nat Struct Mol Biol* 2005, 12:1037–1044.

59. Shi Y, Manley JL. A complex signaling pathway regulates SRp38 phosphorylation and pre-mRNA splicing in response to heat shock. *Mol Cell* 2007, 28:79–90.

60. Shi Y, Nishida K, Campigli Di Giammartino D, Manley JL. Heat shock-induced SRSF10 dephosphorylation displays thermotolerance mediated by Hsp27. *Mol Cell Biol* 2011, 3:458–465.

61. Ling IF, Estus S. Role of SFRS13A in low-density lipoprotein receptor splicing. *Hum Mutat* 2010, 6:702–709.

62. Komatsu M, Kominami E, Arahata K, Tsukahara T. Cloning and characterization of two neural-salient serine/arginine-rich (NSSR) proteins involved in the

regulation of alternative splicing in neurones. *Genes Cells* 1999, 10:593–606.

63. Feng Y, Valley MT, Lazar J, Yang AL, Bronson RT, Firestein S, Coetzee WA, Manley JL. SRp38 regulates alternative splicing and is required for Ca(2+) handling in the embryonic heart. *Dev Cell* 2009, 4:528–538.

64. Straub T, Grue P, Uhse A, Lisby M, Knudsen BR, Tange TO, Westergaard O, Boege F. The RNA-splicing factor PSF/p54 controls DNA-topoisomerase I activity by a direct interaction. *J Biol Chem* 1998, 41:26261–26264.

65. Sakashita E, Tatsumi S, Werner D, Endo H, Mayeda A. Human RNPS1 and its associated factors: a versatile alternative pre-mRNA splicing regulator in vivo. *Mol Cell Biol* 2004, 3:1174–1187.

66. Cowper AE, Caceres JF, Mayeda A, Screaton GR. Serine-arginine (SR) protein-like factors that antagonize authentic SR proteins and regulate alternative splicing. *J Biol Chem* 2001, 52:48908–48914.

67. Krainer AR, Conway GC, Kozak D. The essential pre-mRNA splicing factor SF2 influences 5' splice site selection by activating proximal sites. *Cell* 1990, 62:35–42.

68. Ge H, Manley JL. A protein factor, ASF, controls cell-specific alternative splicing of SV40 early pre-mRNA in vitro. *Cell* 1990, 62:25–34.

69. Roth MB, Zahler AM, Stolk JA. A conserved family of nuclear phosphoproteins localized to sites of polymerase II transcription. *J Cell Biol* 1991, 115:587–596.

70. Mayeda A, Zahler AM, Krainer AR, Roth MB. Two members of a conserved family of nuclear phosphoproteins are involved in pre-mRNA splicing. *Proc Natl Acad Sci U S A* 1992, 89:1301–1304.

71. Fu XD, Maniatis T. The 35-kDa mammalian splicing factor SC35 mediates specific interactions between U1 and U2 small nuclear ribonucleoprotein particles at the 3' splice site. *Proc Natl Acad Sci U S A* 1992, 89:1725–1729.

72. Zahler AM, Lane WS, Stolk JA, Roth MB. SR proteins: a conserved family of pre-mRNA splicing factors. *Genes Dev* 1992, 6:837–847.

73. Roth MB, Murphy C, Gall JG. A monoclonal antibody that recognizes a phosphorylated epitope stains lampbrush chromosome loops and small granules in the amphibian germinal vesicle. *J Cell Biol* 1990, 111:2217–2223.

74. Roth MB, Gall JG. Monoclonal antibodies that recognize transcription unit proteins on newt lampbrush chromosomes. *J Cell Biol* 1987, 105:1047–1054.

75. Gui JF, Lane WS, Fu XD. A serine kinase regulates intracellular localization of splicing factors in the cell cycle. *Nature* 1994, 369:678–682.

76. Colwill K, Pawson T, Andrews B, Prasad J, Manley JL, Bell JC, Duncan PI. The Clk/Sty protein kinase phosphorylates SR splicing factors and regulates their intranuclear distribution. *EMBO J* 1996, 15:265–275.

77. Aubol BE, Plocinik RM, Hagopian JC, Ma CT, McGlone ML, Bandyopadhyay R, Fu XD, Adams JA. Partitioning RS domain phosphorylation in an SR protein through the CLK and SRPK protein kinases. *J Mol Biol* 2013, 425:2894–2909.

78. Ma CT, Hagopian JC, Ghosh G, Fu XD, Adams JA. Regiospecific phosphorylation control of the SR protein ASF/SF2 by SRPK1. *J Mol Biol* 2009, 390:618–634.

79. Colwill K, Feng LL, Yeakley JM, Gish GD, Caceres JF, Pawson T, Fu XD. SRPK1 and Clk/Sty protein kinases show distinct substrate specificities for serine/arginine-rich splicing factors. *J Biol Chem* 1996, 271:24569–24575.

80. Huynh N, Ma CT, Giang N, Hagopian J, Ngo J, Adams J, Ghosh G. Allosteric interactions direct binding and phosphorylation of ASF/SF2 by SRPK1. *Biochemistry* 2009, 48:11432–11440.

81. Mermoud JE, Cohen PT, Lamond AI. Regulation of mammalian spliceosome assembly by a protein phosphorylation mechanism. *EMBO J* 1994, 13:5679–5688.

82. Kanopka A, Muhlemann O, Petersen-Mahrt S, Estmer C, Ohrmalm C, Akusjarvi G. Regulation of adenovirus alternative RNA splicing by dephosphorylation of SR proteins. *Nature* 1998, 393:185–187.

83. Ma CT, Ghosh G, Fu XD, Adams JA. Mechanism of dephosphorylation of the SR protein ASF/SF2 by protein phosphatase 1. *J Mol Biol* 2010, 403:386–404.

84. Cho S, Hoang A, Sinha R, Zhong XY, Fu XD, Krainer AR, Ghosh G. Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. *Proc Natl Acad Sci U S A* 2011, 108:8233–8238.

85. Misteli T, Caceres JF, Clement JQ, Krainer AR, Wilkinson MF, Spector DL. Serine phosphorylation of SR proteins is required for their recruitment to sites of transcription in vivo. *J Cell Biol* 1998, 143:297–307.

86. Tacke R, Chen Y, Manley JL. Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer. *Proc Natl Acad Sci U S A* 1997, 94:1148–1153.

87. Xiao SH, Manley JL. Phosphorylation of the ASF/SF2 RS domain affects both protein-protein and protein-RNA interactions and is necessary for splicing. *Genes Dev* 1997, 11:334–344.

88. Xiang S, Gapsys V, Kim HY, Bessonov S, Hsiao HH, Mohlmann S, Klaukien V, Ficner R, Becker S, Urlaub H, et al. Phosphorylation drives a dynamic switch in serine/arginine-rich proteins. *Structure* 2013, 21:2162–2174.

89. Caceres JF, Screaton GR, Krainer AR. A specific subset of SR proteins shuttles continuously between

the nucleus and the cytoplasm. *Genes Dev* 1998, 12:55–66.

90. Ngo JC, Chakrabarti S, Ding JH, Velazquez-Dones A, Nolen B, Aubol BE, Adams JA, Fu XD, Ghosh G. Interplay between SRPK and Clk/Sty kinases in phosphorylation of the splicing factor ASF/SF2 is regulated by a docking motif in ASF/SF2. *Mol Cell* 2005, 20:77–89.

91. Sanford JR, Ellis JD, Cazalla D, Caceres JF. Reversible phosphorylation differentially affects nuclear and cytoplasmic functions of splicing factor 2/alternative splicing factor. *Proc Natl Acad Sci U S A* 2005, 102:15042–15047.

92. Kataoka N, Bachorik JL, Dreyfuss G. Transportin-SR, a nuclear import receptor for SR proteins. *J Cell Biol* 1999, 145:1145–1152.

93. Lai MC, Lin RI, Huang SY, Tsai CW, Tarn WY. A human importin-β family protein, transportin-SR2, interacts with the phosphorylated RS domain of SR proteins. *J Biol Chem* 2000, 275:7950–7957.

94. Sanford JR, Bruzik JP. Developmental regulation of SR protein phosphorylation and activity. *Genes Dev* 1999, 13:1513–1518.

95. Sanford JR, Bruzik JP. Regulation of SR protein localization during development. *Proc Natl Acad Sci U S A* 2001, 98:10184–10189.

96. Bridge E, Xia DX, Carmo-Fonseca M, Cardinali B, Lamond AI, Pettersson U. Dynamic organization of splicing factors in adenovirus-infected cells. *J Virol* 1995, 69:281–290.

97. Zhou Z, Qiu J, Liu W, Zhou Y, Plocinik RM, Li H, Hu Q, Ghosh G, Adams JA, Rosenfeld MG, et al. The Akt-SRPK-SR axis constitutes a major pathway in transducing EGF signaling to regulate alternative splicing in the nucleus. *Mol Cell* 2012, 47:422–433.

98. Edmond V, Moysan E, Khochbin S, Matthias P, Brambilla C, Brambilla E, Gazzeri S, Eymin B. Acetylation and phosphorylation of SRSF2 control cell fate decision in response to cisplatin. *EMBO J* 2011, 30:510–523.

99. Bressan GC, Moraes EC, Manfiolli AO, Kuniyoshi TM, Passos DO, Gomes MD, Kobarg J. Arginine methylation analysis of the splicing-associated SR protein SFRS9/SRP30C. *Cell Mol Biol Lett* 2009, 14:657–669.

100. Chen YC, Milliman EJ, Goulet I, Cote J, Jackson CA, Vollbracht JA, Yu MC. Protein arginine methylation facilitates cotranscriptional recruitment of pre-mRNA splicing factors. *Mol Cell Biol* 2010, 30:5245–5256.

101. Lavigueur A, La Branche H, Kornblihtt AR, Chabot B. A splicing enhancer in the human fibronectin alternate ED1 exon interacts with SR proteins and stimulates U2 snRNP binding. *Genes Dev* 1993, 7:2405–2417.

102. Sun Q, Mayeda A, Hampson RK, Krainer AR, Rottman FM. General splicing factor SF2/ASF promotes alternative splicing by binding to an exonic splicing enhancer. *Genes Dev* 1993, 7:2598–2608.

103. Tian M, Maniatis T. A splicing enhancer complex controls alternative splicing of doublesex pre-mRNA. *Cell* 1993, 74:105–114.

104. Chiara MD, Gozani O, Bennett M, Champion-Arnaud P, Palandjian L, Reed R. Identification of proteins that interact with exon sequences, splice sites, and the branchpoint sequence during each stage of spliceosome assembly. *Mol Cell Biol* 1996, 16:3317–3326.

105. Staknis D, Reed R. SR proteins promote the first specific recognition of Pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex. *Mol Cell Biol* 1994, 14:7670–7682.

106. Kohtz JD, Jamison SF, Will CL, Zuo P, Luhrmann R, Garcia-Blanco MA, Manley JL. Protein-protein interactions and 5′-splice-site recognition in mammalian mRNA precursors. *Nature* 1994, 368:119–124.

107. Wu JY, Maniatis T. Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* 1993, 75:1061–1070.

108. Xiao SH, Manley JL. Phosphorylation-dephosphorylation differentially affects activities of splicing factor ASF/SF2. *EMBO J* 1998, 17:6359–6367.

109. Zhu J, Krainer AR. Pre-mRNA splicing in the absence of an SR protein RS domain. *Genes Dev* 2000, 14:3166–3178.

110. Wang Z, Hoffmann HM, Grabowski PJ. Intrinsic U2AF binding is modulated by exon enhancer signals in parallel with changes in splicing activity. *RNA* 1995, 1:21–35.

111. Guth S, Tange TO, Kellenberger E, Valcarcel J. Dual function for U2AF(35) in AG-dependent pre-mRNA splicing. *Mol Cell Biol* 2001, 21:7673–7681.

112. Graveley BR, Hertel KJ, Maniatis T. The role of U2AF35 and U2AF65 in enhancer-dependent splicing. *RNA* 2001, 7:806–818.

113. Shen H, Kan JL, Green MR. Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. *Mol Cell* 2004, 13:367–376.

114. Roscigno RF, Garcia-Blanco MA. SR proteins escort the U4/U6.U5 tri-snRNP to the spliceosome. *RNA* 1995, 1:692–706.

115. Schneider M, Will CL, Anokhina M, Tazi J, Urlaub H, Luhrmann R. Exon definition complexes contain the tri-snRNP and can be directly converted into B-like precatalytic splicing complexes. *Mol Cell* 2010, 38:223–235.

116. Shen H, Green MR. A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. *Mol Cell* 2004, 16:363–373.

117. Cao W, Jamison SF, Garcia-Blanco MA. Both phosphorylation and dephosphorylation of ASF/SF2 are required for pre-mRNA splicing in vitro. *RNA* 1997, 3:1456–1467.

118. Amin EM, Oltean S, Hua J, Gammons MV, Hamdollah-Zadeh M, Welsh GI, Cheung MK, Ni L, Kase S, Rennel ES, et al. WT1 mutants reveal SRPK1 to be a downstream angiogenesis target by altering VEGF splicing. *Cancer Cell* 2011, 20:768–780.

119. Berget SM. Exon recognition in vertebrate splicing. *J Biol Chem* 1995, 270:2411–2414.

120. De Conti L, Baralle M, Buratti E. Exon and intron definition in pre-mRNA splicing. *WIREs RNA* 2013, 4:49–60.

121. Stark JM, Bazett-Jones DP, Herfort M, Roth MB. SR proteins are sufficient for exon bridging across an intron. *Proc Natl Acad Sci U S A* 1998, 95:2163–2168.

122. Erkelenz S, Mueller WF, Evans MS, Busch A, Schoneweis K, Hertel KJ, Schaal H. Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA* 2013, 19:96–102.

123. Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G. Comparative analysis identifies exonic splicing regulatory sequences—the complex definition of enhancers and silencers. *Mol Cell* 2006, 22:769–781.

124. Sanford JR, Coutinho P, Hackett JA, Wang X, Ranahan W, Caceres JF. Identification of nuclear and cytoplasmic mRNA targets for the shuttling protein SF2/ASF. *PLoS One* 2008, 3:e3369.

125. Han J, Ding JH, Byeon CW, Kim JH, Hertel KJ, Jeong S, Fu XD. SR proteins induce alternative exon skipping through their activities on the flanking constitutive exons. *Mol Cell Biol* 2011, 31:793–802.

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

127. Maggiora P, Marchio S, Stella MC, Giai M, Belfiore A, De Bortoli M, Di Renzo MF, Costantino A, Sismondi P, Comoglio PM. Overexpression of the RON gene in human breast carcinoma. *Oncogene* 1998, 16:2927–2933.

128. Ghigna C, Giordano S, Shen H, Benvenuto F, Castiglioni F, Comoglio PM, Green MR, Riva S, Biamonti G. Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. *Mol Cell* 2005, 20:881–890.

129. Bauren G, Jiang WQ, Bernholm K, Gu F, Wieslander L. Demonstration of a dynamic, transcription-dependent organization of pre-mRNA splicing factors in polytene nuclei. *J Cell Biol* 1996, 133:929–941.

130. Bjork P, Jin S, Zhao J, Singh OP, Persson JO, Hellman U, Wieslander L. Specific combinations of SR proteins associate with single pre-messenger RNAs in vivo and contribute different functions. *J Cell Biol* 2009, 184:555–568.

131. Caceres JF, Stamm S, Helfman DM, Krainer AR. Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. *Science* 1994, 265:1706–1709.

132. Eperon IC, Makarova OV, Mayeda A, Munroe SH, Caceres JF, Hayward DG, Krainer AR. Selection of alternative 5′ splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. *Mol Cell Biol* 2000, 20:8303–8318.

133. Zhu J, Mayeda A, Krainer AR. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol Cell* 2001, 8:1351–1361.

134. Hanamura A, Caceres JF, Mayeda A, Franza BR Jr, Krainer AR. Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. *RNA* 1998, 4:430–444.

135. Zerbe LK, Pino I, Pio R, Cosper PF, Dwyer-Nield LD, Meyer AM, Port JD, Montuenga LM, Malkinson AM. Relative amounts of antagonistic splicing factors, hnRNP A1 and ASF/SF2, change during neoplastic lung growth: implications for pre-mRNA processing. *Mol Carcinog* 2004, 41:187–196.

136. Simard MJ, Chabot B. SRp30c is a repressor of 3′ splice site utilization. *Mol Cell Biol* 2002, 22:4001–4010.

137. Shin C, Manley JL. The SR protein SRp38 represses splicing in M phase cells. *Cell* 2002, 111:407–417.

138. Feng Y, Chen M, Manley JL. Phosphorylation switches the general splicing repressor SRp38 to a sequence-specific activator. *Nat Struct Mol Biol* 2008, 15:1040–1048.

139. Long JC, Caceres JF. The SR protein family of splicing factors: master regulators of gene expression. *Biochem J* 2009, 417:15–27.

140. Tacke R, Manley JL. The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *EMBO J* 1995, 14:3540–3551.

141. Liu HX, Zhang M, Krainer AR. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* 1998, 12:1998–2012.

142. Fu XD. Specific commitment of different pre-mRNAs to splicing by single SR proteins. *Nature* 1993, 365:82–85.

143. Lin S, Xiao R, Sun P, Xu X, Fu XD. Dephosphorylation-dependent sorting of SR splicing factors during mRNP maturation. *Mol Cell* 2005, 20:413–425.

144. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. CLIP identifies Nova-regulated RNA networks in the brain. *Science* 2003, 302:1212–1215.

145. Sanford JR, Wang X, Mort M, Vanduyn N, Cooper DN, Mooney SD, Edenberg HJ, Liu Y. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res* 2009, 19:381–394.

146. Bernard D, Prasanth KV, Tripathi V, Colasse S, Nakamura T, Xuan Z, Zhang MQ, Sedel F, Jourdren L, Coulpier F, et al. A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. *EMBO J* 2010, 29:3082–3093.

147. Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA, et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell* 2010, 39:925–938.

148. Royce-Tolland ME, Andersen AA, Koyfman HR, Talbot DJ, Wutz A, Tonks ID, Kay GF, Panning B. The A-repeat links ASF/SF2-dependent Xist RNA processing with random choice during X inactivation. *Nat Struct Mol Biol* 2010, 17:948–954.

149. Wu H, Sun S, Tu K, Gao Y, Xie B, Krainer AR, Zhu J. A splicing-independent function of SF2/ASF in microRNA processing. *Mol Cell* 2010, 38:67–77.

150. Anko ML, Muller-McNicoll M, Brandl H, Curk T, Gorup C, Henry I, Ule J, Neugebauer KM. The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biol* 2012, 13:R17.

151. Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, Shiue L, Clark TA, Blume JE, Ares M Jr. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev* 2007, 21:708–718.

152. Martins SB, Rino J, Carvalho T, Carvalho C, Yoshida M, Klose JM, de Almeida SF, Carmo-Fonseca M. Spliceosome assembly is coupled to RNA polymerase II dynamics at the 3′ end of human genes. *Nat Struct Mol Biol* 2011, 18:1115–1123.

153. Lee KM, Tarn WY. Coupling pre-mRNA processing to transcription on the RNA factory assembly line. *RNA Biol* 2013, 10:380–390.

154. Misteli T, Caceres JF, Spector DL. The dynamics of a pre-mRNA splicing factor in living cells. *Nature* 1997, 387:523–527.

155. Yuryev A, Patturajan M, Litingtung Y, Joshi RV, Gentile C, Gebara M, Corden JL. The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins. *Proc Natl Acad Sci U S A* 1996, 93:6975–6980.

156. Kim E, Du L, Bregman DB, Warren SL. Splicing factors associate with hyperphosphorylated RNA polymerase II in the absence of pre-mRNA. *J Cell Biol* 1997, 136:19–28.

157. Du L, Warren SL. A functional interaction between the carboxy-terminal domain of RNA polymerase II and pre-mRNA splicing. *J Cell Biol* 1997, 136:5–18.

158. Misteli T, Spector DL. RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo. *Mol Cell* 1999, 3:697–705.

159. de la Mata M, Kornblihtt AR. RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. *Nat Struct Mol Biol* 2006, 13:973–980.

160. Sapra AK, Anko ML, Grishina I, Lorenz M, Pabis M, Poser I, Rollins J, Weiland EM, Neugebauer KM. SR protein family members display diverse activities in the formation of nascent and mature mRNPs in vivo. *Mol Cell* 2009, 34:179–190.

161. Lin S, Coutinho-Mansfield G, Wang D, Pandit S, Fu XD. The splicing factor SC35 has an active role in transcriptional elongation. *Nat Struct Mol Biol* 2008, 15:819–826.

162. Ji X, Zhou Y, Pandit S, Huang J, Li H, Lin CY, Xiao R, Burge CB, Fu XD. SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. *Cell* 2013, 153:855–868.

163. Twyffels L, Gueydan C, Kruys V. Shuttling SR proteins: more than splicing factors. *FEBS J* 2011, 278:3246–3255.

164. Huang Y, Steitz JA. Splicing factors SRp20 and 9G8 promote the nucleocytoplasmic export of mRNA. *Mol Cell* 2001, 7:899–905.

165. Huang Y, Gattoni R, Stevenin J, Steitz JA. SR splicing factors serve as adapter proteins for TAP-dependent mRNA export. *Mol Cell* 2003, 11:837–843.

166. Lai MC, Tarn WY. Hypophosphorylated ASF/SF2 binds TAP and is present in messenger ribonucleoproteins. *J Biol Chem* 2004, 279:31745–31749.

167. Escudero-Paunetto L, Li L, Hernandez FP, Sandri-Goldin RM. SR proteins SRp20 and 9G8 contribute to efficient export of herpes simplex virus 1 mRNAs. *Virology* 2010, 401:155–164.

168. Swartz JE, Bor YC, Misawa Y, Rekosh D, Hammarskjold ML. The shuttling SR protein 9G8 plays a role in translation of unspliced mRNA containing a constitutive transport element. *J Biol Chem* 2007, 282:19844–19853.

169. Sanford JR, Gray NK, Beckmann K, Caceres JF. A novel role for shuttling SR proteins in mRNA translation. *Genes Dev* 2004, 18:755–768.

170. Michlewski G, Sanford JR, Caceres JF. The splicing factor SF2/ASF regulates translation initiation by enhancing phosphorylation of 4E-BP1. *Mol Cell* 2008, 30:179–189.

171. Maslon MM, Heras SR, Bellora N, Eyras E, Caceres JF. The translational landscape of the splicing factor SRSF1 and its role in mitosis. *eLife* 2014, 3:020208.

172. Bedard KM, Daijogo S, Semler BL. A nucleo-cytoplasmic SR protein functions in viral IRES-mediated translation initiation. *EMBO J* 2007, 26:459–467.

173. Swanson CM, Sherer NM, Malim MH. SRp40 and SRp55 promote the translation of unspliced human immunodeficiency virus type 1 RNA. *J Virol* 2010, 84:6748–6759.

174. Liu KJ, Harland RM. Inhibition of neurogenesis by SRp38, a neuroD-regulated RNA-binding protein. *Development* 2005, 132:1511–1523.

175. Dreyfuss G, Kim VN, Kataoka N. Messenger-RNA-binding proteins and the messages they carry. *Nat Rev Mol Cell Biol* 2002, 3:195–205.

176. Sterne-Weiler T, Martinez-Nunez RT, Howard JM, Cvitovik I, Katzman S, Tariq MA, Pourmand N, Sanford JR. Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res* 2013, 23: 1615–1623.

177. Lemaire R, Prasad J, Kashima T, Gustafson J, Manley JL, Lafyatis R. Stability of a PKCI-1-related mRNA is controlled by the splicing factor ASF/SF2: a novel function for SR proteins. *Genes Dev* 2002, 16:594–607.

178. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* 2007, 446:926–929.

179. Zhang Z, Krainer AR. Involvement of SR proteins in mRNA surveillance. *Mol Cell* 2004, 16:597–607.

180. Sato H, Hosoda N, Maquat LE. Efficiency of the pioneer round of translation affects the cellular site of nonsense-mediated mRNA decay. *Mol Cell* 2008, 29:255–262.

181. Le Hir H, Gatfield D, Izaurralde E, Moore MJ. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *EMBO J* 2001, 20:4987–4997.

182. Nott A, Le Hir H, Moore MJ. Splicing enhances translation in mammalian cells: an additional function of the exon junction complex. *Genes Dev* 2004, 18:210–222.

183. Singh G, Kucukural A, Cenik C, Leszyk JD, Shaffer SA, Weng Z, Moore MJ. The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. *Cell* 2012, 151:750–764.

184. Silver DL, Watkins-Chow DE, Schreck KC, Pierfelice TJ, Larson DM, Burnetti AJ, Liaw HJ, Myung K, Walsh CA, Gaiano N, et al. The exon junction complex component Magoh controls brain size by regulating neural stem cell division. *Nat Neurosci* 2010, 13:551–558.

185. Huang M, Rech JE, Northington SJ, Flicker PF, Mayeda A, Krainer AR, LeStourgeon WM. The C-protein tetramer binds 230 to 240 nucleotides of pre-mRNA and nucleates the assembly of 40S heterogeneous nuclear ribonucleoprotein particles. *Mol Cell Biol* 1994, 14:518–533.

186. Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* 2010, 17:909–915.

187. Zarnack K, Konig J, Tajnik M, Martincorena I, Eustermann S, Stevant I, Reyes A, Anders S, Luscombe NM, Ule J. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* 2013, 152:453–466.

188. Baltz AG, Munschauer M, Schwanhausser B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, Milek M, et al. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell* 2012, 46:674–690.

189. Schueler M, Munschauer M, Gregersen LH, Finzel A, Loewer A, Chen W, Landthaler M, Dieterich C. Differential protein occupancy profiling of the mRNA transcriptome. *Genome Biol* 2014, 15:R15.

Chapter 3

hnRNP A1 modulates the association of U2AF65 with Alu-derived RNAs

Abstract

Heterogeneous nuclear ribonucleoprotein A1 (hnRNP A1) is an important splicing regulator but its mechanisms of action on a global remain enigmatic. We used individual-nucleotide resolution crosslinking immunoprecipitation (iCLIP) to determine how enforced expression of hnRNP A1 affects global protein-RNA interactions of U2 snRNA auxiliary factor 2 (U2AF2) and the serine arginine-rich protein, SRSF1. We observed significant changes in the distribution of U2AF2 crosslinking sites relative to the 3' splice sites of cassette exons but not constitutive exons upon hnRNP A1 over expression. By contrast, SRSF1 crosslinking patterns relative to splice sites are independent of hnRNP A1 expression levels. We also observed an hnRNPA1-dependent increase in U2AF2 but not SRSF1 crosslinking to intronic Alu elements that are most proximal to exons. We propose that Alu-derived RNA elements function in the hnRNP A1-dependent remodeling of U2AF2-RNA interactions.

Introduction

In human genes, non-coding sequences (introns) separate small protein coding regions (exons). Splice site sequences define boundaries between exons and introns. The splicesosome cuts introns from pre-messenger RNA (pre-mRNA) and pastes together the exon sequences. This process initiates with 3' and 5' splice site recognition by the U2 snRNP auxiliary factor (U2AF, composed of U2AF1 and U2AF2 proteins)(Ruskin et al., 1988) and the U1 snRNP (Mount et al., 1983). RNA binding proteins such as hnRNPA1 and SRSF1 modulate these interactions to generate alternative mRNA isoforms (Buvoli et al., 1992; Caceres et al., 1994; Chiou et al., 2013; Ge and Manley, 1990; Krainer et al., 1990; Sun et al., 1993; Tavanez et al., 2012; Wu and Maniatis, 1993; Xiao and Manley, 1997; Zhu et al., 2001). Unfortunately, the generality of the mechanisms explaining how hnRNPA1 and SRSF1 regulate splicing remain enigmatic (Blanchette and Chabot, 1999; Eperon et al., 2000; Okunola and Krainer, 2009; Zahler et al., 2004; Zhu et al., 2001).

Results

To determine how hnRNPA1 influences 3' splice site recognition genome-wide, we perturbed its' expression in HEK293 cells and assayed SRSF1, U2AF2 and hnRNPA1 protein-RNA interactions using individual nucleotide resolution crosslinking immunopreciptitation and high throughput sequencing (iCLIP-

44

seq)(Konig et al., 2010). Induction of hnRNPA1 results in an approximately 2-3 fold increase compared to the endogenous protein (and relative to EWSR1) and has no appreciable affect on SRSF1 or U2AF2 steady state protein levels (Supplemental Figure 1) or hnRNP C (Supplemental Figure 8). We used iCLIP to purify hnRNPA1-, SRSF1-, and U2AF2-RNA complexes, in at least duplicate, from control and hnRNPA1 over-expressing cells (Figure 1A). As expected, the immunoprecipitated material was both UV- and antibody-dependent, nuclease sensitive and produced robust sequencing libraries (supplemental Table 1) with low duplication rates (Supplemental Figure 2). After identification of peaks (supplemental Figure 3) using CLIPper (Lovci et al., 2013) the distribution of hnRNP A1 peaks between different gene regions is largely unchanged upon hnRNP A1 over expression (Figure 1B). However, we observed differences between control and hnRNP A1 over expression cell lines for both U2AF2 and SRSF1 peaks (Figure 1C, supplementary Figure 4). Most notably, the proportion of U2AF2 peaks located near coding exons or in exon proximal intronic regions was reduced whereas intronic peaks located more than 500 nt from exons (distal intron) increased. A similar trend was observed for SRSF1 peaks, where a reduction in CDS and concomitant increase in distal intron peaks was evident upon hnRNP A1 over expression (Figure 1B). To determine if hnRNP A1 influences the RNA binding specificity of U2AF2 and SRSF1 we searched for over represented RNA sequences within the binding site peaks (Figure 1C, supplementary Figure 5). In control cells, U2AF2 peaks are characterized by a pyrimidine-AG motif, closely resembling authentic 3'splice sites. By contrast, a more pyrimidine-rich motif

45

is observed in peaks from hnRNPA1 over expression cells. SRSF1 motifs show only modest differences between control and hnRNPA1 over expression cells. There is little difference in the motifs observed at hnRNPA1 peaks from either the control or hnRNPA1 over expression cells.

Previous work demonstrated that hnRNPA1 influences binding of SRSF1 and U2AF65 near 3' splice sites (Buvoli et al., 1992; Tavanez et al., 2012; Wu and Maniatis, 1993). To test this hypothesis we determined how titration of hnRNP A1 affected the distribution of SRSF1- and U2AF2-RNA crosslinks relative to 3' splice sites of constitutive or alternative cassette exons. As suggested by the peak analysis (Fig. 2) there are no differences in hnRNPA1 crosslinking site between control and over expression cells. SRSF1 crosslinking to exonic sequences was modestly reduced in the hnRNPA1 over expression cells compared to control, but the positional distribution of the SRSF1 sites relative to the 3'ss was largely unchanged for constitutive and skipped exons (Figure 2B and E, red and blue lines, respectively). By contrast, U2AF2 crosslinking distribution relative to the 3'ss was substantially altered in hnRNPA1 over expressing cells compared to the control, where a characteristic peak is observed over the 3'ss of both constitutive and skipped exons (Figure 2C and F, blue lines). By contrast, in cells overexpressing hnRNPA1, U2AF2 crosslinking density near alternative exons shifts downstream of the 3'ss and the peak is substantially reduced (Figure 2F, red line).

To determine if there is a direct relationship between hnRNPA1 and changes in U2AF2 or SRSF1 association with transcripts, we examined regions flanking the

3'ss of skipped exons with significant hnRNPA1, U2AF2 and SRSF1 crosslinking in both conditions (Figure 2F). In regions with no detectable hnRNPA1 crosslinks, the change in U2AF2 crosslinking exhibits a bimodal distribution, which corresponds to regions flanking the 3'ss that show either increased or decreased U2AF2 crosslinking in hnRNPA1 over expression cells relative to control cells (Figure 2F, blue). By contrast, U2AF2 crosslinking to the vicinity of the 3'ss is significantly reduced when direct association of hnRNPA1 is also evident (Figure 2F, pink). Changes in SRSF1 crosslinking appears to be independent of direct hnRNPA1-RNA interactions (Figure 2G). To determine if changes in U2AF2 crosslinking correlated with hnRNPA1-dependent splicing regulation we sequenced polyA+ selected RNA libraries from control and hnRNPA1 overexpression cells. Of the 267 hnRNPA1-regulated cassette exons, the majority exhibited increased levels of exon skipping upon hnRNPA1 overexpression (supplemental Figure 6, bar graph). A total of 83 hnRNPA1 regulated splicing events also exhibited changes in U2AF2 intronic crosslinking (supplemental figure 6, Table). 60% of hnRNPA1-dependent exon skipping events also exhibit depletion of U2AF2 crosslinking within 200 bp of the 3'ss, whereas 40% show hnRNPA1-dependent increases in U2AF2 crosslinking in the same region. For example, we observe hnRNPA1-dependent U2AF2 redistribution and alternative splicing in COG4 (Conserved oligomeric Golgi complex subunit 4; Fig. 2F) and SRSF6 (serine/arginine-rich splicing factor 6 or SRp55; Fig. 2H). In both cases, U2AF2 crosslinking near the 3' splice sites is reduced in the cell lines over expressing hnRNPA1 (Fig. 2G-I)

Previous work by Zarnack et al. demonstrated that hnRNP proteins, such as hnRNP C, can antagonize binding of U2AF2 to *Alu* element, to repress their exonization (Zarnack et al., 2013). We asked if hnRNP A1 similarly repressed U2AF2 crosslinking to Alu elements by measuring global distribution of crosslinks for each protein overlapping of antisense Alu elements throughout intronic regions with titration of hnRNP A1 levels. Surprisingly upon overexpression of hnRNP A1, we detected a dramatic increase in U2AF2 crosslinking to antisense Alu-containing RNA transcripts compared to control cells (Fig 3A). Conversely, hnRNP A1 crosslinking globally decreases over Alu elements with over-expression. By contrast to U2AF2, crosslinking of SRSF1 to antisense Alu elements shows no appreciable changes, suggesting that the affect of hnRNPA1 is specific to U2AF2.

These results suggest that Alu elements with increased U2AF2 crosslinking are located in *cis-* relative to skipped exons (Alu elements upstream or downstream of a splicing event) (Fig. 3B). To test this hypothesis we compared the proportion of U2AF2 crosslinks within Alu-elements relative to flanking sequences across individual exon skipping events in control or hnRNPA1 over expression cells. The scatter plot shown in Fig 3C, demonstrate that the proportion of U2AF2 crosslinks present in Alu elements increases significantly across virtually all exon skipping events, upon hnRNPA1 over expression, whereas the proportion of hnRNPA1 crosslinks are decreased (Supplemental Figure 7). By contrast, the proportion of SRSF1 crosslinks to Alu elements are refractory to changes in hnRNPA1 expression levels (Supplemental Figure 7). These data demonstrate a global change in U2AF2-

Alu association and refute the hypothesis that a few spurious Alu-elements are responsible for the signal observed in Fig 3A. We found that 54% of hnRNPA1-dependent exon skipping events exhibited redistribution of U2AF2 to adjacent Alu elements (Supplemental Figure 6, Table). An example from the PIEZO1 gene is shown in Fig. 3D. Taken together, these data suggest a novel mechanism by which nearby Alu elements regulate alternative exon inclusion.

Alu elements influence alternative splicing, although the mechanisms are poorly understood (Gal-Mark et al., 2009; Lev-Maor et al., 2008; Pastor and Pagani, 2011; Schwartz et al., 2009; Sorek et al., 2002). We investigated the positions of Alu-elements with hnRNPA1-dependent changes in U2AF2 crosslinking relative to the 3' splice site of constitutive or skipped exons. As expected, we observed that Alu elements are closer to skipped than constitutive exons ($p < 1.4e-47$, Wilcoxon rank-sum test, Fig. 4A, compare green and yellow boxes). But yet, those Alu elements with hnRNPA1-dependent increases in U2AF crosslinking are significantly closer to exons than those that are unchanged ($p < 9.5e-93$, Fig 4A). Taken together our data suggest the intriguing hypothesis that Alu-elements may function as cis-regulatory elements that compete with authentic exons for binding to splicing factors.

Alu elements influence gene expression in diverse ways(Chen and Carmichael, 2009; Gong and Maquat, 2011; Hasler and Strub, 2006; Pastor and Pagani, 2011). The results presented here implicate Alu elements in splicing

regulation. The proximity of hnRNPA1-responsive Alu-U2AF2 interaction sites to exons supports this hypothesis. Could these rogue binding sites compete with adjacent exons for U2AF2 binding (Cho et al., 2015; Shao et al., 2014)? Zarnack et al.. demonstrated that hnRNPC competes with U2AF2 to repress inclusion of Alu-derived exons in mRNA. We find that hnRNPA1 overexpression correlates with increased U2AF2 association with Alu-derived RNA sequences. We hypothesize that Alu elements function as a sink for U2AF2. In this model, U2AF2 dissociation from Alu-derived sequences maybe prevented by hnRNPA1. Alternatively, hnRNPA1 may alter U2AF2 RNA binding specificity thereby enhancing association with Alu elements. Taken together, our data demonstrate that Alu-derived sequences function as RNA regulatory elements that respond to changes to the intracellular concentration of splicing factors. These results suggest the intriguing hypothesis that retrotransposons contribute to species-specific differences in alternative splicing throughout the primate lineage (Fig. 4B).

Methods and Materials

iCLIP method

iCLIP was performed as previously described (König J et al.., 2010). Briefly, TREX FLP-in HEK293T cells (Invitrogen) lacking or containing a stable, inducible T7-tagged version of hnRNP A1. Cells were treated with tetracyclin for 24 hr and then irradiated with UV-C light to form irreversible covalent cross-link between

50

proteins and nucleic acids *in vivo*. After cell lysis, RNA was partially fragmented using low concentrations of Micrococcal nuclease, and U2AF65-, SRSF1-, or hnRNP A1–RNA complexes were immunopurified with α-U2AF65, (MC3;SCBT), α-SRSF1 (96;SCBT), and α-hnRNP A1 (4B10;SCBT) antibodies immobilized on protein A–coated magnetic beads (Life Technologies), respectively. After stringent washing and dephosphorylation (Fast AP, Fermentas), RNAs were ligated at their 3′ ends with a pre-adenylated RNA adaptor (Bioo Scientific) and radioactively labeled to allow visualization. Samples were run using MOPS-based protein gel electrophoresis (in-house recipe) and transferred to a nitrocellulose membrane. Protein-RNA complexes migrating 15 -80 kDa above free protein were cut from the membrane, and RNA was recovered from the membrane by proteinase K digestion under denaturing (3.5 M Urea) conditions. The oligonucleotides for reverse transcription contained two inversely oriented adaptor regions adapted from the Bioo NEXTflex small RNA library preparation kit (Bioo Scientific), separated by a BamHI restriction site as well as a barcode region at their 5′ end containing a 4-nt experiment-specific barcode within a 5-nt random barcode to mark individual cDNA molecules. cDNA molecules were size-purified using denaturing PAGE gel electrophoresis, circularized by CircLigase II (Epicenter), annealed to an oligonucleotide complementary to the restriction site and cut using BamHI (NEB). Linearized cDNAs were then PCR-amplified using (Immomix PCR Master Mix, Bioline) with primers (Bioo) complementary to the adaptor regions and were subjected to high-throughput

sequencing using Illumina HiSeq. A more detailed description of the iCLIP protocol has been published (Huppertz I et al.., 2014)

iCLIP Analysis

Following transcriptomic and genomic alignment with 'Tophat2' (Kim et al.., 2013), individual reads were truncated to their 5' ends to represent the site of crosslinking consistent with the iCLIP methodology. For all samples only overlapping peak regions found to be reproducible in two out of three replicates (or both duplicates where applies) were considered to be biologically reproducible candidates for further analysis. Reproducible data from each replicate was then aggregated into one data set and 5' end crosslinks were summed at each position. To determine background from the iCLIP data sets , those genomic positions that showed overlap in all three aggregate data sets were determined and a five nucleotide mask was created for each repetitive 5' position to create a filter track. Aggregated data was then filtered using this generate filter track and used for downstream analyses. This filtering pipeline was adapted from previous described methodology (Flynn et al., 2015; Friedersdorf and Keene, 2014)

CLIPper (CLIP-seq peak enrichment; https://github.com/YeoLab/clipper), was used to determine genomic distribution of RNA crosslinking peaks as well as

identify clusters representing binding sites for hnRNP A1, U2AF2, and SRSF1 for each condition as previously described (Lovci et al., 2013)

RBP Binding Analysis

40,769 cassette exons are extracted from MISO human genome (hg19) alternative events annotation version 2. 200,880 constitutive exons are extracted from RefSeq gene annotation by excluding the exons that overlap with cassette exons. Gene differential expression analysis is done by edgeR. 40,952 constitutive exons that are not significantly differential expressed (FDR > 0.05) are used in further analysis. For each RNA binding protein in each cell line, the iCLIP reads of all the replicates are merged together. The start position of reads are considered as crosslinking sites. The number of reads near 3' splice site (100bp into the intron, 50bp into the exon) of each exon is calculated based on a 10bp window. The raw read counts is normalized by the total library size. Exons with more than 20 reads in the 150bp region are shown in the plot.

The binding changes of U2AF and SF2 near 3' splice sites are further analyzed with edgeR. Read counts are calculated for 200bp intron regions near 3' splice sites of cassette exons. For each RBP, the regions with more than one count per million (CPM) in at least half of the replicates in either of the cell line are used for binding change analysis.

Splicing Change Identification

Whole transctriptome sequencing is performed on wild type cell line and hnRNPA1 over-expressed cell line with duplicates. Reads are mapped to human reference genome (hg19) with TopHat2. Mapped reads of duplicates are merged together for splicing analysis. Splicing change is analyzed with MISO. The MISO result is filtered with parameters: --num-inc 1 --num-exc 1 --num-sum-exc 10 --delta-psi 0.20 --bayes-factor 10. After filtering, 267 skipped exon events are left for further analysis.

Motif Analysis

For each condition, the iCLIP data of replicates are merged. Binding peaks are called with CLIPper (details needed.). The peaks are divided into different categories based on genomic regions including CDS, intron and UTR. Each category is further divided based on whether overlap with Alu elements. 50bp sequences of peak region (crosslinking site ± 25bp) are extracted. A strand-specific MEME-ChIP analysis is performed to find the enriched motifs with width between 6bp to 10bp.

RBP Binding Near Alu Elements

315,974 anti-sense Alu elements are extracted from RepeatMasker. The merged iCLIP data for each condition is down-sampled to 1M reads. The total number of sense strand reads are calculated for Alu and nearby regions (250bp from Alu boundary). For each cassette exon events (cassette exon + up/downstream introns + up/downstream exons), the number of reads in anti-sense Alu elements and the total number of reads in the whole event are calculated separately. The proportion of reads fall into anti-sense Alu elements for each event is used to represent the RBP binding change in Alu regions.

Figure 1. Crosslinking immunoprecipitation of hnRNPA1, SRSF1, and U2AF2 under hnRNP A1 modulation. (A) Examples of iCLIP autoradiographs for each protein under control and overexpression of hnRNP A1. Protein-RNA complex shifts are UV-, antibody- and Micrococcal nuclease-sensitive. Bars denote region of nitrocellulose blot excised for RNA isolation for iCLIP library preparation. **(B)** CLIPper analysis of iCLIP RNA distribution for hnRNP A1, SRSF1, and U2AF2 for control and hnRNP A1 overexpression conditions. **(C)** Top HOMER consensus binding motifs for hnRNP A1, SRSF1, and U2AF2 for control and hnRNP A1 overexpression conditions.
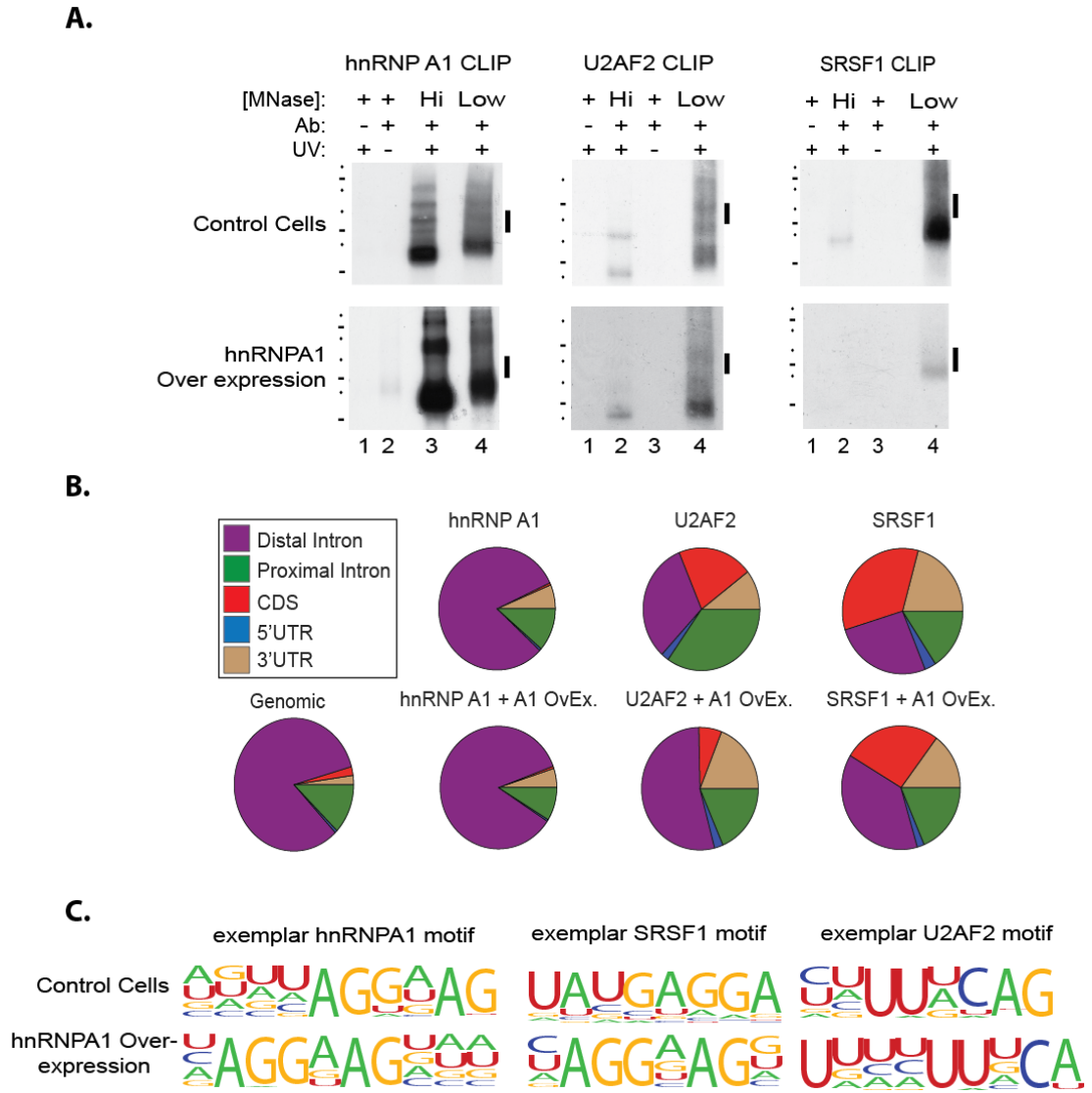
A.

B.

C.

Fig. 1

Figure 2. hnRNP A1 induced redistribution of U2AF2 crosslinking near 3' splice sites. (A,B) Normalized crosslinking distribution for hnRNPA1 (left panel), SRSF1 (middle panel) and U2AF2 (right panel) in wild type (blue line) and hnRNPA1 over-expression cell lines (red line) with 95% confidence interval (grey area). Data is divided between constitutive (A) and cassette (B) exons. **(C,D)** Nature log fold change distribution of U2AF (C) and SRSF1 (D) within 200bp intron regions near 3' splice sites of cassette exons. Blue bars corresponds to annotated alternative splicing events with no evidence of hnRNPA1 crosslinking in either condition and pink represents annotated events with detectable hnRNP A1 crosslinking. **(E)** Bar graph depicting the number of alternative cassette exons differentially expressed upon hnRNPA1 over expression. **(F-I)** Examples of hnRNPA1-dependent modulation of U2AF2 crosslinking and alternative splicing. UCSC genome browser examples of two genes COG4 (F) and SRSF6 (H) and CLIP read coverage data for U2AF65 under control and hnRNP A1 over-expression. Sashimi plots representing MISO analysis of RNA sequencing data from samples used for iCLIP. Each plot for COG4 (G) and SRSF6 (I) represents splicing data corresponding to the exon show in the USCS genome browser snapshots.
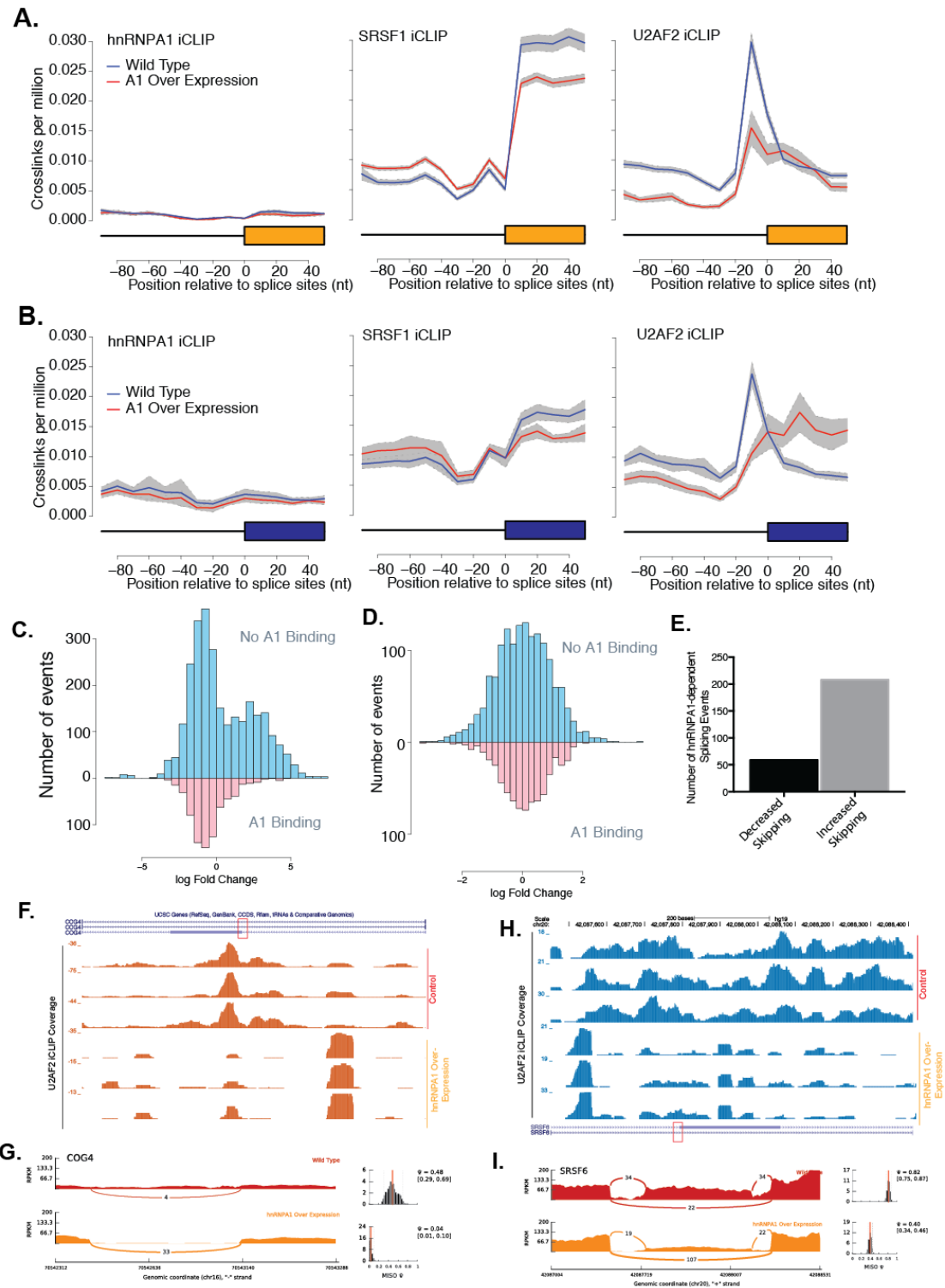
Fig. 2

Figure 3. hnRNP A1 over-expression correlates with global redistribution of U2AF2 signal to *Alu* RNA elements. (A) Aggregated read counts on *Alu* elements and nearby regions for U2AF2 (left panel), hnRNPA1 (middle panel) and SRSF1 (right panel). Blue represents wild-type binding of the given RNA binding protein and red represents hnRNP A1 overexpression of the log10 number of iCLIP read counts across all antisense-Alu elements. **(B)** Model representing two potential modes by which U2AF2 may associate with Alu RNA: *trans-* competition suggests U2AF2 binds to Alu elements on other RNAs, while a *cis*-competition suggest U2AF2 binds to *Alu* elements within the same RNAs that a particular exon is associated. **(C)** Scatter plot of all human cassette exons measuring the proportion of U2AF2 iCLIP crosslinks found within *Alu* elements within the cassette exon event over the total number of crosslinks found within the event. Proportions from control and hnRNP A1 over-expression samples are compared for each individual cassette exon event. **(D)** Example of hnRNPA1-dependent modulation of U2AF2-Alu interaction. UCSC genome browser snapshot depicting U2AF2 redistribution to anti-sense *Alu* elements in PIEZO1. iCLIP data (read coverage) for U2AF65 in control and hnRNP A1 over-expression cell lines. **(E)** Sashimi plot representing MISO analysis of RNA sequencing data from samples used for iCLIP. This plot for PIEZO1 represents splicing data corresponding to the exon show in the USCS genome browser snapshots.
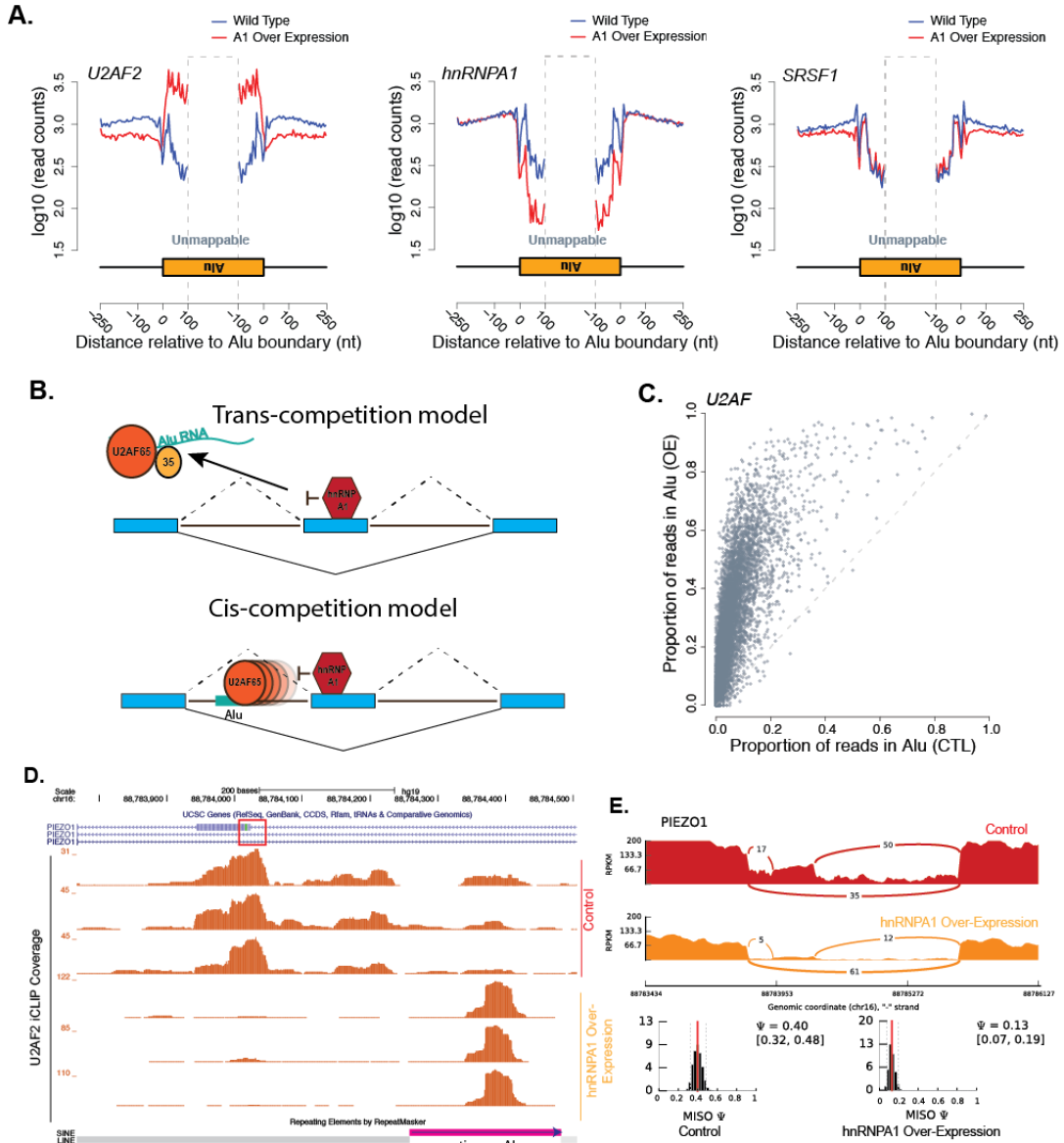
Fig. 3

Figure 4. hnRNPA1 promotes redistribution of U2AF2 crosslinking to exon proximal *Alu* elements. (A) Box plot representing the distance of *Alu* elements from cassette exons (blue) and constitutive exons (orange) that show no change in U2AF2 cross-linking versus those that show an increase in U2AF2 crosslinking. **(B)** Schematic describing a role for Alu-elements in the evolution of primate-specific alternative splicing.
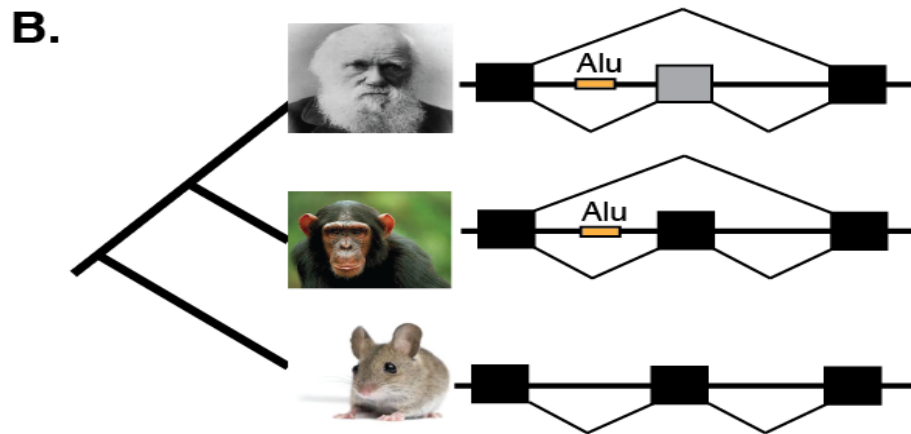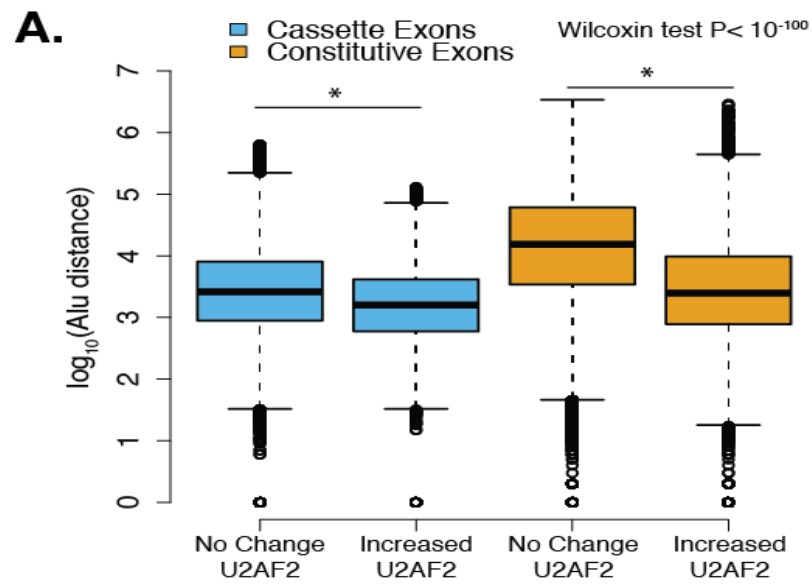
Fig. 4

Supplemental Figure 1. Western blot analysis of control and hnRNP A1 over-expression cell lines. Nuclear extracts were subjected to SDS-PAGE and transferred to nitrocellulose blotting paper. Samples were interrogated with anitbodies for α-T7 peptide tag (with which overexpressed hnRNP A1 has been tagged), α-hnRNP A1 (4B10; Santa Cruz Biotechnology), α-SRSF1 (96;Santa Cruz Biotechnology), α-U2AF2 (MC3;Santa Cruz Biotechnology), and α-EWSR1 (C-9; Santa Cruz Biotechnology) as positive control.
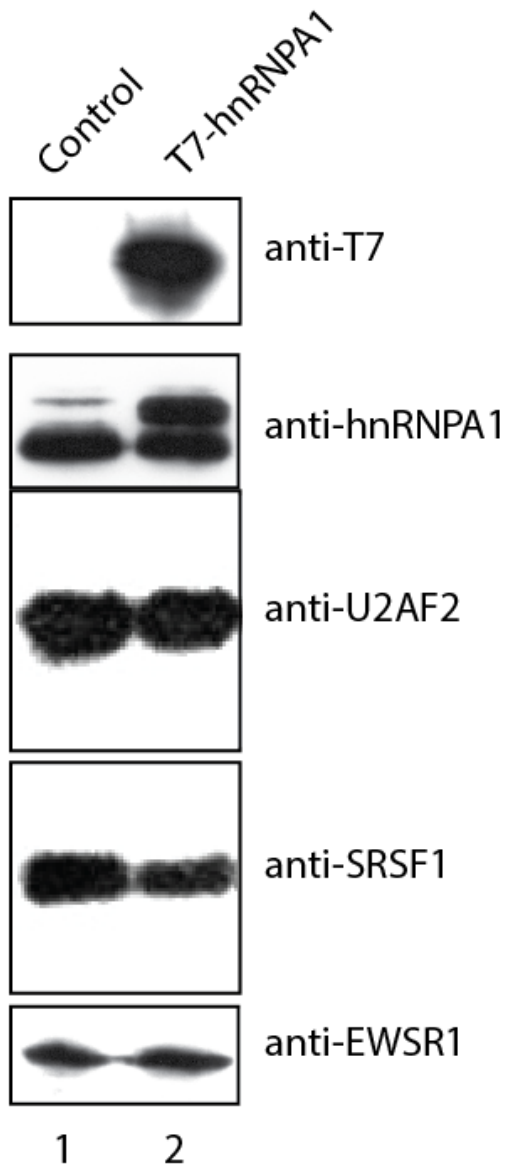
Supplemental Fig. 1

Supplemental Figure 2: Average Duplication Rates of iCLIP replicate libraries. Dot plot showing the PCR duplication rate for each sequencing library replicate for each iCLIP experiment.. "Control" refers to control HEK293 cell lines and "OverExpr" refers to cell lines in which hnRNP A1 overexpression has been induced.

Average Duplication Rates of iCLIP replicate libraries
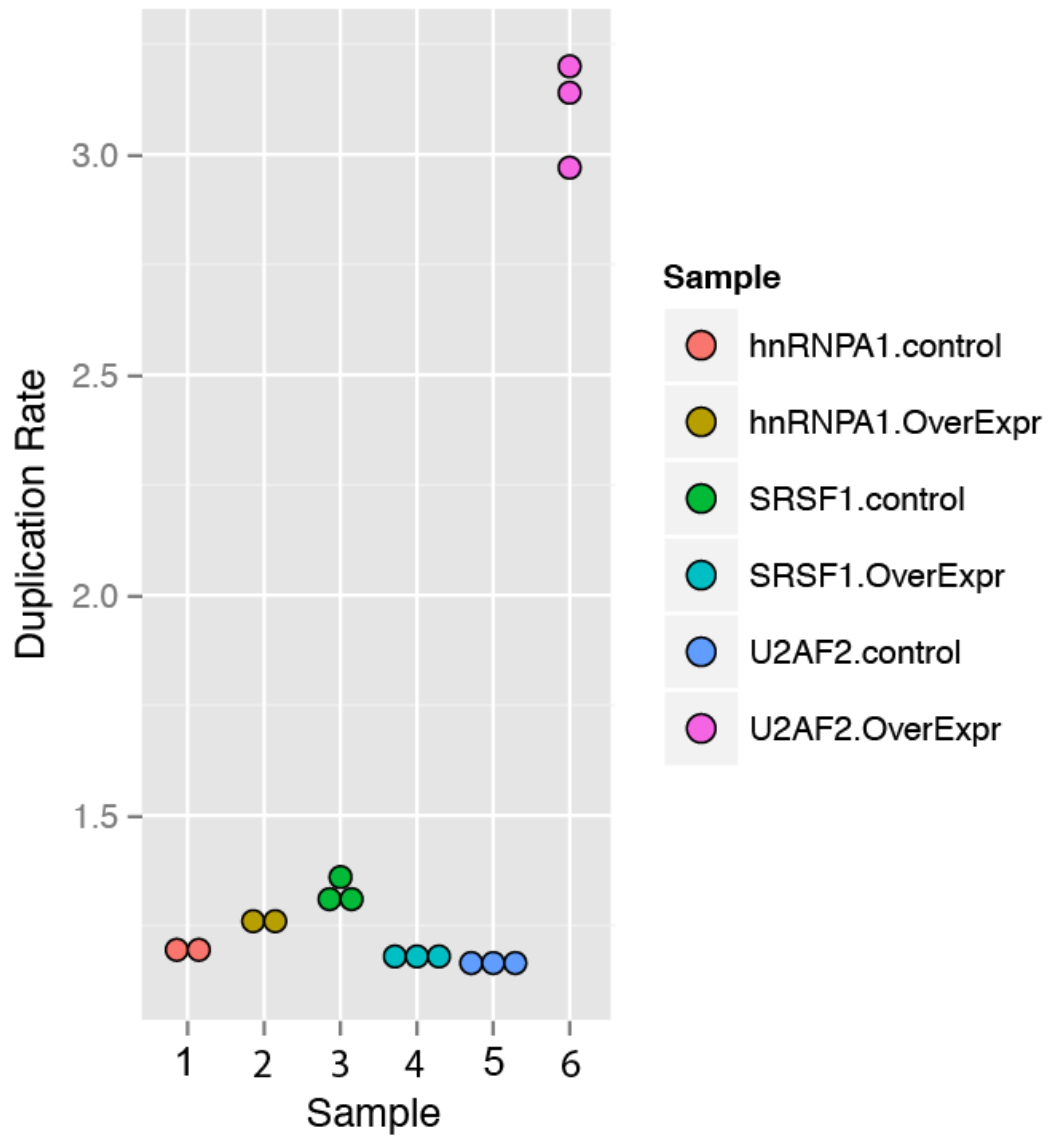
Supplemental Figure 2

Supplemental Table 1. Summary of iCLIP reads for SRSF1, U2AF2, and hnRNP A1 from control and hnRNP A1 over-expression cells. Table is a summary of reads from each iCLIP library from each protein under both conditions. Table also contains information on average read length, number of reads dedicated to peaks, and the average density of peaks called by CLIPper software. Statistical analysis are also provided with regards to number of reads dedicated to statistically significant peaks at various p-value cutoffs.

Sample	Mapped Reads	Number of Peaks	Median Number of Tags/ Peak	Average Number of Tags/ Peak	Median Length of Tag	Average Length of Tag	Read Density Per Peak	Fraction of Data in Peaks	pval0001	pval001	pval01	pval02	pval05
Aggregate SRSF1_A1 overexpression	19.29M	878041	6	13.3	22bp	20bp	665.00 RPK	60.56%	431K	587K	872K	877K	877K
Aggregate SRSF1_control	2.41M	91787	6	14	25bp	22bp	636.36 RPK	53.16%	46343	59724	91783	91783	91783
Aggregate hnRNP A1_A1 overexpression	0.68M	51059	5	7.8	23bp	19bp	410.53 RPK	58.94%	20497	29359	51059	51059	51059
Aggregate hnRNP A1_control	4.73M	319801	5	9.8	23bp	21bp	466.67 RPK	66.33%	141K	201K	319K	319K	319K
Aggregate U2AF65_A1 overexpression	12.92M	383889	8	22	12bp	17bp	1294.12 RPK	65.23%	223K	280K	383K	383K	383K
Aggregate U2AF2_control	2.60M	141046	5	9.5	24bp	21bp	452.38 RPK	51.59%	56977	77423	141K	141K	141K

Supplemental Figure 3. Characteristics of hnRNPA1, SRSF1 and U2AF peaks.

Read density and length distributions for peaks called from iCLIP data using CLIPper

for hnRNP A1, SRSF1, and U2AF2 for control and hnRNP A1 overexpression

conditions.

hnRNP A1 Peaks (Control Cells)
Read Densities Around Peaks

SRSF1 Peaks (Control Cells)
Read Densities Around Peaks

hnRNP A1 Peak (Over-expression cells)
Read Densities Around Peaks

SRSF1 Peak (Over-expression cells)
Read Densities Around Peaks

U2AF2 Peaks (Control Cells)
Read Densities Around Peaks

U2AF2 Peak (Over-expression cells)
Read Densities Around Peaks
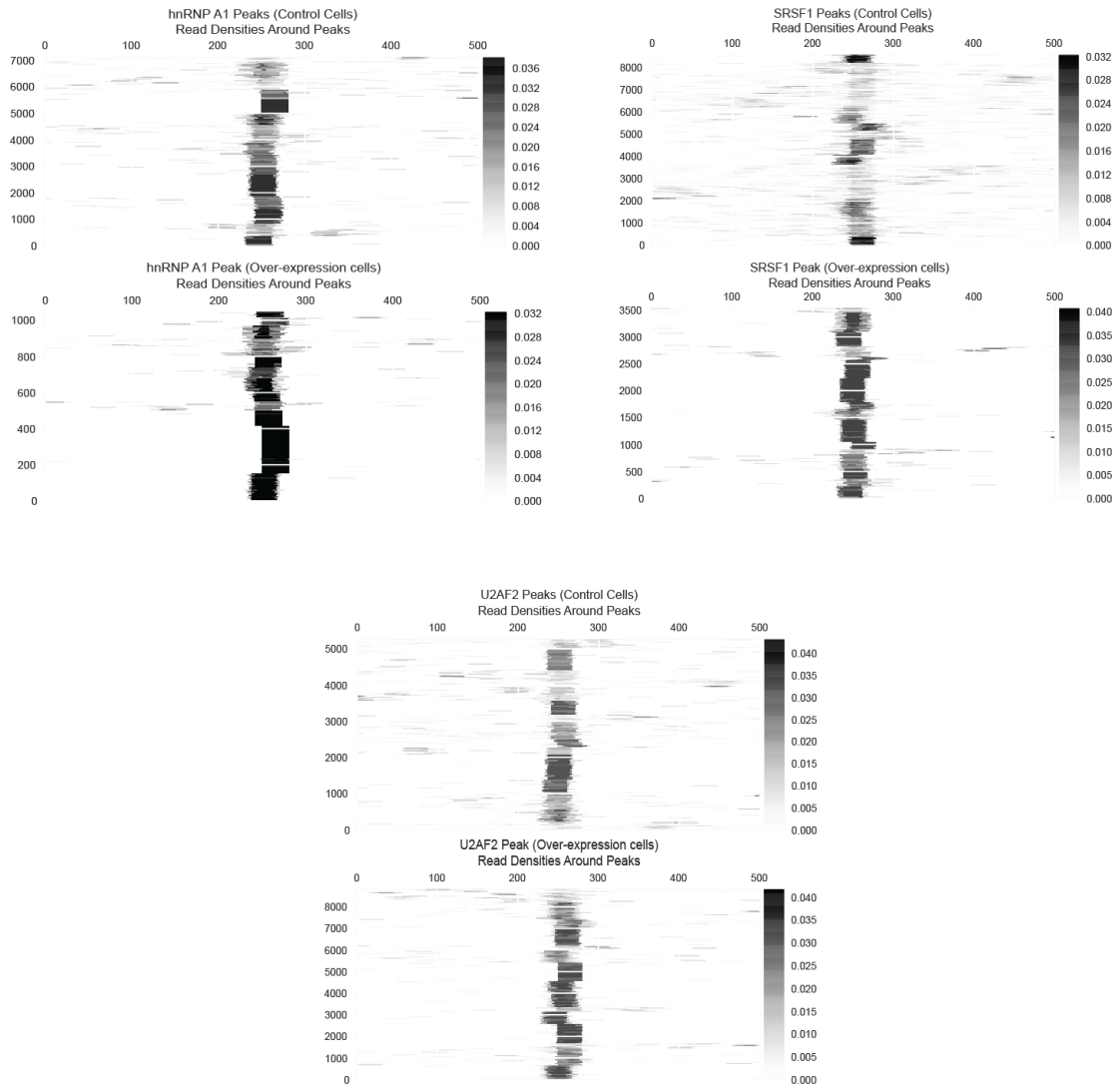
Supplemental Figure 3

Supplemental Figure 4: Distribution of SRSF1, U2AF2, and hnRNP A1 peaks near 3' and 5' splice sites. Peak distribution (not normalized) of SRSF1, U2AF2 and hnRNP A1 iCLIP peaks called by CLIPper under control and hnRNP A1 overexpression conditions near 3' (left column) and 5' (right column) splice sites.
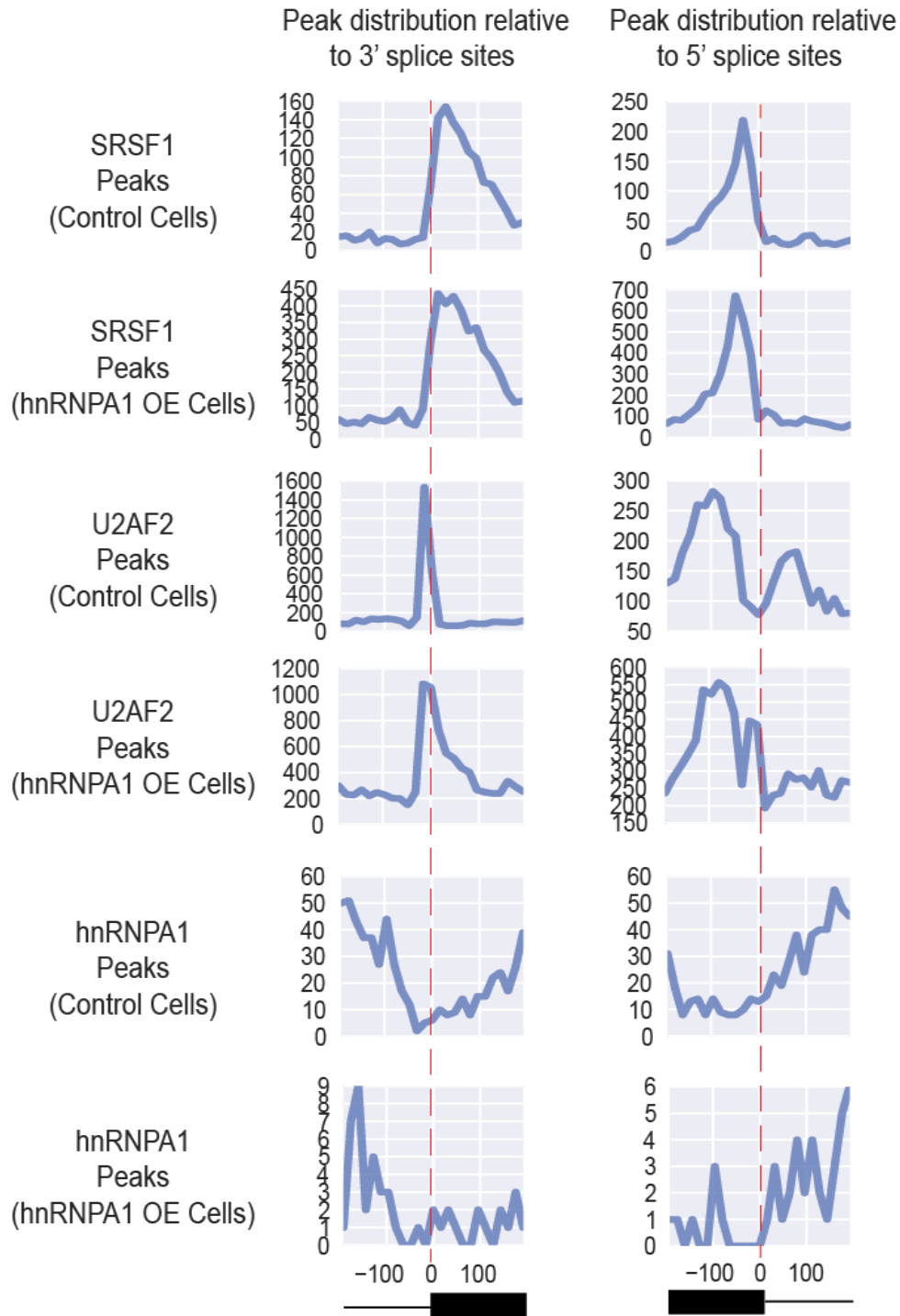
Supplemental Fig. 3

Supplemental Figure 5: Additional HOMER consensus motifs for hnRNP A1, SRSF1, AND U2AF65. (A) Motifs enriched in non-Alu peaks. **(B)** Motifs enriched in Alu-derived peaks.

A.

hnRNPA1 (control cells)

hnRNPA1 (hnRNPA1 over expression)

SRSF1 (control cells)

SRSF1 (hnRNPA1 overexpression)

U2AF2 (control cells)

U2AF2 (hnRNPA1 overexpression)

B.

hnRNPA1 (control cells)

hnRNPA1 (hnRNPA1 over expression)

SRSF1 (control cells)

SRSF1 (hnRNPA1 overexpression)

U2AF2 (control cells)

U2AF2 (hnRNPA1 overexpression)
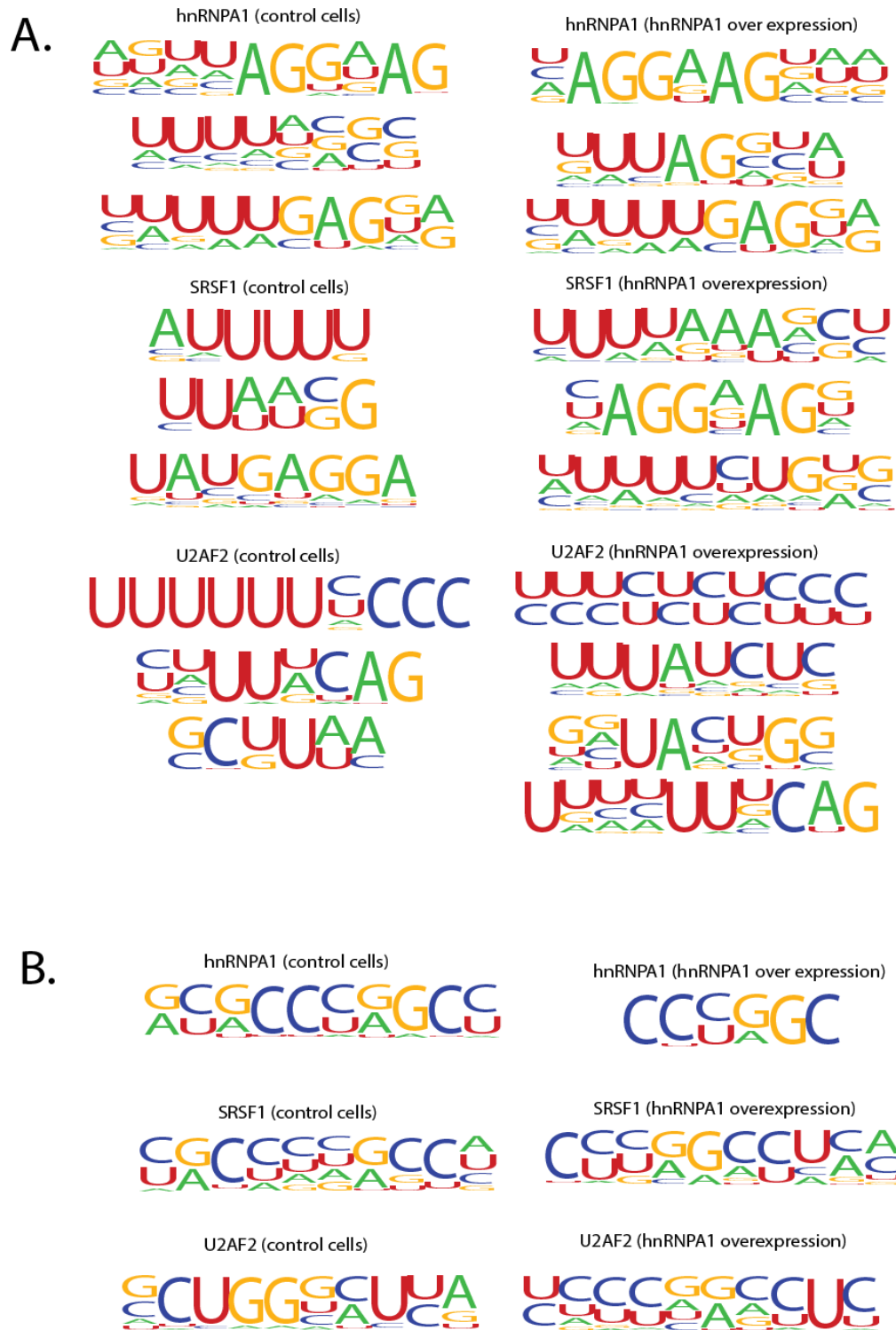
Supplemental Fig. 4

Supplemental Figure 6: MISO analysis of RNA-seq from iCLIPs reveals hnRNPA1-dependent splicing changes. Bar graph representing number of total splicing events detected using MISO analysis for splicing changes. Out of more than 250 events, the majority of events were identified as events involving increased skipping of alternative exons. The table represents a breakdown of these events were changes in U2AF2 crosslinking are also detected, as well as those changes in which a redistribution of U2AF2 to a nearby Alu elements.
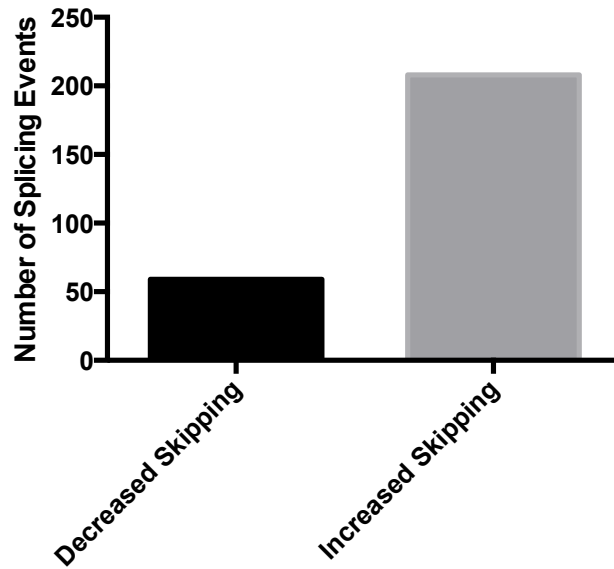
Table 1. summary hnRNPA1-dependent changes in exon skipping and U2AF2 positioning

	hnRNPA1-dependent change in U2AF2 crosslinking near 3'ss			
	Total # of Events	Increased U2AF2	Decreased U2AF2	Redistirbution to Alu
hnRNPA1-dependent exon inclusion	9	3	6	7
hnRNPA1-dependent exon skipping	74	30	44	40

Supplemental Figure 7: Effects of hnRNP A1 over-expression on the proportion of hnRNP A1 and SRSF1 crosslinking sites in *Alu* RNA elements. Scatter plot of all human cassette exons measuring the proportion of hnRNP A1 and SRSF1 iCLIP crosslinks found within *Alu* elements relative to the total number of crosslinks observed throughout the alterative event.

A. *hnRNPA1*

B. *SRSF1*
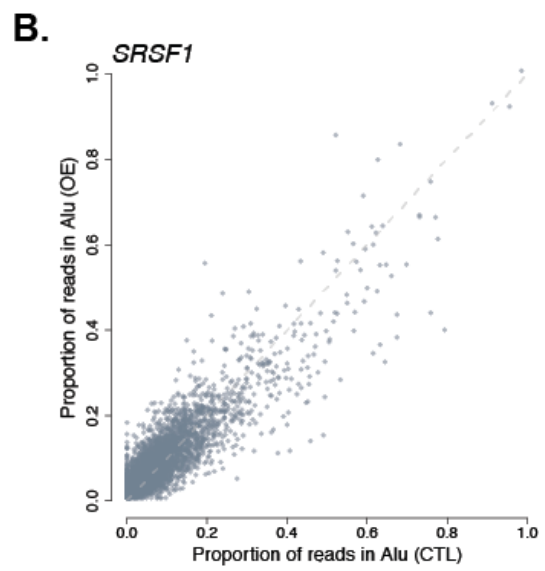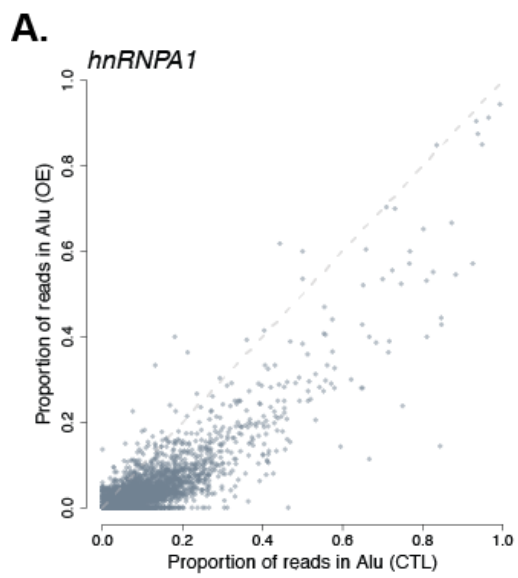
Supplemental Figure 8: Overexpression of hnRNP A1 does not confer appreciable changes in hnRNP C1/C2 expression in HEK293T cells. Western blot of TREX HEK293T control cells and those containing a tetracycline-inducible T7-tagged version of hnRNP A1 (T7-A1). After 24 h cell lysate were subjected to immunoblotting for hnRNP A1, T7 protein tag, hnRNP C1/C2, and EWS as a loading control. Experiment was performed in duplicate
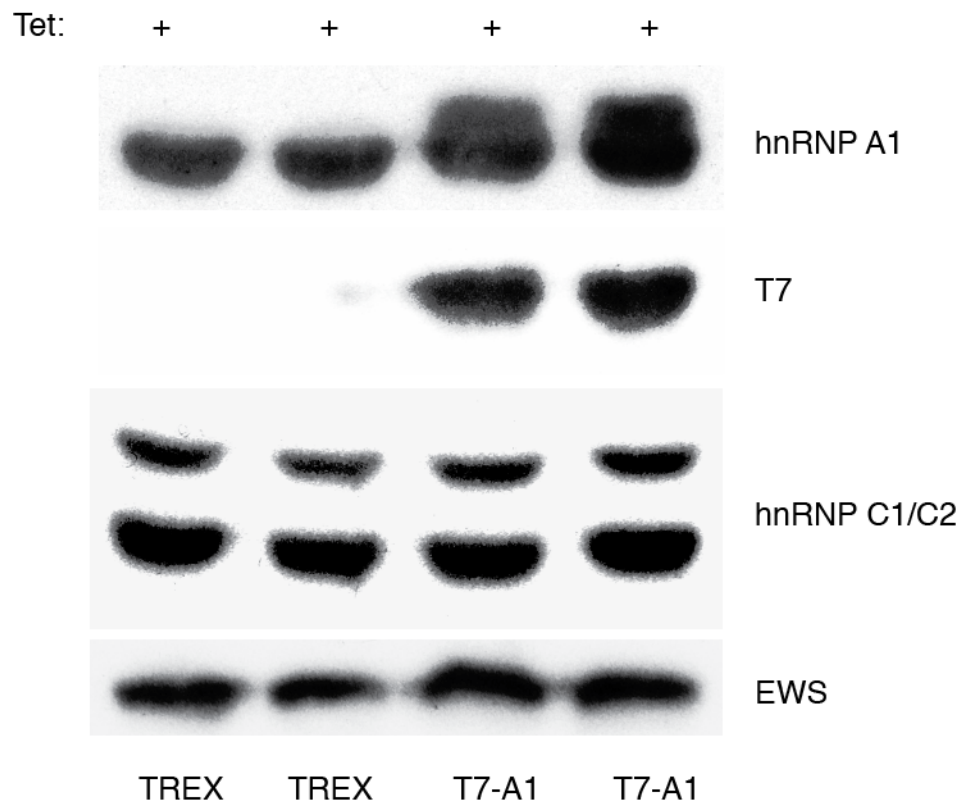
Chapter 4

IGF2BP3 controls cancer cell invasiveness by modulating

RISC function

Abstract

The Insulin-like growth factor 2 mRNA binding protein 3 (IGF2BP3) is an onco-fetal RNA binding protein that is aberrantly expressed in diverse cancers including pancreatic ductal adenocarcinoma and B-acute lymphocytic leukemia. However, its regulatory targets and its role(s) in pathogenesis are poorly understood. Here we demonstrate that IGF2BP3 contributes to the invasive phenotype of pancreatic ductal adenocarcinoma cells (PDAC) and is involved in the direct regulation of at least 410 transcripts encoding proteins with functions related to cell migration, proliferation and adhesion. RNA binding site maps generated by individual-nucleotide resolution crosslinking immunoprecipitation (iCLIP) in three distinct cancer cell lines revealed positional overlap of IGF2BP3 binding sites with the target sequences of several cancer-associated miRNA families thereby highlighting a role for IGF2BP3 in microRNA-mediated gene expression regulation in oncogenesis. Our results demonstrate for the first time that IGF2BP3 not only antagonizes microRNAs but can also promote the association of the RNA induced silencing complex (RISC) with specific transcripts to trigger their decay. Collectively, we propose that IGF2BP3 controls a post-transcriptional malignacyassociated RNA regulon by serving as a bimodal regulator of (both positive and negative modulation) microRNA functionality in cancer.

Introduction

Post-transcriptional gene regulation is extensively controlled by trans-acting RNA binding proteins (RBPs) and noncoding RNAs, including ~22 nt micro(mi)RNAs (Gerstberger et al., 2014b). On a global level RBPs and miRNAs are hypothesized to coordinately regulate the expression of mRNAs encoding functionally related proteins. These post-transcriptional regulons are hypothesized to function analogously to polycistronic operons, but at the RNA level (Blackinton and Keene, 2014; Keene, 2007). The regulatory activities of RBPs and miRNAs often converge on the 3' untranslated regions (3'UTRs) of mRNAs where their binding sites can be closely associated, leading to competitive binding or cooperative interactions (Jens and Rajewsky, 2015; Moore, 2005). For example, the RBP HuR often binds to 3'UTRs and 24% of its target sites overlap predicted miRNA target sites, suggesting that many target transcripts may be co-ordinately regulated by miRNAs and HuR (Uren et al., 2011). A possible consequence of binding site juxtaposition is that RBPs can modulate miRNA-mediated gene regulation and vice versa (Ho and Marsden, 2014; Jiang et al., 2013). Although several antagonistic and cooperative examples have been described (Dassi et al., 2013; Kedde et al., 2010; Kim et al., 2009), the underlying molecular mechanisms and generality of this interplay are unclear.

The Insulin Growth Factor 2 mRNA Binding Proteins (IGF2BP1, -2 and -3) are a family of structurally and functionally related RBPs with tissue-specific and developmentally regulated expression patterns (Hansen et al., 2004; Yaniv and

Yisraeli, 2002). IGF2BP3 is of particular interest because it is undetectable in most adult tissue but strongly expressed in embryos and in diverse tumor types (Mueller-Pillasch et al., 1999; Wagner et al., 2003). For example, it is up-regulated in 90% of pancreatic ductal adenocarcinomas, suggesting that it may have a role in initiation or progression of cancer (Esposito et al., 2014; Findeis-Hosey and Xu, 2011; Kobel et al., 2009; Schaeffer et al., 2010). Several groups have proposed that elevated IGF2BP3 expression is prognostic of decreased survival rates and malignancy in PDAC, colorectal cancer, ovarian and B-acute lymphocytic leukemia (B-ALL) patients (Kobel et al., 2009; Lochhead et al., 2012; Stoskus et al., 2011). Indeed, Taniuchi et al. demonstrated aberrant IGF2BP3 expression promotes metastasis in xenograft assays in nude mice (Taniuchi et al., 2014). It is likely that this role of IGF2BP3 in cancer metastasis mirrors a normal function of IGF2BP3-mediated cell migration during embryogenesis (Li et al., 2014b). Taken together, the clinical and experimental data demonstrate that IGF2BP3 expression promotes tumor metastasis and invasion *in vivo*.

As evidence mounts that IGF2BP3 is a *bona fide* pathoprotein, it is remarkable that several fundamental questions remain unanswered. For example, despite being implicated in several cytoplasmic steps of post-transcriptional gene expression such as mRNA stability, localization and translation, the underlying molecular mechanisms are still enigmatic (Gu et al., 2012; Jonson et al., 2014; Nielsen et al., 1999; Vikesaa et al., 2006). Additionally, the RNA targets of IGF2BP3 in cancer cells and the specific sequences to which it binds are also open questions.

Recently, transcripts bound by exogenous, epitope tagged IGF2BP3 (using PAR-CLIP) and also the endogenous protein (RNA immunoprecipitation (RIP)) in HEK293 and the PDAC cell line S2-013 have been reported (Hafner et al., 2010; Taniuchi et al., 2014). However neither study provided detailed insights into IGF2BP3-mediated mechanisms of post-transcriptional gene regulation. Here, we elucidate global mRNA targets and the mechanisms of IGF2BP3-dependent regulation in established pancreatic ductal adenocarcinoma (PDAC) cell models PANC1 and PL45 (Deer et al., 2010; Li et al., 2013). Using a combination of genome wide approaches including RIP-seq, iCLIP and RNA-Seq we show that IGF2BP3 preferentially binds the 3'UTRs of mRNA transcripts encoding proteins with functions in many cancer-related pathways. Interestingly, analysis of the *in situ* binding map for IGF2BP3 revealed that it predominantly interacts with its mRNA targets at positions that significantly overlap target sites of miRNAs with functional links to malignancy and invasiveness. This is suggestive of co-regulation or cross-talk of IGF2BP3 with miRNA mediated gene expression regulation in cancer. Indeed, we demonstrate that IGF2BP3 modulates the levels of its target gene transcripts, at least in part by affecting their association with the RISC complex. Taken together, our results establish a role for IGF2BP3 as an important regulator of an extensive posttranscriptional regulatory network associated with tumor metastasis.

Results

IGF2BP3 promotes PDAC cell invasion *in vitro* and modulates a malignancy-associated RNA regulon.

We used an *in vitro* invasion assay to determine whether depletion of IGF2BP3 affects the invasive behavior of two distinct PDAC cell lines, PANC1 and PL45. IGF2BP3-depleted cells exhibited significantly reduced cell invasion as compared to control cells (Student's t-test, $p<0.01$ for both lines, Figure 1A and supplementary Figure 1). To understand the molecular basis of IGF2BP3 on PDAC cell invasiveness, we analyzed the gene expression profiles of IGF2BP3-depleted and control PANC1 cell lines using high-throughput RNA sequencing (RNA-Seq). 2,795 genes exhibited greater than 2 fold changes in their steady state mRNA levels upon IGF2BP3-depletion. A strong enrichment for protein functions related to regulation of the extracellular matrix, cell motility and migration, cellular adhesion, signal transduction pathways and integral membrane proteins was observed among the differentially expressed mRNAs (Supplemental figure 2B). RNA immunoprecipitation and high throughput sequencing (RIP-seq) was employed to identify mRNAs that are associated with IGF2BP3 *in vivo*. 2,223 transcripts from the PDAC cell line PL45 are significantly enriched in the anti-IGF2BP3 immunoprecipitations relative to the control sera (Supplemental figures 3, 4A and supplemental table 2). To identify the targettranscripts that are most likely directly

87

regulated by IGF2BP3, we focused on genes whose expression was affected upon

IGF2BP3 depletion and were also found to be associated with IGF2BP3 in RIP-Seq

assays (Figure 1B). This approach revealed 410 transcripts which are enriched in

genes whose protein products function in cellular migration, epithelial cell

proliferation, the actin cytoskeleton, signal transduction pathways, cell adhesion and

development (adjusted P value < 0.05, Tables 1 and 2, supplemental table 3).

A single nucleotide resolution map of IGF2BP3-RNA interactions

To elucidate the *in situ* RNA binding specificity of IGF2BP3 in cancer cells

we performed iCLIP analysis on the PL45 cell line. Additionally, we performed

iCLIP analysis of IGF2BP3 on human B-acute lymphocytic leukemia (B-ALL) cell

lines REH and RS4;11; representative of another cancer type that exhibits elevated

IGF2BP3 expression. Nuclease-sensitive IGF2BP3-RNA complexes were

immunopurified from whole cell extracts of UV-irradiated PANC1 (not shown),

PL45, RS4;11 and REH cells (Supplemental Figure 5). iCLIP sequencing libraries

were prepared from two independent immunoprecipitation experiments from PL45,

and three replicates from REH and RS4;11 cells. We identified 244 and 335 mRNA

targets that directly crosslinked with IGF2BP3 in the iCLIP assay in PL45 and B-

ALL cells, respectively. A complete description of the IGF2BP3 mRNA targets in B-

ALL cells will be published elsewhere. We observed that IGF2BP3-associated

crosslinks in PL45 and in the B-ALL cell lines occurred in higher frequency within

exonic sequences as compared to simulated crosslinking sites randomly distributed throughout the genome. Importantly, this crosslink distribution is specific since it differs markedly from the distribution of hnRNPA1, another RNA binding protein (analyzed in parallel in HEK293 cells) that preferentially binds to intronic sequences (Figure 2A). Overall, our analysis revealed that the IGF2BP3 binding sites in exons are preferentially distributed in the 3'UTRs and/or last exons of the target transcripts (Figure 2B) particularly in regions proximal to the translational stop codon (Figure 2C). Interestingly, this pattern is similar to the exonic binding of hnRNPA1 where also crosslinks are frequently located at the 3'UTRsof the targets. However, the exonic hnRNPA1 binding sites are more distal to the stop codon compared to the IGF2BP3 sites in the 3'UTRs of their respective targets (Figure 2C).

Next, we used IGF2BP3 crosslink sites in PL45, REH and RS4;11 cells to identify over-represented motifs using a pentamer-clustering approach (See Methods). Of four over-represented motifs detected within a 10 nt window of each crosslinking site, only one closely resembles a previously defined IGF2BP3 consensus binding sequence derived from an *in vitro* binding assay and PAR-CLIP (Hafner et al.. 2010b). By contrast the remaining motifs appear to be distinct from these previously characterized binding sites and provide evidence for cell type specific RNA binding specificity. The two motifs identified using B-ALL cell lines REH and RS4;11 are more similar to each other than to the PL45 motifs, however common pentamers are enriched in all three cell lines suggesting that there is overlap between the binding specificity for IGF2BP3 in three different cell lines and also

across different cancer types (Figure 2D and E and supplemental table 7). The

apparent dual specificity of IGF2BP3 consensus sites in PL45 cells is consistent with

previous results suggesting a bipartite binding preference for IGF2BP family proteins

(Gu et al.. 2012). Additionally, we used the RIP-Seq data to cross-validate and

expand the results obtained from iCLIP in PL45 cells. Approximately 70% (166 of

244 genes) of transcripts identified by iCLIP overlap with those detected by RIP-Seq

analysis (Supplemental Figure 4B; $P < 1.0$ e-90 Fisher's Exact test). Consistent with

these results and those described for RNA-seq and RIP-seq comparisons

(Supplemental Figure 2B and Tables 1,2), Gene Ontology enrichment analysis of the

166 genes reveals a functionally coherent set of IGF2BP3 targets related to cellular

adhesion, migration and remodeling of the extracellular matrix. Figure 2F-H shows

genome browser screen shots of representative examples of 3'UTRs that were

identified in both iCLIP and RIP-seq experiments as IGF2BP3 targets. RNA-seq

analysis for differential expression from both Pacn1 and PL45 cells cross-referenced

with RNA immuprecipitation sequencing (RIP-seq), revealing a total of 164 genes

that overlapped from the two RIP-seq experiments and showed differential expression

in either Panc1 cells, PL45 cells, or both (Supplemental Figure 7A). Validation of

differential expression of a handful of these genes was performed as well showing

similar effects compared to the DEseq analysis (Supplemental Figure 7B). GO

analysis of these genes showed enrichment for biological processes (Supplemental

Figure 7C). involved with cell motility, migragtion, and locomotive movement, while

KEGG analysis showed subsets of genes associated with known cancer pathways

(Supplemental Figure 7D).

IGF2BP3 binding sites overlap miRNA target sites

Based on the positional distribution of IGF2BP3 crosslinking sites within

3'UTRs (Figure 2B and C), we hypothesized that other prominent cis-regulatory

features associated with 3'UTR-mediated gene regulation might also overlap with

IGF2BP3 binding sites. To investigate this possibility, we examined the distribution

of IGF2BP3 crosslinking sites relative to miRNA target sites in the three different

cancer cell lines. After correcting for the uniform background distribution of

simulated crosslink sites, we find that IGF2BP3 crosslink density in all three cell lines

is highly enriched within a 25 bp window centered on predicted miRNA target

sequences (Figure 3A). By contrast, the density of hnRNPA1 crosslinking sites are

uniformly distributed relative to miRNA target sites. These observations indicate that

IGF2BP3 binding is specifically enriched over predicted miRNA target sites.

To determine if IGF2BP3 actually shares sequence specificity with miRNAs

we independently scored 221 miRNA seed-pairing sites (Xie et al., 2005) by their

similarity to IGF2BP3-bound pentamers derived from iCLIP data derived from

PDAC and B-ALL cells (see Methods). For 3'UTRs containing both IGF2BP3 iCLIP

and miRNA target sites we observed a greater degree of sequence similarity between

the IGF2BP3 motif and miRNA target site sequences as compared to the miRNA

target sites present in 3'UTRs that lack evidence of IGF2BP3 interactions (Figure 3B

and C, supplemental table 8; **P**<0.00153 Wilcoxon rank-sum test). We noted that specific microRNA target sites are over represented in both the iCLIP data from PL45 cells and in the 3'UTRs of transcripts that are differentially expressed upon IGF2BP3-depletion in PANC1 cells (Figure 3C, grey boxes). The absence of some microRNA target sites in the PL45 iCLIP data also suggests cell type-specific RNA recognition elements. To test this hypothesis, we calculated the number of shared IGF2BP3 mRNA targets (identified in B-ALL or PL45 cell lines) among all pairs of miRNA family gene sets. To investigate the extent of redundancy among these sets, we performed unsupervised clustering, plotting individual members and cluster exemplars by the number of B-ALL or PL45 iCLIP mRNA targets (Figure 3D and supplemental table 8). We observed that miRNA families intersecting IGF2BP3 iCLIP targets are optimally placed into eight clusters containing distinct sets of mRNA targets (Figure 3D). Some of the clusters appear to represent cell type specific iCLIP targets, such as those corresponding to miR-19A and -19B (TTTGCAC, Figure 3D) in B-ALL cells and miR-200B, -200C and -429 in PDAC cells (CAGTATT, Figure 3D) and are enriched in iCLIP and differentially expressed genes in PDAC cells (Figure 3D). By contrast, clusters corresponding to miR-1, -9, -128A, -128B and -206 (ACATTCC, ACCAAAG and CACTGTG) have similar numbers of targets in both PDAC and B-ALL cells and significant similarity with the IGF2BP3 motif score (Figure 3C and D). These data highlight the potential reach of IGF2BP3 in many distinct post-transcriptional regulatory pathways and suggest an intriguing mechanism for IGF2BP3 action by modulating miRNA-mediated gene regulation. To test this

92

hypothesis, we determined the impact of IGF2BP3-depletion on steady-state levels of transcripts in which miRNA- and IGF2BP3-target sites overlap, compared to those transcripts that are only targeted by miRNAs in PANC1 cells. We observed reduced steady-state expression levels in IGF2BP3-depleted cells relative to control for transcripts containing both IGF2BP3 and miRNA targets sites in their 3'UTRs as compared to transcripts containing only miR-target sites (Figure 3E and supplemental table 8). These data suggest that IGF2BP3 may attenuate miRNA-mediated decay.

IGF2BP3 promotes Ago2-mRNA interactions

The data presented above suggest that IGF2BP3 can antagonize miRNAs and increase steady state levels of specific transcripts (Figure 3). Interestingly, we also identified many transcripts with elevated steady state levels in IGF2BP3-depleted cells, suggesting that some transcripts are destabilized by IGF2BP3. For example, the mRNA encoding CLDN1, a tight junction component, is dramatically stabilized in PL45 cells depleted of IGF2BP3 (Supplemental figure 6). By contrast the mRNA encoding CD44, which is stabilized by IGF2BP3 (Vikesaa et al., 2006), has an accelerated decay rate in IGF2BP3-depleted cells (Supplemental figure 6). To determine if IGF2BP3-dependent decreases in steady state mRNA levels are also mediated by modulation of the RNA-induced silencing complex (RISC), we performed Ago2-RIP from control and IGF2BP3-depleted PANC1 cells (Figure 4A). This assay allowed us to investigate the extent to which IGF2BP3 positively or

negatively influences the association of the RISC with mRNA. Subsequently we

quantified co-precipitation of the IGF2BP3 target CLDN1, ZFP36L1 and DCBLD2

by RT-qPCR. As controls for the Ago2-RIP, we used the TBP and IGF2BP3 mRNAs.

TBP mRNA was not detected in any of the IGF2BP3 protein-RNA interaction assays,

and its steady-state mRNA levels were not affected by IGF2BP3 depletion. By

contrast, IGF2BP3 mRNA is targeted by shRNA, which we reasoned, should increase

the association of this transcript with RISC. A representative immunoprecipitation

from control or IGF2BP3-depleted cells (NT or KD, respectively) with either

nonspecific rabbit IgG or an anti-Ago2 serum is shown in Figure 4B. We first

compared cytosolic levels of TBP, IGF2BP3, ZFP36L1, DCBLD2 and CLDN1

mRNAs in control or IGF2BP3 depleted cells. As expected from the RNA-seq data,

the steady state level of TBP mRNA is unaffected by IGF2BP3 depletion, whereas

the mRNAs for ZFP36L1, DCBLD2 and CLDN1 are significantly elevated,

respectively (Figure 4C, E, G). We next assayed relative levels of CLDN1, ZFP36L1,

DCBLD2, IGF2BP3 and TBP mRNA in the Ago2 immunoprecipitates from both

control and knock down cells and compared them to control IPs performed with

rabbit IgG (Figure 4D, F, H, J and L). All five mRNAs tested were significantly

enriched relative to the nonspecific control antibody. However, we observed a

significant reduction in the relative levels of ZFP36L1, DCBLD2 and CLDN1

mRNAs in the Ago2 IP from IGF2BP3-depleted cells relative to control (Figure 4H,

J, L, P< 0.0161, 0.0135 and 0.0144, respectively). By contrast Ago2 antibodies

immunoprecipitated similar levels of IGF2BP3 and TBP mRNA from both cell lines

94

(Figure 4D and F). Together these data suggest that IGF2BP3 influences the association of the RISC with CLDN1, ZFP36L1 and DCBLD2 transcripts or may alter RISC function.

Discussion

In this study we report a coherent set of direct mRNA targets for IGF2BP3 in pancreatic cancer cell lines and provide new insight into the role of IGF2BP3 in gene regulation. Our results demonstrate that IGF2BP3 has different effects on the steady-state levels of specific transcripts. Based on analysis of high-throughput protein-RNA interaction data coupled with gene expression profiling and Ago2 RIP, we propose that IGF2BP3 modulates miRNA function to control the expression of a malignancy-associated RNA regulon. (graphical abstract). The subcellular localization of all three IGF2BP paralogs is predominantly within cytoplasmic foci, which may correspond to RNA transport granules (Bell et al., 2013; Wachter et al., 2013). With recent studies, a model is emerging in which the IGF2BP family promotes the formation of stable RNA-protein complexes where transcripts are protected from translation and miRNA medicated decay (Jonson et al., 2014). However, experiments presented in figure 4 demonstrate that depletion of IGF2BP3 correlates with increased steady state CLDN1 mRNA levels and its' decreased association with Ago2. Taken together with the work of Jønson et al., our data demonstrates that IGF2BP3 may have transcript specific effects on the association with RISC. Several studies demonstrate that IGF2BP3 can

promote mRNA stability, as described for the CD44 mRNA (Vikesaa et al., 2006).

One mechanism suggested by our iCLIP data is that in certain contexts miRNAs and

IGF2BP3 may antagonize each other by competing for binding site within 3'UTRs.

The high degree of sequence similarity between the consensus IGF2BP3 RNA

recognition elements and target sites of several different miRNA families support this

hypothesis (Figure 3) as does the concomitant decrease in expression of mRNAs that

are targeted by both regulatory systems (Figure 3E). Similar modes of regulation have

been described for RBPs such as HuR and Dnd1 (Kedde et al., 2007; Kim et al.,

2009). However, our experiments also reveal an unexpected role for IGF2BP3 in

promoting the association of mRNAs with Ago2 and thereby triggering mRNA

decay. Given that several of the miRNA families presented in Figure 3C correspond

to those with known roles in cancer cell biology it is likely that the positive and

negative modulation of RISC contributes directly to malignancy in both PDAC and

B-ALL cells (Mogilyansky and Rigoutsos, 2013; Yu et al., 2010).

In this study, we identify a new set of IGF2BP3 mRNA targets that are

involved in cancer-related pathways including focal adhesions, adherens junctions,

regulation of the actin-cytoskeleton and cell migration. Additionally, we describe a

single nucleotideresolution RNA map that reveals positional- and sequence

specificity of IGF2BP3-RNA interactions that are similar in PDAC and leukemia cell

lines. Our work provides evidence that IGF2BP3 and miRNAs converge on the

3'UTRs of these transcripts to coordinately up- and down-regulate programs of gene

expression that are associated with malignancy and invasiveness. The results further

show that this dual activity of IGF2BP3 likely operates through its ability to protect transcripts from – or enhance – miRNA-mediated post-transcriptional gene silencing.

Experimental Procedures

Cell lines and constructs

Human PDAC and B-ALL cell lines were obtained from American Type Culture Collection and cultured following recommended conditions (Manassas, VA). The PDAC cells were stably transfected with plasmids expressing either shRNA targeting IGF2BP3 or a non-targeting control (Santa Cruz Biotechnologies, Santa Cruz, CA) using Lipofectamine 2000 (Life Technologies) and selected with puromycin (1 ug/mL). Multiple clones were screened for IGF2BP3 depletion and phenotypic characterization.

Invasion Assay

In vitro invasion assay was performed using Transwell inserts with 8.0-**μ**m pore size (BDbiosciences). After optimization, 5 x 10$_5$ PDAC cells were suspended in serum-free medium and seeded onto the Transwell inserts pre-coated with matrigel insert. The Transwell inserts were then placed into 6-well plates containing supplemented medium containing 10% FBS. After a 72 hour incubation, the upper

surface of the Transwellinserts was wiped with a cotton swab and the invaded cells were fixed and stained with Diff-Quick stain (IMEB, San Marcos, CA). The number of invading cells was counted under an inverted microscope (x50) in four randomly selected fields per well.

Western blot

Western blotting was performed as previously described (Sterne-Weiler et al.. 2011). The following antibodies were used: anti-IGF2BP3 (clone D7, Santa Cruz Biotechnology), anti-GAPDH (SIGMA), anti-hnRNPA1 (clone 4B10, Santa Cruz Biotechnology), anti-4EBP (Santa Cruz Biotechnology).

Gene expression profiling of PDAC cells by RNA-Seq

RNA was purified from both cytosolic fractions of IGF2BP3-depleted or control PANC1 cells using TRI-Reagent LS (Sigma), converted to double stranded libraries using the TruSeq polyA+ kit (Illumina Inc., San Diego, CA) and sequenced using Illumina HiSeq2000 platform. RIP-Seq assay 400 μl of Protein A sepharose (50% slurry) was washed five times with NT2 buffer (50 mM Tris–HCl pH 7.4, 1 M Tris–HCl, 150 mM NaCl, 1 mM $MgCl_2$, 0.05% NP40) and resuspended in 1 ml of NT2 plus 5% BSA and 10 μg of anti-IGF2BP3 (RN009P, MBL Inc.) or normal rabbit IgG from R&D (AB-105-C). Beads plus antibodies were incubated overnight at 4°C

with rotation and washed five times with cold NT2 buffer. Lysates were prepared from semi-confluent PL45 cells in polysomal lysis buffer (10 mM HEPES pH 7.0, 100 mM KCl, 5 mM MgCl2, 0.5% NP40, 2 mM dithiothreitol) containing proteinase and RNA inhibitors. After centrifugation for 10 min, supernatant was adjusted to 2 mg/ml and 6 ml of lysate were combined with the bead/antibody and rotated at room temperature for 3–5 h. Beads were washed five times with cold NT2. Proteins were extracted with 25 μl (20 mg/ml) proteinase K in 600 μl of 1Å~ buffer at 50°C for 30 min. Samples were vortexed for 1 min and beads pelleted by centrifugation. The supernantant was extracted with 700 μl of acid phenol–chloroform and precipitated with sodium acetate and isopropanol. RNA was recovered by centrifugation, washed and resuspended in 13 μl of RNase free water. Quantity and quality were checked with Nanodrop and Bioanalyzer. RNA samples were sequenced in a HighSeq 2000 machine according to Illumina protocols.

iCLIP assay

iCLIP was performed as previously described (Konig et al., 2010) (König J et al.., 2010), but with severalmodifications. After crosslinking and cell lysis, RNA was partially fragmented using low concentrations of Micrococcal nuclease, and IGF2BP3–RNA complexes were immunopurified with **α**-IGF2BP3 antibody immobilized on protein A–coated magnetic beads (Life Technologies). The oligonucleotides for reverse transcription contained two inversely oriented adaptor

regions adapted from the NEXTflex Small RNA-Seq library preparation kit (Bioo Scientific), separated by a BamHI restriction site as well as a barcode region at their 5′ end containing a 4-nt experiment-specific barcode within a 5-nt random molecular index to mark individual cDNA molecules. After circularization and linearization, cDNAs were then PCR-amplified using primers (Bioo Scientific) complementary to the adaptor regions and subsequently sequenced using Illumina HiSeq (or MiSeq) platforms.

Gene ontology analysis

To identify over represented classes of functionally related IGF2BP3-mRNA targets we annotated either differentially expressed genes, transcripts identified by iCLIP or RIP-Seq or combinations of the three approaches using ENRICHR (Chen et al., 2013). GO terms associated with biological processes, molecular function, cell compartment and KEGG pathways selected based on an Adjusted P value <0.05. Scatter plots were generated based on semantic similarity between the terms using REVIGO (Supek et al., 2011).

cDNA synthesis and qRT–PCR

cDNA was synthesized from RNA purified from cytosolic extracts, sucrose gradient fractions and RIP using the High capacity reverse transcriptase kit from ABI Scientific. qPCR was perfomed using a Roche Lightcycler 480 (Roche Diagnostics), Titanium Taq (Clontech) and Sybr green dye. Based on melting point analysis primers corresponding to TBP1, CLDN1, IGF2BP3 and 18S rRNA generated a single specific amplicon.. For normalization, TBP was used as a reference gene for comparisons in cytoplasmic RNA whereas 18S rRNA was used in the RIP and sucrose gradient experiments. Relative expression levels were determined using the $\Delta\Delta CT$ method with the Roche Lightcycler Analysis Software Package version 1.5 (Roche Diagnostics). All experiments use a minimum of three technical replicates per sample and at least three biological replicates per analysis. Statistical significance was determined by comparing the mean normalized ratios of each mRNA using a nonparametric T-test (Prism6, GraphPad).

Ago2 RNA immunoprecipitation

To immunoprecipitate Ago2 mouse monoclonal antibody 9E8.2 (Millipore) was tethered to Dynal protein A beads using a rabbit anti-mouse IgG bridging antibody (Jackson ImmunoResearch, Fc$_\gamma$ Fragment Specific). Mouse IgG (Pierce Biotechnology) was used as a negative control for RIP assays. Beads were washed

three times in lysis buffer (25 mM Tris-HCl at pH 8.0, 150 mM NaCl, 2 mM MgCl2, 0.5% NP-40, and 5 mM DTT) to remove the unbound ntibody. Whole cell extracts were prepared from IGF2BP3- depleted or control PANC1 cells by lysing cells on ice for 10 min in 1 mL of fresh lysis buffer with protease inhibitors (Complete Protease Inhibitor Cocktail Tablets, EDTA-free, Roche Applied Science) and RNasin (1/1000 dilution; Applied Biosystems), followed by sonication and centrifugation at 10,000g for 10 minutes. Following the washes in lysis buffer (as described in supplemental experimental procedures), beads were treated with 5U DNase 1 (Promega) for 10 minutes at 37ºC. An aliquot from each IP was used for western blot analysis and RNA was purified from the remainder using TRIzol LS (Invitrogen) as described in a previous publication (Li et al., 2009). Total RNA from cell lysates was isolated using the same procedure, and was also subjected to DNA digestion as described above prior to cDNA synthesis.

iCLIP data analysis pipeline

iCLIP analysis is described in the supplementary experimental procedures. RIP-Seq data analysis pipeline RIP-seq reads were mapped using RMAP (Karolchik et al., 2004; Kent et al., 2002; Pruitt et al., 2014; Smith et al., 2009; Smith et al., 2008; Uren et al., 2012) the hg19 human genome assembly, downloaded from the UCSC genome browser (Karolchik et al., 2004; Kent et al., 2002) website (http://genome.ucsc.edu). Regions outside of RefSeq (Pruitt et al., 2014) annotated

genes were masked out to improve mapping rates. We also constructed a database of junctions using RefSeq annotated transcripts to map junction-spanning reads. Reads were then counted to genes based on collapsing all RefSeq transcripts for each gene to a single super-transcript. Piranha (Uren et al., 2012) was run with 0.5 background fraction and significance threshold of 0.0001 to identify genes enriched in each of the IGF2BP3 IP replicates. The final target set was constructed by taking those genes that were significant in all three replicates.

RNA-seq data analysis pipeline

High-throughput RNA sequencing data generated by Illumina HighSeq 2500 and corresponded to ~50-100M reads per sample was mapped to hg19 build of the human genome (Feb. 2009 GRCh37, NCBI Build 37.1) using Bowtie and TopHat (Langmead et al., 2009; Trapnell et al., 2009). All data collection and parsing was performed with the Perl programming language, and statistical analyses with the R statistical language, version 2.14.1. All external library packages used are available on CPAN or CRAN, respectively. Differentially expressed genes were identified using DEseq (Anders and Huber, 2010) (Anders and Huber 2010).

Dissertation Author Contributions

JMH: Performed all iCLIP experiments in B-ALL and PDAC cells. Performed Ago2

RIP experiment. Performed cross-reference analysis of DEseq and RIP-seq data sets.

Helped with writing

This page has been left intentionally blank for printing purposes.

**FIGURE 1. IGF2BP3 interacts with a metastasis associated RNA regulon**.

A. Bar graph quantifying the invasiveness of control or IGF2BP3-depleted PANC1 or PL45 cells through matrigel filters relative to control filters. The bars represent an average of three independent experiments. Error bars correspond to standard deviation. Statistical significance estimated using unpaired T-test, ***** P**<0.001, *** P**<0.05. B. Venn diagram depicting overlap of gene targets identified by RIP-Seq and by RNA-seq analysis from IGF2BP3-depleted or control PANC1 cells.

A.

PANC1   PL45

B.

RIP-Seq   RNA-Seq
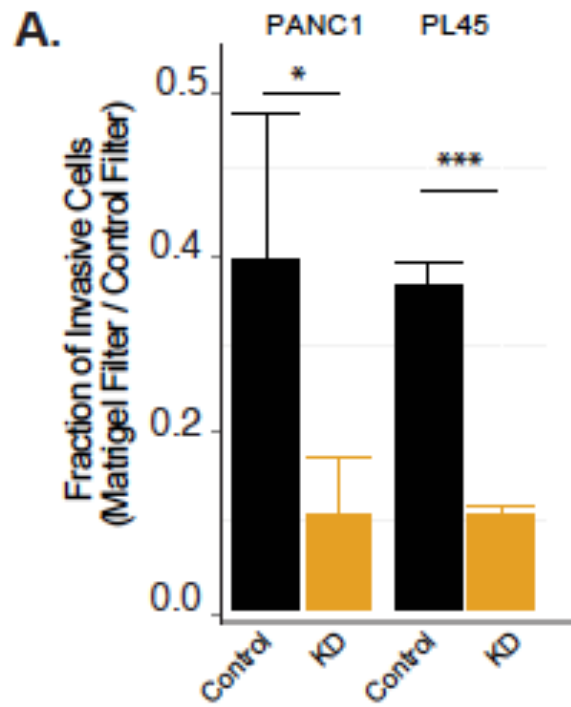
FIGURE 2. Global analysis of IGF2BP3-RNA interactions. A. Bar graph showing genomic distribution of IGF2BP3 (from PL45, REH and RS4;11 cells), hnRNPA1 (HEK293 cells) and simulated (hg19 background) crosslinking sites. The bars represent an average of 2 or 3 biological replicate experiments for PL45 and B-ALL cells, respectively. Standard deviations are indicated by error bars. B. Distribution of peaks called from IGF2BP3-, hnRNPA1- and simulated crosslinking sites within mRNA. C. Meta-analysis of IGF2BP3 and hnRNPA1 iCLIP crosslink sites relative to the stop codon. Relative crosslinking density (normalized count) is shown -/+ SEM. . D. Scatter plot of pentamer enrichment (observed/expected) around IGF2BP3 crosslink sites from B-ALL and PDAC cells. E. Over represented pentamers found near IGF2BP3 crosslinking sites in PL45, REH and RS4;11 cells. F-H. UCSC Genome Browser Snapshots of 3'UTRs bound by IGF2BP3 in PDAC cells. Read coverage from IGF2BP3 iCLIP, IGF2BP3 RIP and control IgG RIP assays are indicated. Vertebrate conservation and microRNA target sites (miR TS) are also depicted.

A. Region

B. Exonic Position

C.

D. Pentamer Enrichment

E.

**F.** ZFP36L1

G. DCBLD2

H. CLDN1
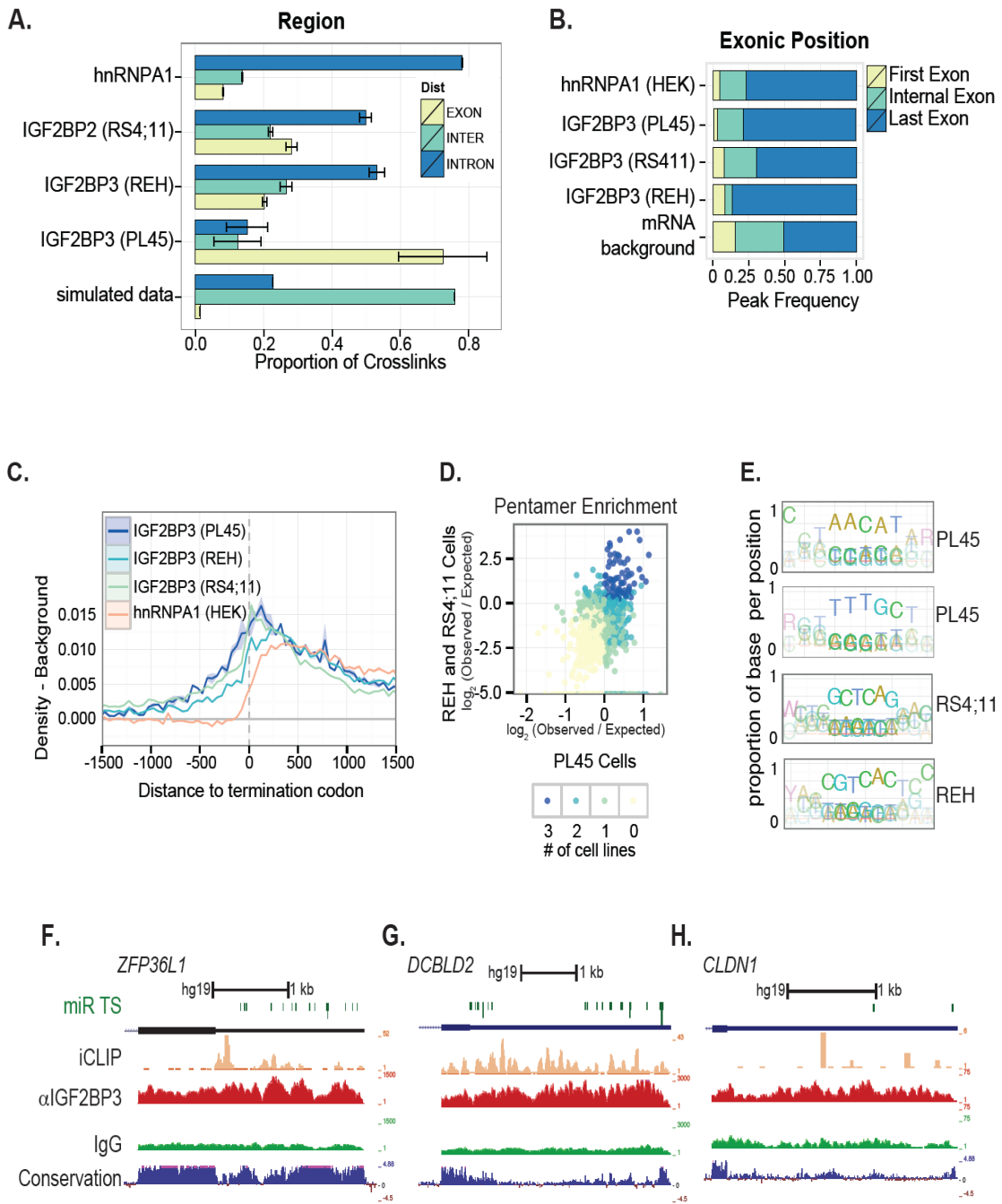
FIGURE 3. IGF2BP3 binding sites overlap miRNA target sites. A. IGF2BP3

crosslinking density from three different cell lines mapped relative to annotated

miRNA target sites. hnRNPA1 crosslinking sites from HEK293 cells are included as

a control. B. Box plot showing similarity score between IGF2BP3 consensus sites and

miRNA seed sites in IGF2BP3-bound transcripts (No overlap) or transcripts where

IGF2BP3 binding sites overlap miRNA target sites (IGF2BP3-miR target overlap). C.

Examples of miRNA families with significant sequence similarity between their

target site and the IGF2BP3 consensus motif (left panel). Enrichment of miRNA

target sites in 3'UTRs identified by iCLIP and by RNA-Seq analysis of control or

IGF2BP3-depleted PANC1 cells. D. Scatter plot showing the number of 3'UTRs

targeted by IGF2BP3 and specific clusters of miRNAs in PL45 and REH/RS4;11 cell

lines. E Box plot showing differentially expressed genes targeted only by miRNAs or

miRNAs that overlap IGF2BP3 binding sites.

A.
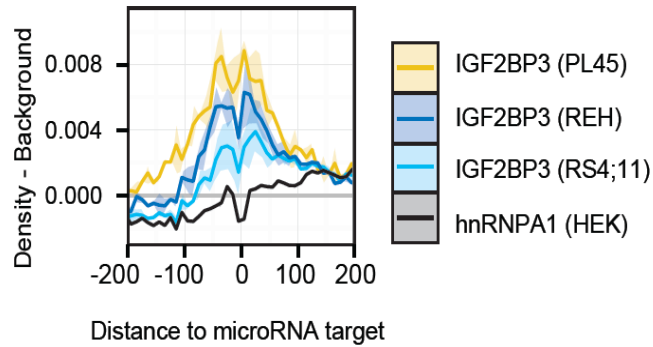
B.
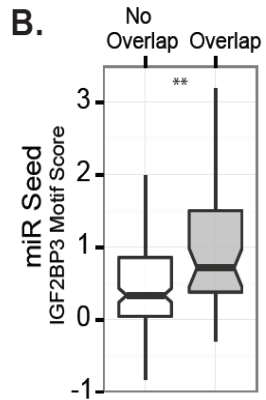
C.
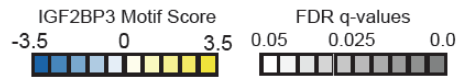
miR Target	Overlap Geneset Enrichment		microRNA Family
GCACTTT			**MIR-17-5P,-20A-B,-106A**
TTTGCAC			**MIR-19A&B**
CACTGTG			**MIR-128A&B**
GTGCCTT			MIR-506
CAGTATT			MIR-200B&C,-429
ACCAAAG			MIR-9
ACATTCC			MIR-1,-206
TGAATGT			MIR-181A-D

iCLIP PL45   PANC1 RNA-seq

D.
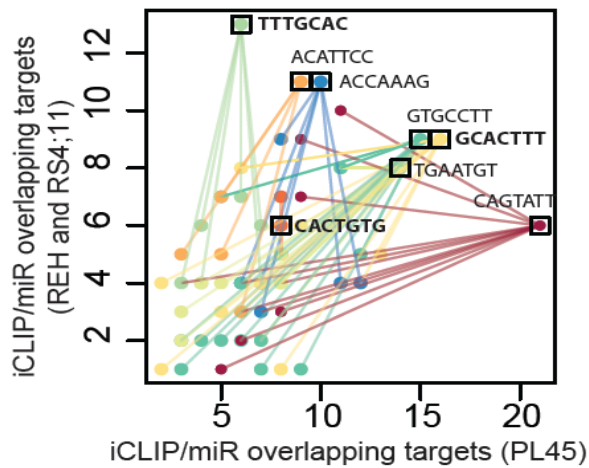
E.
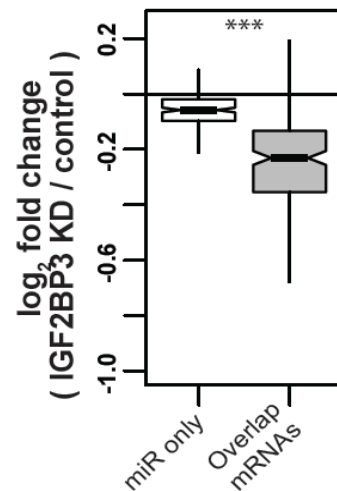
FIGURE 4. IGF2BP3 alters the association of iCLIP-elucidated mRNA targets with RISC. A. Schematic diagram of Ago2 RNA immunoprecipitation experiment. B. Western blot of Ago2 immunoprecipitation from control or IGF2BP3-depleted PANC1 cells. (C,E,G, I, K) RT-qPCR analysis of IGF2BP3, TBP CLDN1, ZFP36L1 and DCBLD2 cytoplasmic mRNA levels, respectively in control and knock down Panc1 cells. Relative quantification was performed using TBP as a reference gene. For relative quantification of TBP mRNA, the geometric mean of CFL, EGFR and RHOA expression was used as a reference. These data estimate the relative levels of each transcript in extracts used as inputs for the Ago2 or control IgG immunoprecipitation assay. (D,F,H,J,L) RT-qPCR analysis of RNA precipitated with α-Ago2 or nonspecific rabbit IgG (Ago2 RIP or IgG RIP, respectively). Relative quantification for each transcript was performed using 18S rRNA as a reference gene. Statistical significance was estimated for each comparison using a unpaired T-test (* $P<0.05$, ** $P<0.01$, *** $P<0.001$, **** $P<0.0001$).
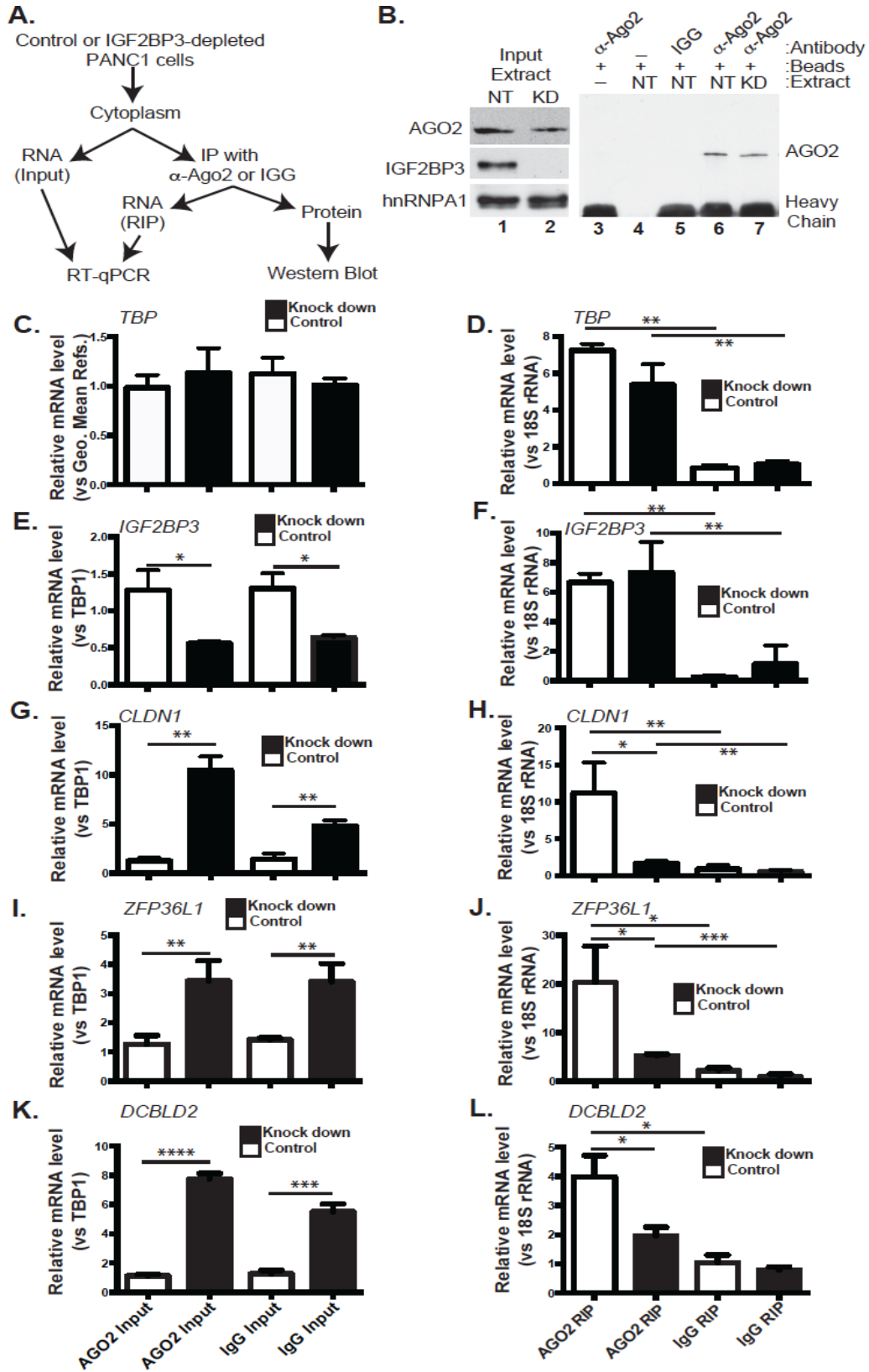
Graphical Abstract. IGF2BP3 iCLIP data and Argonaute 2-RIP data suggest IGF2BP3 may not only antagonize but also promote microRNA association with 3' UTR of specific transcripts via the RNA induced silencing complex (RISC) to trigger their decay. This suggests an intriguing model in which IGF2BP3 controls a post-transcriptional malignancy-associated RNA regulon by serving as a bimodal regulator of (both positive and negative modulation) microRNA functionality in cancer.

mRNA Stabilization

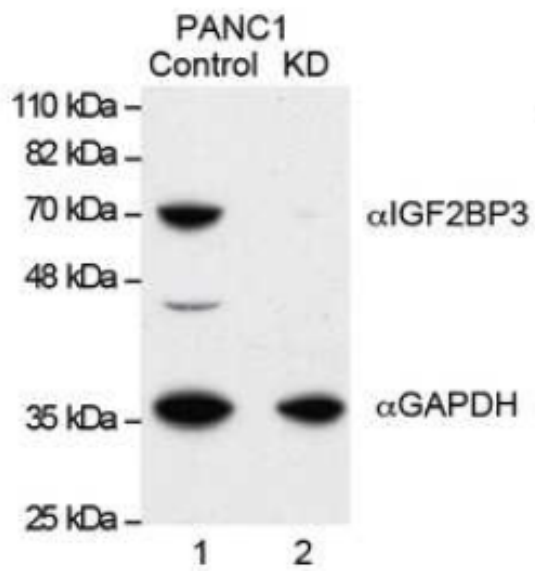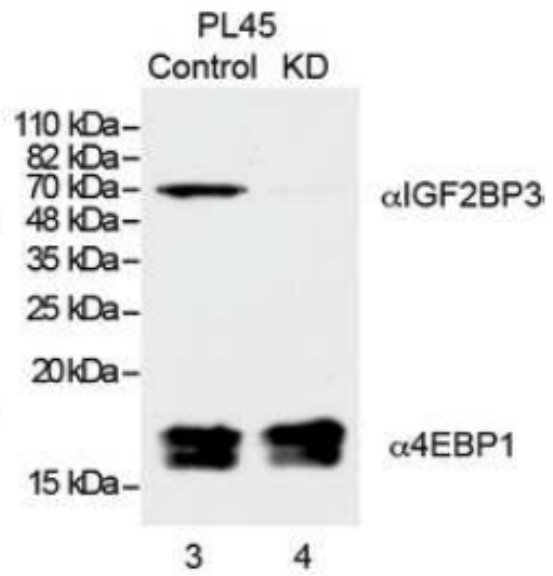mRNA Decay

ORF

ORF

IGF2BP3

RISC

RISC

IGF2BP3

IGF2BP3 RNA Regulon

Nucleus

mRNA

Cytoplasm

IGF2BP3

Adhesion

Migration

Proliferation/
Survival

115
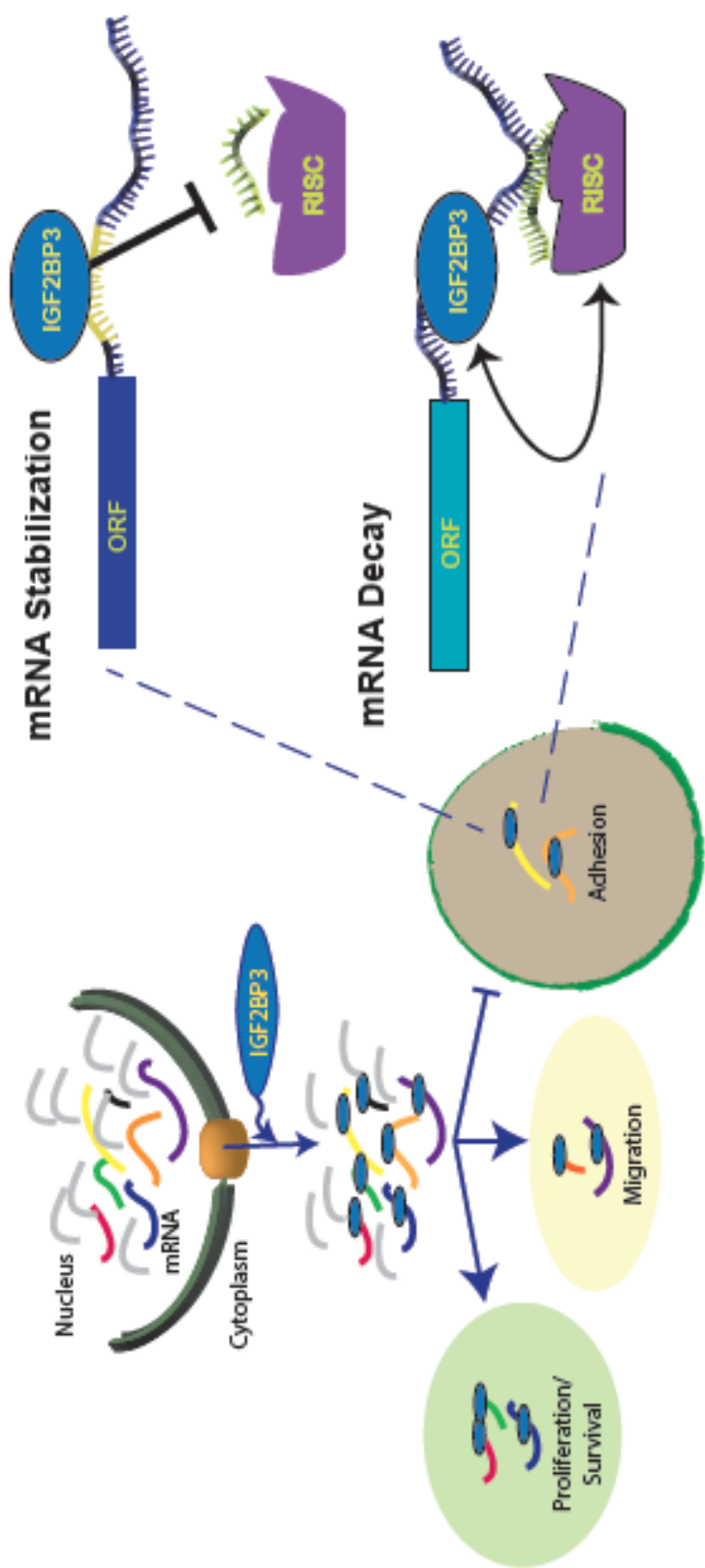
Table 1. Gene Ontology Terms enriched in the IGF2BP3-RNA targets from PDAC cells

Term	GOA	Adj. P-value	Genes
positive regulation of cell migration	GO:0030335	0.0030013474	TFAP2A, CD74, ONECUT2, IRS1, ANXA3, SEMA3A, ITGA2, LPAR1, F3, HIF1A, EGFR, IGF1R, TGFBR2, AMOT, LGALS3, DAB2, RPS6KB1, APC, PODXL, GCNT2, AMOTL1, MET
negative regulation of viral genome replication	GO:0045071	0.006343608	APOBEC3C, IFITM3, IFITM1, IFITM2, APOBEC3F, OAS3, HMGA2, SRPK1
actin cytoskeleton organization	GO:0030036	0.021901813	TMOD1, MEF2A, MEF2C, LIMCH1, TPM1, ARPC5, RND3, AMOT, EHD2, FGD6, EPB41L1, CFL2, ARHGDIB, NF1, CDC42EP2, MYH9, MYH10, ARF6
gland development	GO:0048732	0.030456581	ONECUT2, IRS1, ITGA2, HMGA2, AK4, AHR, PBX1, EGFR, IGF1R, HOXB9, APC, CDH1, NF1, GNPNAT1, MET, ARF6
response to interferon-beta	GO:0035456	0.034390884	IFITM3, IFITM1, IFITM2, XAF1, TLR3
actin filament-based process	GO:0030029	0.034390884	TMOD1, MEF2A, MEF2C, LIMCH1, TPM1, ARPC5, WASL, RND3, AMOT, EHD2, FGD6, EPB41L1, CFL2, ARHGDIB, NF1, CDC42EP2, MYH9, MYH10, ARF6
establishment of cell polarity	GO:0030010	0.057952764	FERMT1, PRKCI, MYH9, AMOTL1, IGF1R, AMOT, ARF6
phosphatidylinositol 3-kinase signaling	GO:0014065	0.078515264	PLEKHA1, ERBB3, IRS1, NF1, IGF1R
response to metal ion	GO:0010038	0.082749545	MEF2A, TFAP2A, MEF2C, NEDD4L, BRAF, ABAT, IQGAP1, EGFR, BMP6, PPP3CA, S100A16, CDH1, NEDD4, ID2, SLC25A23, B2M

Table 2. KEGG Pathways enriched in IGF2BP3 transcript targets in PDAC Cells

KEGG Pathway	P-value	Adjusted P-value	Genes
adherens junction	0.0008	0.0673	SMAD2;CDH1;WASL;IQGAP1;PVRL3;MET;EGFR;IGF1R;TGFBR2
colorectal cancer	0.0016	0.0673	SMAD2;APC;CYCS;BRAF;MET;SOS2;EGFR;IGF1R;TGFBR2
regulation of actin cytoskeleton	0.0017	0.0673	ITGA2;MSN;BRAF;ARPC5;WASL;IQGAP1;GNG12;SSH3;EGFR;APC;CFL2;ITGB8;MYH9;MYH10;SOS2

Supplemental Figure 1. IGF2BP3 expression correlates with increased cell invasion. (A and B) Westernblot analysis of IGF2BP3 level following transfection of short hairpin against IGF2BP3 in PDAC cell lines PANC1 and PL45 (lane 2 and 4) or with a non-trageting shRNA (lane 1 and 3). (B) Bar graph quantifying the fraction of invasive PANC1 and PL45 cells through matrigel filters in the presence (Control) or absence of IGF2BP3 (KD). The bars represent an average of three independent experiments and positive SDs are indicated. *** (Unpaired T.test Pvalue<0.01), * (Unpaired T.test≤0.01). (C) Representative micrograph of DAPI stained cells that migrated through the matrigel filter after 72 hrs.

A.

PANC1
Control   KD

110 kDa –
82 kDa –
70 kDa –   αIGF2BP3
48 kDa –
35 kDa –   αGAPDH
25 kDa –

1   2

B.

PL45
Control   KD

110 kDa –
82 kDa –
70 kDa –   αIGF2BP3
48 kDa –
35 kDa –
25 kDa –
20 kDa –
15 kDa –   α4EBP1

3   4

C.

Ctrl filter   Matrigel filter

shCtrl
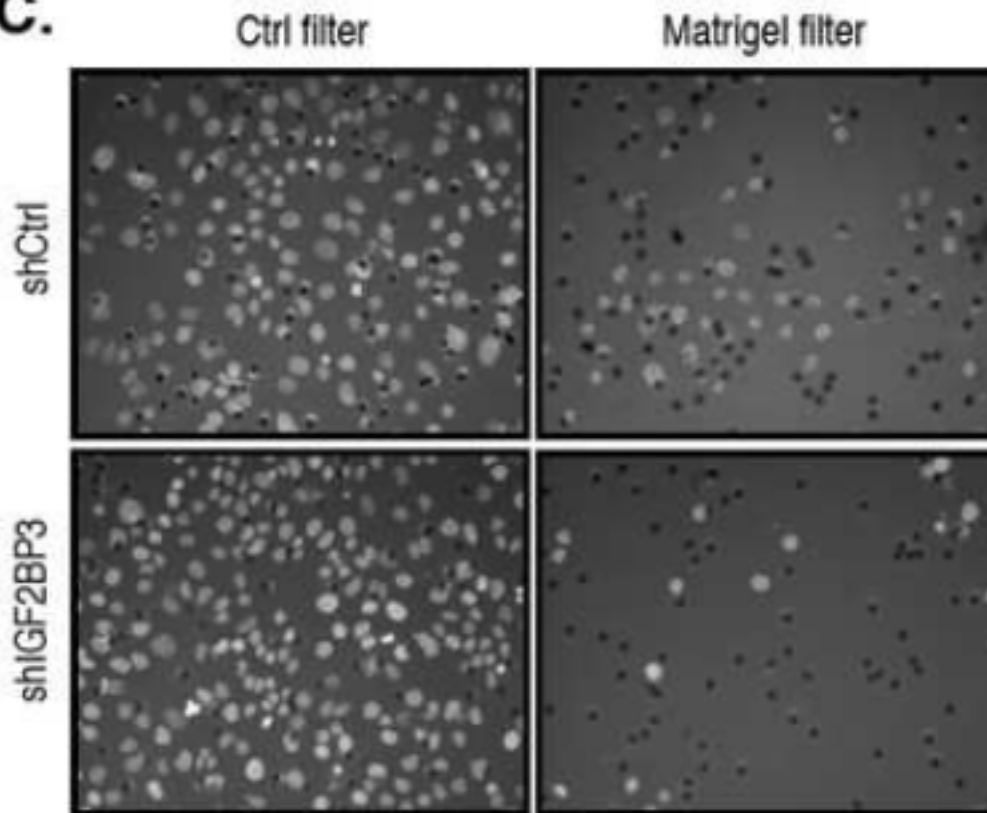
shIGF2BP3

Supplementary Figure 2: Enriched functions of proteins encoded by differentially expressed mRNAs in PANC1 cells. (A) Volcano plot showing differentially expressed genes in control or IGF2BP3-depleted PANC1 cells. Genes that are differentially expressed by ≥4 fold with an adjusted P value <0.1 are coloured orange, nonsignificant genes are in grey. (B) Gene Ontology terms and KEGG Pathways enriched indifferentially expressed genes identified in IGF2BP3-depleted PANC1 cells.

A.

padj < 0.1 and Fold Change > ⟨ ● No ● Yes

B.

Enriched functions of proteins encoded by differentially expressed mRNAs in PANC1 cells

- Biological Process
- Cellular Compartment
- Molecular Function
- KEGG Pathway
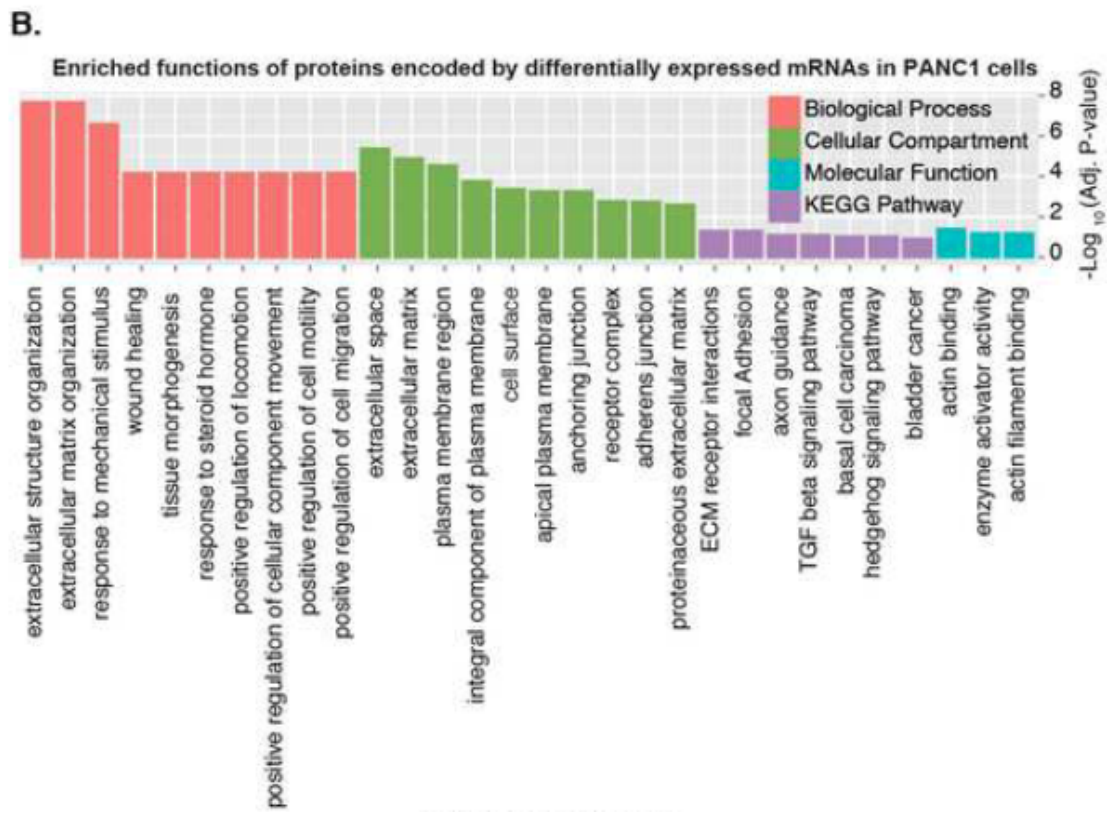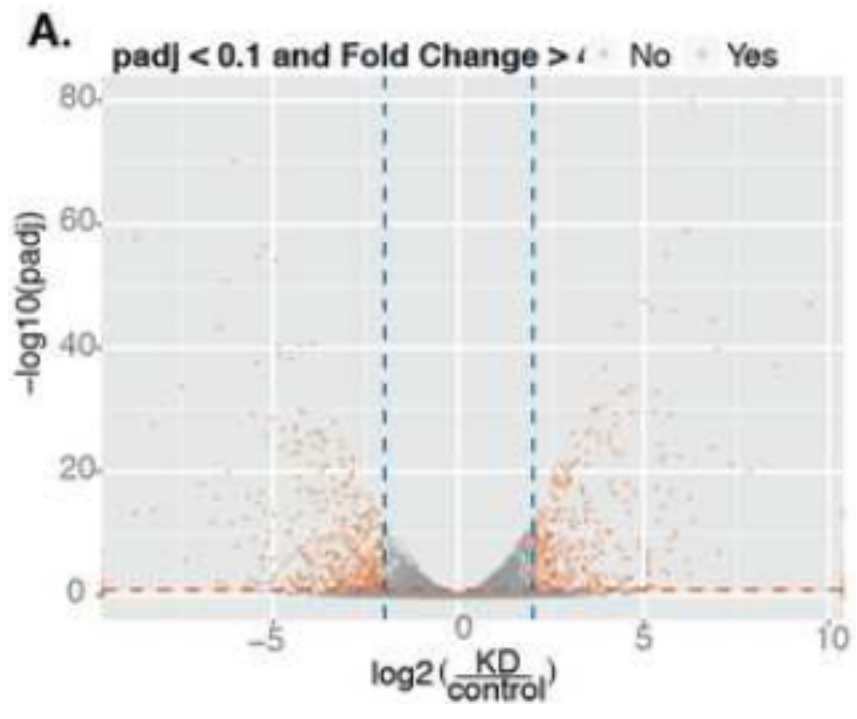
Supplementary Figure 3. Correlation of replicate IGF2BP3 RIP-seq experiments. Read counts for mRNAs identified in IGF2BP3 (IMP3) RIP experiments or control immunoprecipitations with IGG.
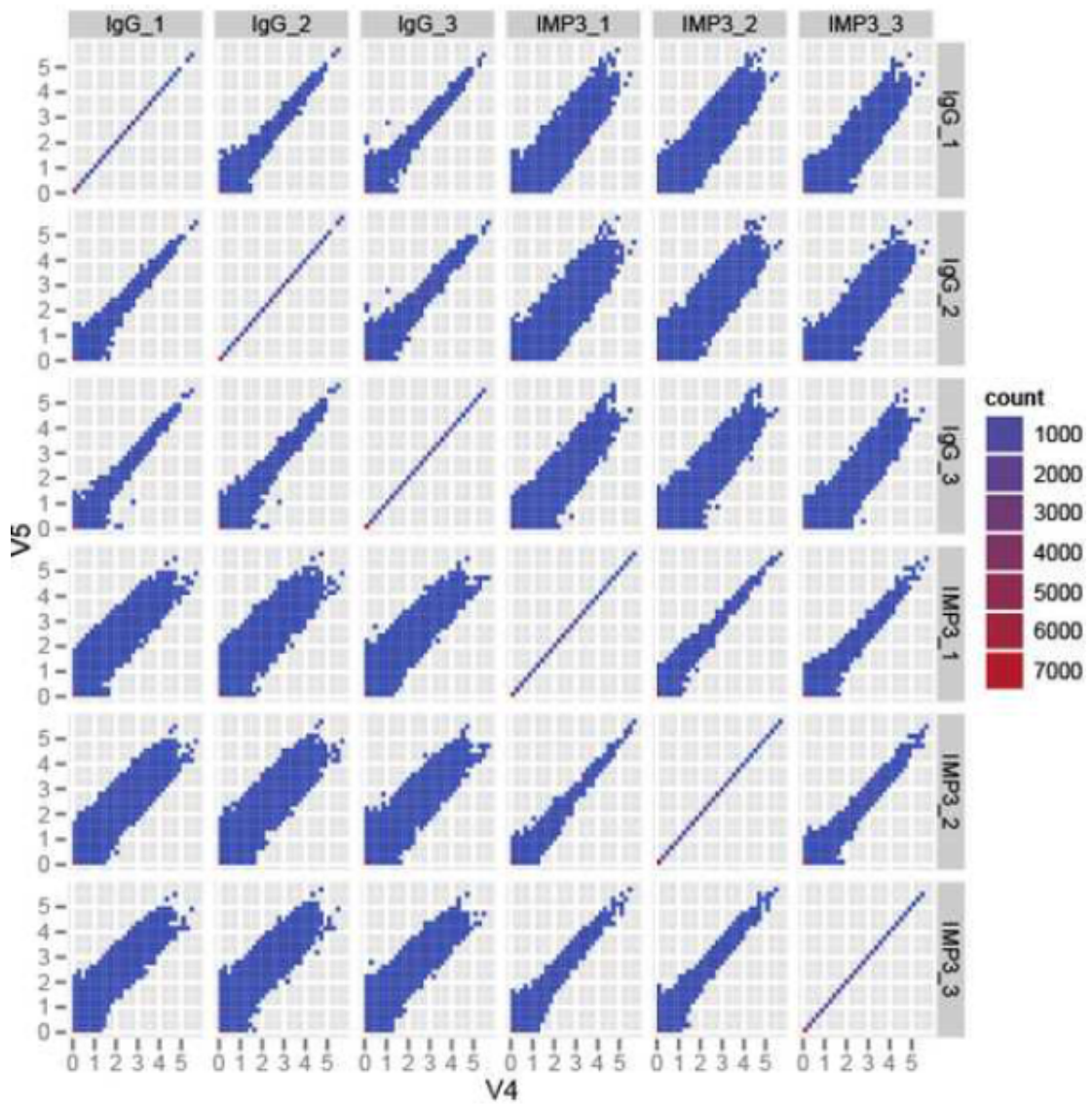
Supplementary Figure 4. Global analysis of IGF2BP3-RNA interactions by RIP and iCLIP-seq. (A) Volcano plot of transcripts enriched in anti-IGF2BP3-RIP versus control IgG. (B) Same plot as (A) with transcripts identified by iCLIP shaded in dark blue or dark green. (C) KEGG pathways enriched in the set of 166 cross-validated mRNA targets. (D,E) Genome browser snap shots of two cross-validated IGF2BP3 mRNA targets.
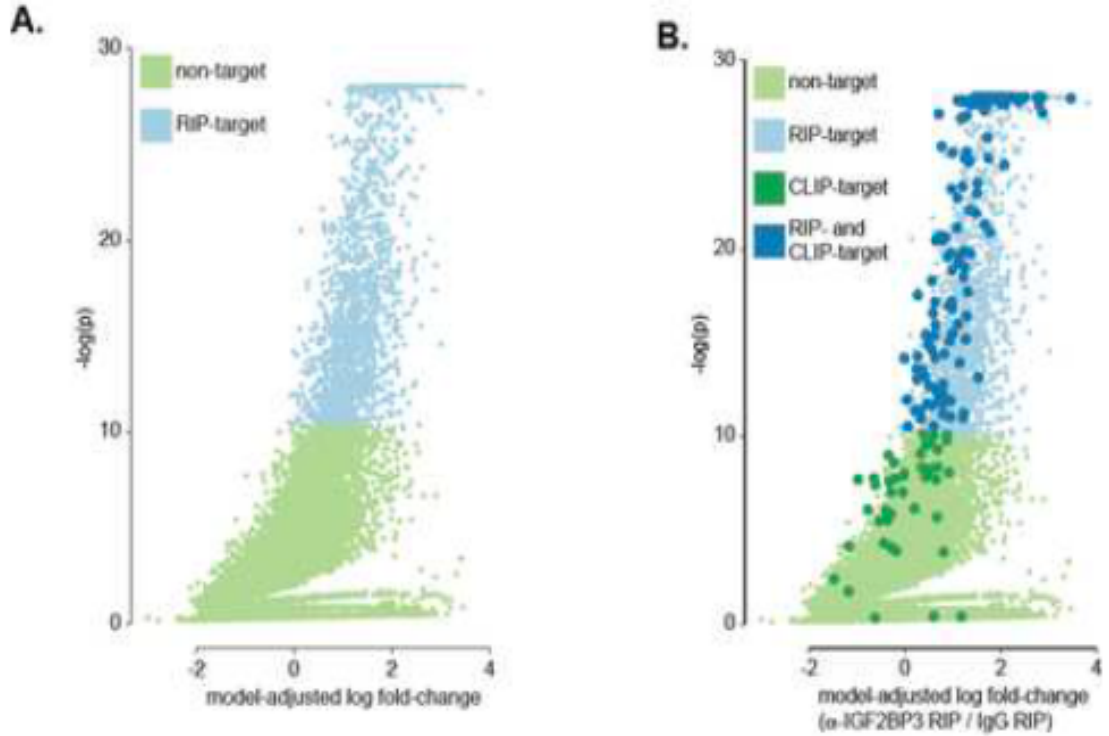
C. Pathways enriched in cross-validated IGF2BP3 mRNA targets
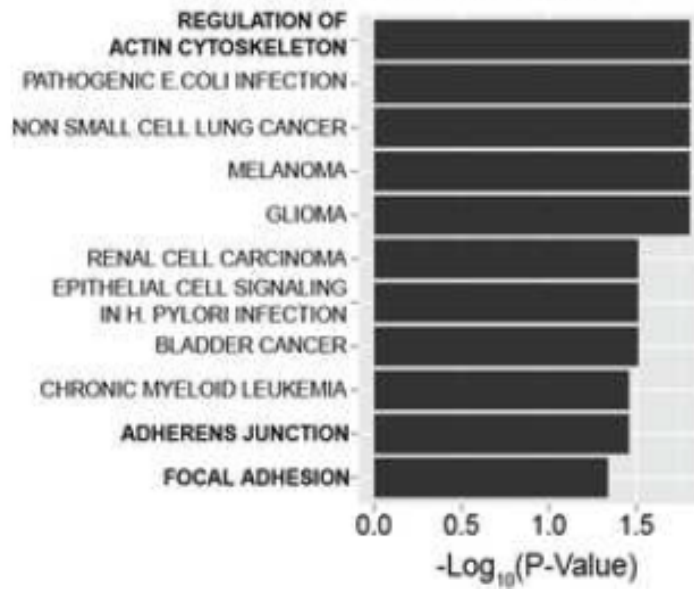
Supplementary Figure 5. Purification of IGF2BP3-RNA complexes by crosslinking immunoprecipitation. (A) Autoradiograph of IGF2BP3-RNA ^{32}P-labeled complexes after 15 U and 0.5 U microconuclease digestion (lane 3 and 4) in PL45 cells treated with UV. Black bar, region of the gel excised for library generation. (B) Western blot confirming immunoprecipitation of endogenous IGF2BP3 protein (lane3).

A.

	Input	SN	Pellet
	15U	15U	15U
	+	+	+
	-	+	+
	-	-	-
	1	2	3

B.

MNase:	15U	15U	15U	0.5U
Extract:	+	+	+	+
Ab:	-	+	+	+
UV:	+	-	+	+
	1	2	3	4

C.

RS411

	Input	IP

IGF2BP3
HC
*

IP: α-IMP3 (MBL)
Blot: α-IMP3 (Bethyl)

D. RS411 B-Cell IGF2BP3 CLIP

[MNase]:	+	+	+++	+
Ab:	-	+	+	+
UV:	+	-	+	+

IGF2BP3-
RNA
Complex

| 1 | 2 | 3 | 4 |

E.

REH

	Input	IP

GF2BP3
HC

IP: α-IMP3 (MBL)
Blot: α-IMP3 (Bethyl)

F. REH B-Cell IGF2BP3 CLIP

[MNase]:	+	+	+++	+
Ab:	-	+	+	+
UV:	+	-	+	+

IGF2BP3-
RNA
Complex

| 1 | 2 | 3 | 4 |

127
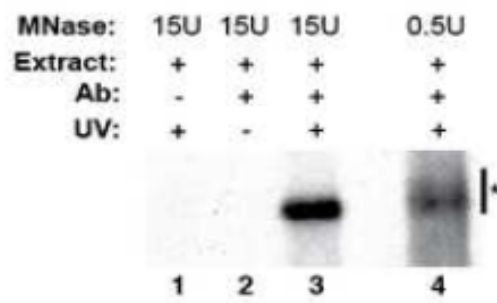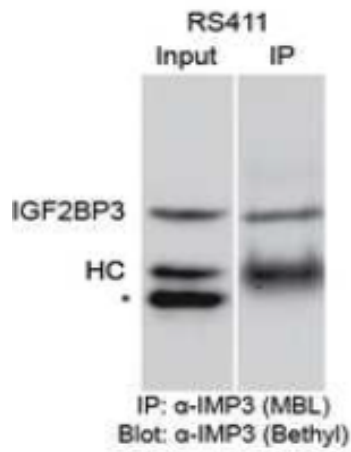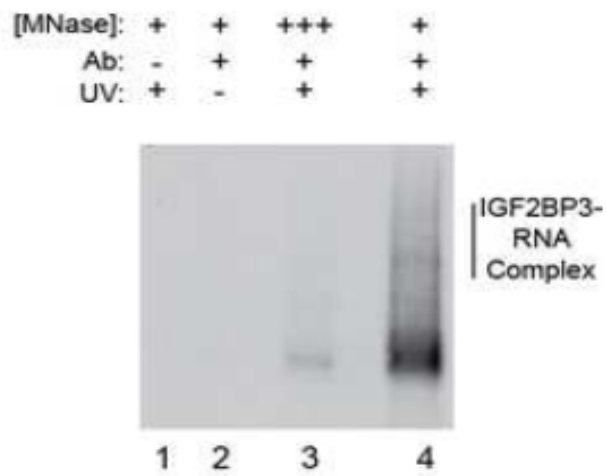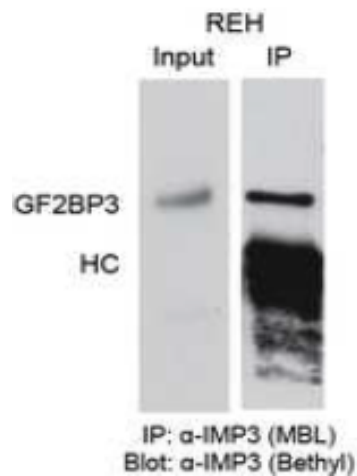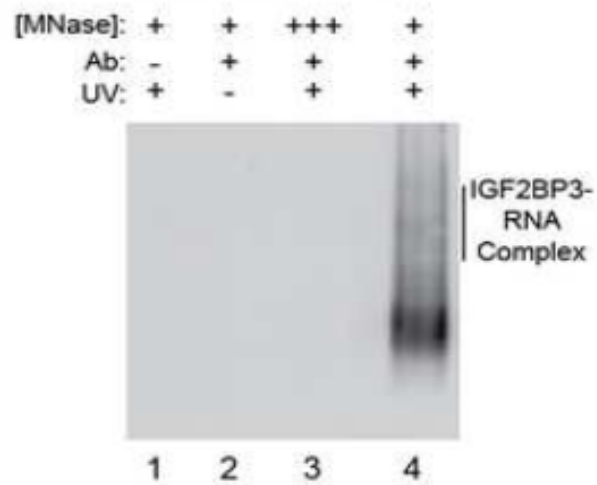
Supplementary Figure 6: mRNA decay assays for IGF2BP3 target transcripts in the PDAC cell line PL45. A. Left panel, Genome Browser snapshots of read coverage from replicate RIP seq experiments for CD44 3'UTR. Right panel, relative standard curve RT-qPCR analysis of CD44 mRNA in control of IGF2BP3-depleted PL45 cells. B. As in (A) except screen shots and stability assays focus on the CLDN1 mRNA.
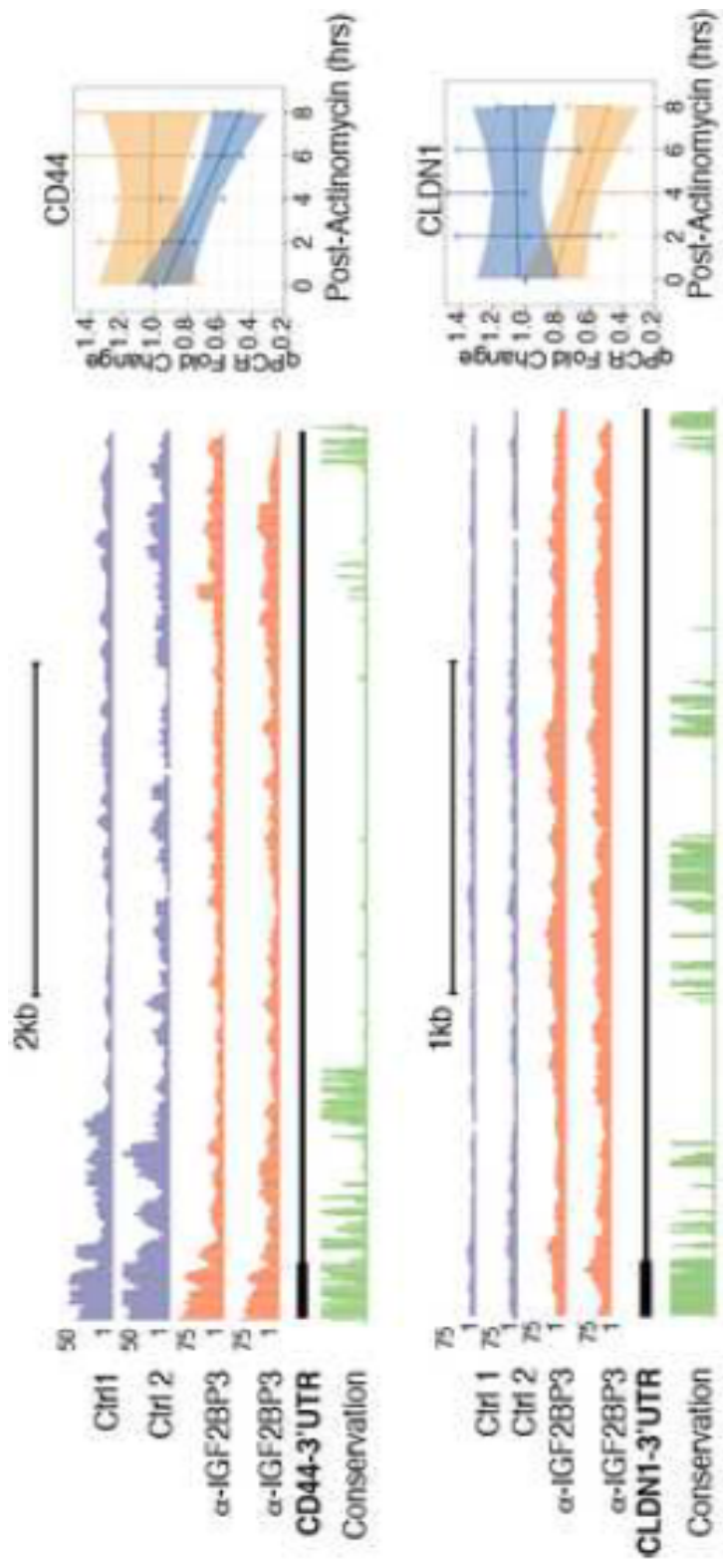
Supplemental Figure 7: Aggregated Cross-validation of IGF2BP3 iCLIP targets with IGF2BP3-sensitive differentially-expressed genes in PL45 and Panc1 PDAC cells. (A) Volcano plot of differentially-expressed genes (blue dots) determined using DESeq analysis (Anders and Huber. Genome Biol. 2010) on RNA-seq samples from control and constitutive knock down for IGF2BP3 in PL45 and Panc1 cells. Those differentially-expressed genes identified as IGF2BP3 targets by iCLIP have been highlighted (orange dots). Dots demarcated by black outlines are those genes determined to be associated with oncogenesis, cell invasion and motility, as well as focal adhesion/adherens junctions by gene ontology analysis. Dotted lines represent cutoffs of -/+ 1.5 fold-change in expression (vertical red lines) and p-value < 0.05 cutoff (horizontal black line). (B) qPCR analysis of selected genes showing differential expression and that are targeted by IGF2BP3 in iCLIP experiments. mRNA levels are normalized to control gene TBP. (C,D) Gene Ontology analysis of gene subgroup showing differences in expression IGF2BP3 mRNA targets using Enrichr gene list enrichment analysis webtool (Chen et al, BMC Bioinformatics. 2013). Term lists used in this analysis were GO Biological_Processes (C) and KEGG (Kyoto Encyclopedia of Genes and Genomes) (D) to determine enriched processes and pathways from our cross-validated list of 164 IGF2BP3-targeted and –sensitive genes. Vertical dotted lines represent p-value cutoff ($p < 0.05$).

A.

B.

C. GO Biological process terms
- positive regulation of cellular component movement
- positive regulation of cell motility
- positive regulation of cell migration
- positive regulation of locomotion
- leukocyte differentiation
- response to interferon–alpha
- regulation of cell projection organization
- cellular response to calcium ion
- negative regulation of viral genome replication
- muscle tissue development

D.
- colorectal cancer
- adherens junction
- regulation of actin cytoskeleton
- pancreatic cancer

KEGG pathways Terms

131

Chapter 5

The RNA binding protein IGF2BP3 promotes hematopoietic progenitor cell proliferation by targeting leukemogenic pathways

Abstract

Post-transcriptional control of gene expression plays important roles in defining normal and pathological cellular phenotypes. Amongst mechanisms of post-transcriptional regulation, RNA binding proteins (RBPs) have recently been shown to play important roles. However, *in vivo* roles for RBPs are not well understood. Here, we identified the RBP Insulin-like growth factor 2 mRNA binding protein 3 (IGF2BP3) to be specifically overexpressed in MLL-rearranged B-acute lymphoblastic leukemia (B-ALL), which constitutes a subtype of this malignancy associated with poor prognosis and a risk of high relapse. IGF2BP3 was required for the survival of B-ALL cell lines, and knockdown led to decreased proliferation and increased apoptosis. Enforced expression of IGF2BP3 provided murine bone marrow cells with a strong survival advantage, led to a proliferation of hematopoietic stem and progenitor cells and skewing of hematopoietic development to the B-cell/myeloid lineage. Cross-link immunoprecipitation and high-throughput sequencing, uncovered the transcriptome regulated by IGF2BP3; including novel direct targets, MYC and CDK6. These were regulated following experimental alteration of IGF2BP3 expression *in vivo*, and are regulated via elements within their 3'untranslated regions. Hence, IGF2BP3 mediated targeting of oncogenic transcripts may represent a critical pathogenetic mechanism operant in MLL-rearranged B-ALL, highlighting IGF2BP3 and its cognate RNA binding partners as potential therapeutic targets in this disease.

Introduction

Oncogenesis in early B-cell progenitors results in B cell acute lymphoblastic leukemia (B-ALL), the most prevalent hematological neoplasm in children and young adults (Mullighan, 2012). The majority of B-ALL cases exhibit genetic alterations including recurring chromosomal rearrangements, which contribute to the heterogeneity of the observed clinical behavior (Mullighan and Downing, 2009). Specifically, B-ALL with chromosomal rearrangements of the Mixed Lineage Leukemia (MLL) gene accounts for 5-6% of all B-ALL cases and is associated with poor prognosis and risk of early relapse after treatment (Krivtsov and Armstrong, 2007). *MLL*, which encodes a H3K4 methyltransferase, plays a critical role in the transcriptional dysregulation that occurs during leukemogenesis (Krivtsov and Armstrong, 2007; Krivtsov et al., 2008). Previously demonstrated targets of MLL include genes critical in cell survival and proliferation, such as BCL2, MYC, and CDK6 (Jiang et al., 2012; Placke et al., 2014; Robinson et al., 2008). Additionally, MLL is known to regulate hematopoiesis and its expression correlates with the maintenance of hematopoietic stem cell (HSC) self-renewal and differentiation (Ernst et al., 2004; Jude et al., 2007). In line with such a role in normal HSC function, MLL fusion proteins induce *HOXA9* and *MEIS1*, generating leukemia that displays "stem cell-like" properties (Imamura et al., 2002; Rozovskaia et al., 2001; Somervaille et al., 2009). These findings demonstrate an intimate connection between the

134

dysregulation of gene expression and malignant transformation and highlight the importance of investigating key players in the regulation of gene expression.

Simplistically, gene expression may be regulated at the transcriptional and post-transcriptional level. Recent work has revealed the complexity of the latter mechanism, which includes not only sequences intrinsic to the regulated mRNA but also other factors such as microRNAs, RNA binding proteins (RBPs) and non-coding RNA (Palanichamy and Rao, 2014). However, the role of gene expression regulation by RBPs in the malignant transformation of B-cells is not understood. In an effort to identify critical RBP-mediated regulation in B-ALL, we began by examining a high throughput dataset generated in our laboratory, identifying the Insulin-like growth factor 2 mRNA-binding protein 3 (IGF2BP3) as one of the top dysregulated genes in MLL-translocated B-ALL. IGF2BP3 belongs to a family of mRNA-binding proteins which consists of three structurally and functionally related paralogs (IGF2BP1, IGF2BP2, and IGF2BP3) that influence the cytoplasmic fate of mRNAs through localization, stability, and translation (Bell et al., 2013; Lederer et al., 2014). IGF2BP3 is an oncofetal protein with high expression during embryogenesis, low expression in adult tissues, and re-expression in malignant tissues. In epithelial cancer, IGF2BP3 expression is associated with a range of neoplastic phenotypes (Fadare et al., 2013; Kobel et al., 2009; Schaeffer et al., 2010; Suvasini et al., 2011). However, many of these studies have been largely correlative and a bona fide functional role of IGF2BP3, or any RBP, in B-cell oncogenesis has not been

established and *in vivo* evidence for a role in oncogenesis and/or development remains limited.

In this study, we sought to delineate the function of IGF2BP3 in B cell leukemogenesis. Providing the first *in vivo* functional role for IGF2BP3, we overexpressed IGF2BP3 in the bone marrow of lethally irradiated mice and found that it plays a critical role in the maintenance of hematopoietic stem and progenitor cells— recapitulating some features of MLL-rearranged B-ALL. We also found IGF2BP3 was essential for the survival of B-ALL cell lines. We used individual nucleotide resolution crosslinking immunoprecipitation (iCLIP) to capture the in situ specificity of protein-RNA interactions and to reveal the positional context of protein-binding sites across the transcriptome. In total, we identified IGF2BP3 binding sites in nearly 1,000 transcripts in two B-ALL cell lines. IGF2BP3 crosslinking sites are strongly enriched in the 3'UTRs of target transcripts. Of the many IGF2BP3 target transcripts, we demonstrated IGF2BP3-mediated enhancement of the expression of oncogenic targets CDK6 and MYC in B-ALL cells and hematopoietic progenitor cells *in vivo*. Together, our studies suggest that IGF2BP3 mediated upregulation of a variety of oncogenic targets may represent a key pathogenetic mechanism operant in MLL-rearranged B-ALL.

Results

IGF2BP3 is differentially expressed in MLL rearranged B-ALL

To begin our study of RBPs in B-ALL, we initially analyzed a microarray to identify RNA binding proteins that are differentially expressed in MLL rearranged B-ALL, using an annotated list based on empirical evidence of RNA binding as well as predicted RNA binding function. In our initial sample set, we hybridized 20 cases of B-ALL with three different common translocations (TEL-AML1, E2A-PBX1 and various translocations involving MLL, but mainly MLL-AF4), followed by a confirmatory set of 24 independent cases (for a total n=44). Following correction for multiple hypotheses testing, we performed unsupervised hierarchical clustering with significantly differentially expressed protein-coding genes (adj. p-value ≤0.01). This generated a list of RNA binding proteins differentially expressed between the three subtypes of B-ALL. In the list of RNA binding proteins whose expression was highest in MLL rearranged leukemia, IGF2BP3 was among the top candidates (Figure 1A). MLL rearranged leukemias show a "stem-cell" signature with high expression of stemness associated genes like HOXA9, MEIS1 and CD44 (Imamura et al., 2002; Tsutsumi et al., 2003). Concordant with this, we observed that HOXA9, MEIS1A, CDK6 and MYC, putative targets of the oncogenic MLL fusion protein, were significantly over expressed in the MLL group when compared to the other two subsets (Figure S1A-D). By performing qPCR on a large cohort of B-ALL patient-

derived bone marrows, we confirmed that IGF2BP3 and CD44 were highly expressed in the MLL group (total n=134) (Figure 1B-C). Additionally, IGF2BP3 expression was significantly higher in all B-ALL samples when compared to CD19+ B-cells isolated from healthy donors (Figure 1B). To examine the dependence of IGF2BP3 on MLL-mediated effects on gene expression, we utilized I-BET151, a BET domain inhibitor which has recently been shown to have specific anti-tumor activity via inhibition of MLL-dependent gene expression regulation (Dawson et al., 2011). Treatment of RS4;11, an MLL-AF4 expressing human B-ALL cell line, with I-BET151, caused a dose dependent decrease in the expression of MYC, CDK6 and IGF2BP3 (Figure 1D). It also caused cell cycle arrest in the G1-S phase as seen after propidium iodide staining (Figure 1E-F). These experiments confirm the over expression of IGF2BP3 in B-ALL, with the highest expression seen in MLL rearranged B-ALL. In line with IGF2BP3 being downstream of MLL fusion proteins, a fall in IGF2BP3 mRNA levels along with other MLL-AF4 targets is seen after BET inhibition.

IGF2BP3 loss-of-function causes apoptosis in B-ALL cells

Given the oncogenic expression pattern of IG2BP3 in human B-ALL, we proceeded to examine its expression in four different B-ALL cell lines, including 697 (E2A-PBX1 translocated), RS4;11, REH (TEL-AML1 translocated), and NALM6 (Figure 2A).To examine the effects of IGF2BP3 knockdown, we used a lentiviral

vector expressing two different miRNA-formatted siRNA sequences to transduce RS4;11 cells (Figure 2B). Both siRNAs caused decreased IGF2BP3 expression as seen by RT-qPCR and near-complete abrogation by Western Blot (Figure 2C and 7H). Propidium iodide staining showed an increase in the apoptotic sub-G1 fraction and MTS assay shows a significant reduction in cell proliferation with IGF2BP3 knockdown by both siRNAs compared to the empty control vector, confirming the dependence of B-ALL cell lines on IGF2BP3 for survival (Figure 2D-E). Deletion of the IGF2BP3 locus using the CRISPR-Cas9 system was also undertaken in the RS4;11 cell line. We utilized the LentiCRISPR vector and two different guide strands, Cr1 and Cr2, (Shalem et al., 2014) to target the IGF2BP3 locus for deletion. Complete abrogation of IGF2BP3 protein expression was seen on Western blot with the guide RNA Cr2, whereas residual protein was detected with Cr1 (Figure 2F). CRISPR mediated deletion was also confirmed by a T7 endonuclease assay (Figure S2D-E). Cr2 demonstrated reduced cell proliferation by MTS assay, increased sub-G1 staining and increased Annexin-V positivity (Figure 2G-I). To confirm these findings, we also targeted IGF2BP3 for knockdown in NALM6 cells using a lentiviral siRNA expression system (Figure S2A). Reduced IGF2BP3 mRNA levels were observed with both siRNAs, with si2 giving a stronger reduction in cell proliferation as seen by the MTS assay (Figure S2B-C). Together, these findings highlight the importance of IGF2BP3 in maintaining cell survival and proliferation in B-ALL.

Enforced expression of IGF2BP3 leads to high levels of engraftment and increased leukocytes

To directly assess the role of IGF2BP3 in the hematopoietic system, we undertook an *in vivo* experiment to examine the effects of its enforced expression. We initially cloned the human or mouse coding sequence of IGF2BP3 into an MSCV-based retroviral vector (Figure 3A), and confirmed the functionality of the vector in expressing both IGF2BP3 and the GFP marker. Both the human and mouse coding sequence vectors demonstrated expression of GFP and IGF2BP3 at both the RNA and protein level (Figure 3B-3D and data not shown). A peripheral bleed of these mice at 12 weeks showed significantly increased engraftment in mice with enforced expression of human and mouse IGF2BP3, as measured by the congenic CD45.2 versus CD45.1 FACS markers (Figure 3D-E). Moreover, significantly increased GFP+ leukocyte cells were found, confirming increased engraftment attributable to IGF2BP3 expression (Figure 3G). Additionally, B-cell and myeloid cell counts in the peripheral blood were also increased following complete engraftment (Figure 3H-I). There was no difference in the number of T-cells in the periphery (Figure 3J). Interestingly, the number of platelets and red blood cells were significantly lower with IGF2BP3 enforced expression (Figure 3K-L). Together these findings suggest that IGF2BP3 promotes hematopoietic stem cell engraftment, and skews bone marrow development towards the B/myeloid lineage, and away from T-cells, erythroid cells, and megakaryocytes. These findings, notably the preferential

increase in B and myeloid cells, are interesting in light of the fact that MLL rearrangements are found not only in B-ALL but also in mixed lineage acute leukemia, which most commonly express both B-cell and myeloid markers.

Enforced expression of IGF2BP3 leads to increased progenitors in the bone marrow with higher rates of proliferation

To further characterize these hematopoietic changes, IGF2BP3 overexpressing mice were sacrificed and the hematopoietic organs were collected for analysis at 6 months post-transplant. The percentage of GFP+ cells was significantly higher in the IGF2BP3 overexpressing bone marrow, once again suggesting that IGF2BP3 expressing HSCs show better, sustained engraftment (Figure S3A-B). The overall proportion of myeloid and B-cells in the bone marrow were similar between control and IGF2BP3 expressing mice (Figure S3C-D). RT-qPCR from the RNA collected from the mouse bone marrow confirmed human and mouse IGF2BP3 overexpression, respectively (Figure S3E-F). These changes led us to query whether there were changes in hematopoietic progenitor subsets in the bone marrow. Indeed, enforced expression of IGF2BP3 led to an increase in the fraction of hematopoietic stem cells (HSCs), lymphoid-primed multipotent progenitors (LMPPs) as well as common lymphoid progenitors (CLPs) (Figure 4A-C). We followed the developmental pathway of B-cells by subsetting the various B-cell progenitors in the bone marrow following the schema created by Hardy et al. (Hardy and Shinton, 2004). Among the

Hardy fractions, we observed a significant increase in the number of cells in fractions A and B with no significant differences observed in developmentally subsequent stages (Figure S3H-I, K). Hence, over expression of IGF2BP3 led to an increase in immature hematopoietic fractions starting at the level of the HSC and on to the pro-B-cell stage. To analyze the proliferation rate of the various progenitor cells in the bone marrow, we performed intracellular staining with Ki67 in conjunction with staining for the various hematopoietic progenitor populations. Flow cytometry revealed that the Lin-Sca1+c-Kit+ (LSK) population and the LMPPs in IGF2BP3 overexpressing BM had significantly higher amounts of Ki67 (Figure 4D-E). The CLPs did not show a significant difference in Ki67 expression. (Figure S3G, J). These findings imply an increase in the proliferation rate of the early progenitors (HSCs and LMPPs), secondary to increased IGF2BP3 expression. Presumably, this leads to an increase in their numbers and differentiation into more committed downstream progenitors (CLPs and the earlier Hardy fractions, which are also elevated in number). Hence, the enforced expression of IGF2BP3 causes a preferential increase in numbers and proliferation of early progenitor populations, leading to the observed B- and myeloid biased leukocytosis seen in the periphery.

Enforced expression of IGF2BP3 increases the number of B-cells in the thymus and myeloid cells in the spleen

The normal mouse thymus is composed mostly of T-cell progenitors, but in many murine models of leukemia and lymphoma, it becomes enlarged and overrun by malignant leukocytes (Aifantis et al., 2008). Here, we observed a significantly higher percentage of GFP+B220+ B-cells in the thymus when human or mouse IGF2BP3 was overexpressed, with the effect being more pronounced with mouse IGF2BP3 (Figure 5D). Mice with enforced expression of mouse IGF2BP3 also showed a significant decrease in the number of CD3ε+ T-cells. There was no significant difference in the level of GFP + cells in these thymi, indicating a lineage specific expansion of B-cells (Figure 5A-D). 4/8 of the thymi overexpressing human IGF2BP3 and 1/8 of the thymi overexpressing mouse IGF2BP3 weighed over 50 mg. None of the MIG expressing thymi weighed over 50 mg (Data not shown).

On microscopic examination, IGF2BP3 caused thymic medullary expansion with infiltration by large cells. One of the thymi expressing human IGF2BP3 had complete ablation of the cortico-medullary junction (Figure 5E). These results imply that IGF2BP3 expression might lead to disruption of the thymic architecture and can serve as a precursor to malignant transformation. Interestingly, the spleens were also enlarged following enforced expression of IGF2BP3. Differences in splenic weight were statistically significant for human IGF2BP3 overexpressing group with a trend

noted for the mouse IGF2BP3 group (Figure S4A). IGF2BP3 led to an increase in the number of myeloid cells in the spleen with a significant decrease in the number of CD3ε+ T-cells (Figure S4B-F). Overall, IGF2BP3 appears to tilt the hematopoietic stem cell developmental program towards the B-cell or the myeloid lineage which is evident in the peripheral blood, spleen and thymus. Hence, the changes seen in the bone marrow- increased numbers and proliferation of B-lymphoid and myeloid progenitors- may result in alterations in hematopoietic homeostasis in the periphery.

iCLIP identifies the IGF2BP3-RNA interactome in leukemia-ALL cells

The molecular basis of the action of RBPs has recently been investigated using iCLIP and high throughput sequencing. To gain insight into the role of IGF2BP3 in cell growth and MLL-driven leukemogenesis, we performed an iCLIP assay with this protein. iCLIP exploits the photo-reactivity of nucleic acid and protein residues and nuclease fragmentation of protein bound transcripts to capture protein-RNA interactions occurring in situ. Antibodies against IGF2BP3 were used to immunoprecipitate protein-RNA complexes from control or UV-irradiated RS4;11 and REH cells. The iCLIP experiment was performed in triplicate from the RS4;11 and REH cell lines and a representative autoradiograph is shown in Figure S5A and C. As expected, the immunoprecipitated material is antibody-dependent, UV-dependent and the electrophoretic mobility of the complex is nuclease-sensitive, as predicted for a protein-RNA complex (Figure S5B and D). Co-precipitated RNA was

isolated from nitrocellulose filters, converted to cDNA libraries (Figure S5E) and subject to high-throughput sequencing. After accounting for PCR duplications, we obtained ~1 million reads per replicate of which >70% mapped uniquely to the human genome (Table S1). iCLIP replicates from both RS4;11 and REH cells are strongly correlated, demonstrating the high degree of precision in our approach (Figure S6D-E). The 5' ends of iCLIP sequences correspond to the protein-RNA crosslinking sites. We compared the distributions of crosslinking sites obtained from IGF2BP3 (REH and RS4;11 cells) and hnRNPA1 (HEK cells) to simulated data drawn randomly from the human genome. Compared to hnRNPA1 and the genome, IGF2BP3 crosslink sites are enriched in exons (Figure 6A). We identified peaks using a negative binomial model (see Methods) in each biological replicate from RS4;11 and REH cells. In total, 849 peaks in 669 genes and 1,937 peaks in 1,149 genes were identified in REH and RS4;11 iCLIP experiments, respectively. Of the peaks called within mRNA sequences, the majority is located within the 3'UTRs (Figure 6B: last exon and Figure S6A-C). Given the apparent bias of IGF2BP3 binding sites in 3'UTR of target transcripts, we investigated the positional bias of crosslinking sites at a single nucleotide resolution relative to mRNA stop codons. The crosslink density of IGF2BP3 differed from hnRNPA1 in the 3'UTR, reaching its apex just downstream of the stop codon in both RS4;11 and REH cells. These data suggested that IGF2BP3 binding sites within the 3'UTR specifically target sequences close to the stop codon in both leukemia cell lines (Figure 6C). A search for sequence specificity in cross-linked regions revealed an 8-16 fold enrichment over background

of GCAC tetramer containing motifs in both the REH and RS4;11 datasets (Figure 6D). The ENRICHR tool (Chen et al., 2013) was used to functionally classify IGF2BP3 mRNA targets in REH and RS4;11 cell. In both cell lines, we found that target transcripts were enriched for KEGG pathways related to ribosome biogenesis and translation (Figure 6E and Table S2). By contrast, transcripts classified by ENRICHR as involved in pathogenic E. coli infection and perhaps most importantly, chronic myeloid leukemia (CML) were enriched in RS4;11 but not the REH dataset (Table S3). Amongst the genes in the CML data set, CDK6 and MYC, were also highly overexpressed in MLL-translocated cases of B-ALL (Figure 6E and S1A-B). These data demonstrate, for the first time, a comprehensive IGF2BP3 RNA interaction site atlas from human leukemia cells. This extensive interaction map reveals a strong preference of IGF2BP3 binding non-uniformly to 3'UTRs with pronounced preference for a GCAC-rich consensus motif.

In order to determine if IGF2BP3 iCLIP targets are regulated by IGF2BP3 expression levels, we performed RNA-seq on control and IGF2BP3-depleted RS4;11 cells. Cross-validation of genes differentially expressed by at least 1.5-fold with RS4;11-specific IGF2BP3 iCLIP targets found 269 common genes. Of these targets, the majority showed decrease expression of IGF2BP3 iCLIP targets with IGF2BP3 depletion (154 decrease vs. 59 increase; Figure S8A). OMIM disease gene ontology (GO) analysis of those common genes reveals a cohort of genes found associated with leukemogenesis (black circles in supplementary Figure S8A), including CDK6 and MYC. Biological processes GO analysis for IGF2BP3 iCLIP targets (Figure S8B,C

top bar plots), showing decrease expression revealed genes associated with post-transcriptional control, hematopoietic cell differentiation, and chromatin modification. KEGG GO analysis (Figure S8B,C bottom bar plots) also suggested these targets are involved in cell cycle and a variety of cancer pathways. Identical analyses on IGF2BP3 iCLIP targets showing increased expression with IGF2BP3 depletion revealed genes primarily associated with translation and protein localization.

CDK6 and MYC mRNAs are targets of IGF2BP3

Among the many novel IGF2BP3 mRNA targets, CDK6 and MYC stand out as they are quintessential oncogenes and are very important in the pathogenesis of MLL-translocated B-ALL (Jiang et al., 2012; Placke et al., 2014). Inititally, we validated these two targets in RS4;11 cells via RNA immunoprecipitation, showing enrichment in the IGF2BP3 immunoprecipitation over the IgG control (Figure S9). To test the hypothesis that IGF2BP3 post-transcriptionally regulates the expression of CDK6 and MYC via binding sites in their 3'UTR, we generated a series of vectors using a dual-luciferase reporter system (Figure 7A). The CDK6 3'UTR is 10kb long and was cloned in five separate pieces (CDK6-1 to CDK6-5), while the MYC 3'UTR is ~300 bp long. Cotransfection of the luciferase vectors with MIG or the IGF2BP3 over expressing vector showed an increase in the ratio of luciferase activity in CDK6-3 to CDK6-5. Similarly there was also a significant increase in the luciferase activity

with the MYC UTR after hIGF2BP3 overexpression (Figure 7B). This, along with the iCLIP data, confirms that IGF2BP3 binding to the UTRs of these genes may stabilize the mRNA, enhance translation, or both. Given these results, IGF2BP3 binding would be predicted to lead to increased levels of CDK6 and MYC in cells overexpressing IGF2BP3 and decreased levels in knockdown cells. To elucidate whether IGF2BP3 targets CDK6 and MYC *in vivo*, we performed intracellular staining in bone marrow cells from mice with enforced IGF2BP3 expression and analyzed them by flow cytometry. Bone marrow GFP+ cells derived from human IGF2BP3-expressing mice had increased CDK6 and MYC, as measured by Mean Fluorescence intensity (MFI) (Figure 7C-D). Enforced expression of murine IGF2BP3 led to a trend towards the same, but did not reach statistical significance. On the other hand, GFP negative cells of both groups did not show any difference in the MYC and CDK6 protein levels (Figure 7E-F). To complement these *in vivo* data, we analyzed CDK6 and MYC protein levels in cell lines where IGF2BP3 was knocked down using siRNAs previously. There was a significant decrease in CDK6 protein levels in the RS4;11 cell line after IGF2BP3 knockdown, as seen by Western blotting. Similarly, siRNA mediated knockdown in the murine pre-B 7OZ/3 cell line using a MGP-based miRNA-formatted siRNA vector demonstrated a reduction in mouse CDK6 and MYC protein levels (Figure 7G-H). At the mRNA level, IGF2BP3 overexpression led to a slight but significant increase in the Myc mRNA levels, but not in Cdk6, in bulk bone marrow (Figure S7A-B). It is possible that an increase in Cdk6 mRNA may not be detected due to the heterogeneity of the bone marrow cells

148

or because effects on mRNA stability by IGF2BP3 may be mRNA-specific. Similarly, murine 7OZ/3 cells showed no increase in the Cdk6 mRNA levels when IGF2BP3 was overexpressed (Figure S7C-E). These findings corroborate IGF2BP3 binding and subsequent translational augmentation of these target genes. Overall, these experiments demonstrate both *in vitro* as well as *in vivo* targeting of CDK6 and MYC by IGF2BP3, and indicate a molecular mechanism for the observed phenotypic effects.

Discussion

The molecular mechanism of leukemogenesis mediated by MLL-fusion proteins remained elusive until a few years ago, despite knowledge of the translocation and the resulting fusion proteins for two decades. Increasingly, it is recognized that secondary, non-genetic changes are necessary for elaboration of cancer. For example, we now know that deregulation of epigenetic marks by DOT1L, which is recruited by the MLL-AF4, is a key oncogenic mechanism (Bernt and Armstrong, 2011). In this study, we found that an RNA binding protein, IGF2BP3, is overexpressed in cases of B-ALL that carry a translocation of the MLL gene. We propose that post-transcriptional gene expression deregulation may also play an important role in leukemogenesis. This is borne out by prior studies of another RNA binding protein, Musashi-2, which demonstrated that it was highly expressed in acute myeloid leukemia, and that its overexpression could collaborate with BCR-ABL to promote myeloid leukemogenesis (Kharas et al., 2010). Our findings here extend the repertoire of deregulated RNA binding protein expression to B-ALL, providing new insights into the disease.

The mechanism of IGF2BP3 upregulation in MLL-AF4 expressing leukemia is an important question, as this would define how IGF2BP3 participates in leukemogenesis. Our studies showed that a BET domain inhibitor could downregulate IGF2BP3 in RS4;11, and this is thought to specifically target MLL-

mediated transcription at low doses (Dawson et al., 2011). Previous reports have also shown IGF2BP3 to be upregulated after MLL-AF4 overexpression in murine bone marrow cells (Krivtsov et al., 2008). Interestingly, an indirect ChIP-Seq assay done in SEM cells (which carry the MLL-AF4 translocation) showed binding of the fusion protein to the genomic locus containing IGF2BP3 (Guenther et al., 2008). Therefore, it is tempting to speculate that IGF2BP3 is a direct transcriptional target of MLL fusion proteins. However, IGF2BP3 is also overexpressed in a number of epithelial malignancies, and IGF2BP3 levels were high in most of the B-ALL cell lines that were tested. Previous immunohistochemical studies demonstrate that IGF2BP3 is also highly expressed in various mature B-cell neoplasms (Hartmann et al., 2012; King et al., 2009; Liao et al., 2005; Tang et al., 2013). Furthermore, differential regulation of this protein has been observed in B-ALL (Stoskus et al., 2011). Hence, the mechanism of its upregulation may include other oncogenic pathways.

Previously, knockdown of IGF2BP3 in multiple epithelial cell lines has been shown to reduce cell proliferation and cause apoptosis (Liao et al., 2005; Lu et al., 2011; Suvasini et al., 2011). We found that knockdown or deletion of IGF2BP3 by siRNA or by the Cas9/CRISPR system led to reduced cell proliferation and increased apoptosis in RS4;11, an MLL-AF4 expressing cell line. Similar reduction in cell proliferation was also seen in NALM6, another B-ALL cell line that shows high levels of IGF2BP3. These findings highlight the important role that post-transcriptional gene regulation can play in maintaining the malignant behavior of B-

ALL cells. It will be of great interest to study whether cell lines with low expression levels of IGF2BP3 demonstrate an altered RNA binding repertoire, with an aim towards illuminating key mRNAs that mediate downstream effects of IGF2BP3.

To study the pathogenetic function of this protein, we also created the first *in vivo* model of IGF2BP3 enforced expression in the murine hematopoietic system. We find that IGF2BP3 increases the number of HSCs, LMPPs and CLPs in the bone marrow with a concomitant increase in the proliferation rate of HSCs and LMPPs. IGF2BP3 skewed mouse hematopoiesis towards the B cell and myeloid lineage in the periphery, with leukocytosis (increased B-lymphocyte and myeloid cells) in the peripheral blood, atypical B-cell infiltration into the thymic medulla, and increased myeloid cells in the spleen. Although the mice did not develop overt leukemia by six months post-transplantation, these abnormal developmental features are similar to those seen in MLL-driven leukemogenesis. The spectrum of MLL-driven leukemia includes B-ALL, acute myeloid leukemia and acute leukemia of ambiguous lineage, particularly those with B/myeloid phenotype. Hence, in addition to the observed effects on quintessential oncogenes, the expression of IGF2BP3 may confer lineage specificity/ambiguity in leukemogenesis by regulating post-transcriptional gene expression of key lineage regulators.

Previously identified targets of IGF2BP3 include the mRNAs for IGF-2, CD44, and the transcription factor HMGA2 (Bell et al., 2013; Jonson et al., 2014).

152

Further reports on HMGA2 and IGF2BPs have suggested that they may play a role in the self-renewal potential of fetal hematopoietic stem cells (Copley et al., 2013; Toledano et al., 2012). Thus, IGF2BP3 may function in regulating both developmental and oncogenic processes. To examine the molecular mechanism behind the observed disruptions in cellular and hematopoietic homeostasis, we performed iCLIP-Seq analyses. Remarkably, we identified numerous RNA targets of IGF2BP3 in B-ALL cell lines that were known targets of MLL, including CDK6 and MYC. CDK6 has recently been implicated as a highly important target in B-ALL, and inhibition of CDK6 may for the basis of a new therapeutic intervention in B-ALL (Placke et al., 2014). MYC is a quintessential oncogene, and its overexpression plays a direct, causative role in many B-cell leukemias and lymphoma. Reporter assays confirmed that interaction of IGF2BP3 with the UTRs of MYC and CDK6 are functionally important. *In vivo* and *in vitro* targeting of CDK6 and MYC protein was confirmed in the mouse bone marrow and leukemia cell lines, in both the loss- and gain-of-function settings. Hence, we have validated iCLIP as a powerful technique for uncovering functionally relevant targets. It is important to point out that there are likely other mRNAs that interact with IGF2BP3, including those we identify here, that play important roles in cellular proliferation and/or differentiation.

The iCLIP-Seq analyses provide the first global view of the transcriptome regulated by IGF2BP3. Prior studies have demonstrated single or a few targets for this protein, but our work here shows approximately 1,000 mRNAs bound by this

protein. Enrichment for gene sets known to play a role in chronic myeloid leukemia and pathogenic E. coli infection was found. These gene sets are not specific for these processes, but are likely portions of gene expression programs that are operant in MLL-mediated leukemogenesis. Like other modes of post-transcriptional gene expression regulation, the actions of IGF2BP3 are dependent on the gene expression program mediated by transcription factors, and hence may have a role in stabilizing leukemogenic gene expression. It appears that IGF2BP3 is itself induced by the MLL-AF4 expression program, and that it binds to and upregulates several genes that are induced. In this way, IGF2BP3 appears to be reinforcing certain aspects of the gene expression program, thereby sustaining oncogenesis. Interestingly, however, enforced expression of this protein alone also led to proliferation of hematopoietic progenitor cells, suggesting that specificity of RNA stabilization and/or translational enhancement can direct development and influence lineage choice and proliferation.

As oncogenic mutations and translocations are being catalogued via various high-throughput sequencing approaches, it is also becoming apparent that single genetic abnormalities are insufficient to cause oncogenesis. Prior work has implicated non-genetic mechanisms, particularly, epigenetic regulation, as key factors in the pathway to full-blown oncogenesis. Here, our studies have uncovered a novel post-transcriptional mechanism of stabilizing oncogenic gene products such as CDK6 and MYC. This mechanism requires further study and clarification, and is likely to yield important insights into the nature of gene expression regulation in

154

leukemogenesis. With targeted therapies emerging against CDK6 and MYC, it will become critical to consider the role of post-transcriptional mechanisms in regulating oncogene-mediated gene expression programs. Moreover, novel therapeutic avenues are also suggested by the current study, including the generation of sink RNAs to block the binding of IGF2BP3 to its targets, or small molecule inhibitors of this protein designed to block its RNA binding function. Given the expression of IGF2BP3 in many different types of cancer, it will be of great interest to define whether the repertoire of bound mRNAs is similar in tumors of distinct histogenesis, and whether conserved oncogenic pathways can be targeted in hematologic and non-hematologic malignancies.

Methods and Materials

Patient Samples

The patients consisted of 134 children consecutively admitted to the Pediatric Oncologic Department at the University of Padua, Italy, from 2000 to 2008 with the diagnosis of ALL. The enrollment criteria of the study were: newly diagnosed B-ALL, age range from 0 to 18 years, and written informed consent of the parents following the AIEOP (Italian Association of Pediatric Hematology and Oncology) and the BFM (Berlin-Frankfurt-Muenster) ALL-2000 trial. The diagnosis of B-ALL was established by morphology, immunophenotyping and molecular genetics. The Philadelphia chromosome t(9;22)BCR-ABL, t(12;21)TEL-AML1, t(1;19)E2A-PBX1 and t(4;11)MLL-AF4 were detected by karyotype, FISH, or RT-qPCR. For the initial microarray studies, we utilized 20 patients (7 patients t(4;11) MLL rearranged, 6 t(12;21) TEL-AML1 translocated and 7 t (1; 19) E2A-PBX1 translocated). For further validation, an additional 24 samples with the same translocations were hybridized to microarrays. For independent validation by RT-qPCR, we utilized 90 additional samples of de novo B-ALL without selection criteria. In total, we had gene expression data from the following number of samples: 34 patients with TEL-AML1, 17 with E2A-PBX1, 17 with MLL-translocations, 3 with BCR-ABL, 63 with unknown karyotypes or no common translocations (Total n=134). Peripheral blood mononuclear cells derived from anonymized donors were obtained from the Center

for AIDS Research Virology Core Lab at UCLA, and the Flow Cytometry and Bone Marrow Lab at the Department Pathology and Laboratory Medicine.

Microarray Data Analysis

Agilent SurePrintG3 Human GE 8x60K microarrays (product # G4851A) were hybridized at the Caltech microarray core facility. These arrays target 27,958 Entrez Gene RNAs, based on RefSeq Build 36.3, Ensemble Release 52, Unigene Build 216, GenBank (April 2009), as well as 7,419 lincRNAs, based on the initial discovery set from the Broad Institute (Guttman et al., 2009). Data analysis was implemented in the R statistics package. The data from two microarray experiments (20 in the initial set and 24 in the validation set) were analyzed independently but following the same protocol (R Development Core Team, 2008). The Agilent feature extraction raw data files were loaded into the R environment and analyzed using the R library of Linear Models for Microarray Data (LIMMA) (Smith, 2004). The raw data were preprocessed for background correction and normalized between arrays using the quantile method, summarized by taking the average of replicates for each gene, and subsequently log-transformed.

CD19+ Cell Isolation

Peripheral blood mononuclear cells derived from anonymized healthy donors were obtained from the Centre for AIDS Research Virology Core Lab at UCLA, and the Flow Cytometry and Bone Marrow Lab at the Department Pathology and Laboratory Medicine. All procedures were approved by the local institutional review boards, and the study was considered exempt from review at UCLA. Leukocytes from peripheral blood were isolated using a Ficoll gradient. CD19+ cells were separated using human CD19 Microbeads, LS columns and MACS separator (Miltenyi Biotec). The purity of the sorted CD19+ cells was assessed by FACS. RNA was isolated as per the protocols described below.

Apoptosis, Proliferation and Cell Cycle Analysis

To measure cell proliferation, 2000-4000 cells per well were cultured in 96 well plates. MTS reagents were added according to the manufacturer's instructions (Promega CellTiter 96 Aqueous Non-Radioactive Cell Proliferation Assay kit) and cells were incubated at 37°, 5% CO_2 for 4 hours before absorbance was measured at 490nm. To measure apoptosis, cells were plated, harvested and stained with APC tagged Annexin-V and analyzed by flow cytometry. For cell cycle analysis, cells were collected, washed with PBS, fixed with 70% ethanol and stained with 1X propidium solution in PBS. Cells were subsequently analyzed using flow cytometry.

RT-qPCR

RNA collected from human samples was reverse transcribed using iScript reagent (Quanta Biosciences). RNA from cell lines was reverse transcribed using qScript (Quanta Biosciences). Real Time quantitative PCR was performed with the StepOne Plus Real-Time PCR System (Applied Biosystems) using PerfeCTa SYBR Green FastMix reagent (Quanta Biosciences). The qPCR primer sequences used are listed in Table S4.

Western Blot

Cells were lysed in RIPA buffer (Boston BioProducts) supplemented with Halt Protease and Phosphatase Inhibitor Cocktail (Thermo Scientific). Equal amounts of protein lysate (as quantified by using bicinchoninic acid protein assay, BCA (Thermo Scientific)) were electrophoresed on a 5–12% SDS–PAGE and electroblotted onto a nitrocellulose membrane. Antibodies used were c-MYC Rabbit polyclonal (#9402), CDK6 (DCS83) Rabbit monoclonal (all antibodies from Cell Signaling), IGF2BP3 (N-19, Santa Cruz) and ß-Actin (AC15) mouse monoclonal antibody (Sigma Aldrich). Secondary HRP-conjugated antibodies were purchased from Santa Cruz Biotechnology. SuperSignal West Pico kit (Pierce) was used for enhanced chemiluminescence based detection.

Cell Culture, Plasmids and Spin Infection

The mmu-miR-155 or hsa-miR-21 formatted siRNAs were cloned between the NotI and BamHI sites of a pHAGE6 lentiviral vector (pHAGE6-CMV-siRNA-UBC-ZsGreen) or between the NotI and BamHI sites of a modified pHAGE6 vector downstream of GFP (CMV-GFP-siRNA-UBC-Puromycin).For knockdown in mouse cell lines, a MSCV based vector was used (MSCV-GFP-siRNA-PGK-Puro). The siRNAs were cloned between NotI and XhoI sites downstream of GFP.

For over expression in mice a MIG based vector was used (MSCV-IGF2BP3 CDS-IRES-GFP). The gene of interest was cloned between BglII and XhoI sites. The list of primers used for cloning and sequencing is given in Table S4. For CRISPR-Cas9 mediated targeting, guide RNAs were designed using the Zhang lab website (http://crispr.mit.edu/). Guide RNAs were cloned into the LentiCRISPR vector as done previously (Shalem et al., 2014). Lentivirus production was done in 293T cell line using the helper plasmids pMD2G and psPAX2. RS4; 11 cells were spin-infected at 30°C for 90 minutes in the presence of polybrene. Cells were selected with 5μg/mL of puromycin for 7 days and later used for cell proliferation and apoptosis assays. The human B-ALL cell lines, RS4;11, NALM6, 697 and REH as well as the murine pre-B leukemic cell lines 70Z/3 and WEHI-231 were grown under standard conditions. Lentiviruses were generated using the pHAGE6 constructs and the helper plasmids pHDM-Hgpm2 and pHDM-VSV-G using a 293T packaging cell line as previously described (O'Connell et al., 2010). Retroviruses were generated using the MSCV

based vectors and the pCL-Eco helper plasmid using the 293T cell line (Rao et al., 2010).

Bone Marrow Transplant

CD45.2+ donor mice were injected intraperitoneally with 200 mg/kg of 5-fluorouracil. Mice were sacrificed after 5 days and the bone marrow plated in media enriched with IL-3, IL-6 and mSCF. The bone marrow was infected twice with retroviruses expressing the empty MIG vector or human or mouse IGF2BP3. CD45.1+ recipient mice were lethally irradiated and injected with donor bone marrow 6 hours after irradiation. 8 mice were used per group. These mice were bled at 4, 12 and 16 weeks post bone marrow injection. All mice were housed under pathogen free conditions at the University of California, Los Angeles.

Flow Cytometry

Blood, bone marrow, thymus and spleen were collected from the mice under sterile conditions at 27 weeks post-transplant. Single cell suspensions were lysed in red blood cell lysis buffer. Fluorochrome conjugated antibodies were used for staining (all antibodies obtained from Biolegend). The list of antibodies used is provided in Table S5. For intracellular staining, after initial staining with surface marker antibodies and fixation with 1% paraformaldehyde (PFA), cells were

161

incubated with antibodies against intracellular antigens (Ki67, Cdk6 and Myc) in 1% Triton containing MACS buffer. After 30 minutes of staining at 4^0C, cells were washed twice with PBS and fixed with 1% PFA. Flow cytometry was performed at the UCLA Jonsson Comprehensive Cancer Center (JCCC) and at the BROAD Stem Cell Research Flow Core. Analysis was performed using FlowJo software.

Histopathology

Organs were collected after necropsy and fixed in 10% buffered formalin. These were then embedded in paraffin, processed for hematoxylin and eosin staining by the Translational Pathology Core Laboratory at UCLA. Histopathologic analysis was performed by a board certified hematopathologist (D.S.R).

Statistical Analyses

Figures are graphed as mean with the standard deviation of the mean (SD) for continuous numerical data. Bar graphs are employed to show dichotomized or ordinal-type histopathologic data. Statistical tests were performed using GraphPad Prism software and applied to each experiment as described in the figure legends.

iCLIP

iCLIP was performed as previously described (Konig et al., 2010). Briefly, RS4;11 or REH cells were irradiated with UV-C light to form irreversible covalent cross-link proteins to nucleic acids *in vivo*. After cell lysis, RNA was partially fragmented using low concentrations of Micrococcal nuclease, and IGF2BP3–RNA complexes were immunopurified with anti-IGF2BP3 antibody (MBL International) immobilized on protein A–coated magnetic beads (Life Technologies). After stringent washing and dephosphorylation (FastAP, Fermentas), RNAs were ligated at their 3' ends to a 3' pre-adenylated RNA adaptor and radioactively labeled to allow visualization. Samples were run using MOPS-based protein gel electrophoresis (in-house recipe) and transferred to a nitrocellulose membrane. Protein-RNA complexes migrating 15-80 kDa above free protein were cut from the membrane, and RNA was recovered from the membrane by proteinase K digestion under denaturing (3.5 M Urea) conditions. The oligonucleotides for reverse transcription contained two inversely oriented adaptor regions adapted from the Bioo NEXTflex small RNA library preparation kit (Bioo Scientific), separated by a BamHI restriction site as well as a barcode region at their 5' end containing a 4-nt experiment-specific barcode within a 5-nt random barcode to mark individual cDNA molecules. cDNA molecules were size-purified using denaturing PAGE gel electrophoresis, circularized by CircLigase II (Epicenter), annealed to an oligonucleotide complementary to the restriction site and cut using BamHI (NEB). Linearized cDNAs were then PCR-

amplified using (Immomix PCR Master Mix, Bioline) with primers (Bioo) complementary to the adaptor regions and were subjected to high-throughput sequencing using Illumina HiSeq. A more detailed description of the iCLIP protocol has been published (Huppertz et al., 2014).

iCLIP Data Analysis

Following transcriptomic and genomic alignment with 'Tophat2'(Kim et al., 2013). Individual reads were truncated to their 5' ends to represent the site of crosslinking consistent with the iCLIP methodology. In order to call specific sites of protein-RNA interaction, 30bp regions of enrichment over background were determined using 'Piranha'(Uren et al., 2012) in zero-truncated negative binomial mode with a custom local covariate. The covariate was calculated by uniformly distributing the crosslink number of each 30bp bin across the neighboring 6 bins (3 on each side) to control for regions of overall higher depth rather than site specific pile-up of reads as an indication of protein-RNA interaction. Adjacent bins with significant p-values from Piranha (alpha < 0.05) were combined into single regions. For REH and RS4;11 samples, only overlapping peak regions found to be statistically significant in all three replicates were considered to be biologically reproducible candidates for further analysis. To derive the intragenic distributions of CLIP-seq sites, we queried the UCSC Genome Browser MySQL database for hg19 (Meyer et al., 2013) and determined the nearest overlapping gene based first on CCDS gene

164

annotations (Pruitt et al., 2009) to determine the canonical ORF, and second on Gencode V19 comprehensive for other features (Harrow et al., 2012). Following this, the nearest intragenic anchor (transcription start site, start codon, 5' splice site, 3' splice site, stop codon, and polyadenylation site) was recorded and the spliced distance to the nearest ORF boundary was calculated. As a background control, uniformly distributed crosslink sites were simulated by pseudo random intervals from hg19 calculated using 'bedtools'(Quinlan and Hall, 2010). The intragenic distribution of these sites was determined following the same methodology as the iCLIP crosslink sites.

In order to determine the binding specificity of IGF2BP3, a 10nt window surrounding each crosslink site occurring within a biologically reproducible and statistically significant peak from a `last exon' was extracted and the counts of each n-mer (4 through 6) calculated. As a control, random 20bp intervals from a window of 100 to 300 nucleotides either upstream or downstream of each crosslink site was included as a normalized frequency of n-mers to represent the background probability (p) of observing each n-mer. For each n-mer size, the probability of observing k occurrences of some n-mer out of N total observations is binomially distributed (k ~ Bin(p, N)) and Poisson approximated given sufficiently large values for N (>1000) and small values of p (<0.01). Individual n-mers with Poisson-approximated p-values significant at a 5% false discovery rate (Klipper-Aurbach et al., 1995) were then aligned with one another, and grouped into motifs using k-medoid clustering for

165

optimal values of k (determined using average silhouette width). The frequency of occurrence of each nucleotide was then plotted in a position-specific manner for each motif cluster.

Luciferase Assays

The CDK6 UTR is ~10kb long while the MYC UTR is ~300bp. We cloned the MYC UTR downstream of firefly luciferase in the pMirGlo vector between the SacI and XhoI sites. pMirGlo is a dual luciferase vector having Renilla luciferase downstream of the PGK promoter. The CDK6 UTR was divided into 5 pieces (CDK6 1-5; ~2kb each) and cloned individually downstream of the firefly luciferase. 293T cells were transfected with the pMirGlo or UTR containing reporter vectors along with the MIG empty vector or the MIG-hIGF2BP3 overexpression vector at a 1:10 ratio (50ng: 500ng). Co-transfections were performed with Lipofectamine 2000 (Life Technologies) as per the manufacturer's instructions. Cells were lysed after 24 hours, substrate was added and luminescence was measured on a Glomax-Multi Jr (Promega). The ratio of firefly to Renilla luciferase activity was calculated for all samples. The hIGF2BP3/MIG luminescence for the pMirGlo empty vector was used as a normalization control.

IGF2BP3 RNA Immunoprecipitation

600 uL of Protein A Dynabeads (50% slurry; Life Technologies) were washed four times with 100 mM Sodium Phosphate buffer (NaP; pH 8.1) and divided into two tubes in 1.5 mL NaP buffer. Beads were treated with either 21 μg IgG mouse antibody (Jackson Laboratories) or α-IGF2BP3 (D-7; Santa Cruz Biotechnology) for 1 hr spinning at 4°C. Beads were then washed three times with RSB-100 buffer (10 mM Tris-Cl at pH 7.4, 100 mM NaCl, 2.5 mM MgCl2, 0.5% NP-40)and split into 3 tubes each for replicates. Cytoplasmic lysates were prepared from suspensionRS4;11 cells in RSB-100 buffer containing RNase inhibitors. After centrifugation for 10 min, supernatants were combined with the bead/antibody and rotated at 4°C overnight. Beads were washed five times with cold RSB-100 and re-suspended in 1 mL of buffer. 200 μL of bead slurry was removed for Western Blot analysis. Buffer was removed from rest of beads and resuspended in 200 μL of Proteinase K buffer (100 mM Tris-HCl, pH 7.4, 50 mM NaCl, 10 mM EDTA, pH 8.0). Samples were treated with 10 μL of RQ DNase I (1 U/μL; Promega) and then treated with 10 μL ofProteinase K (20 mg/mL; Ambion), both incubated while shaking for 20 min. at 37°C. The supernantant volume was increased to 400 μL with Proteinase K buffer and the RNA was extracted with addition of 400 μl of acid phenol–chloroform and precipitated with sodium acetate, absolute ethanol, and addition of 0.75 μL of co-precipitate GlycoBlue (Ambion). RNA was recovered by centrifugation, washed and resuspended in 21 μl of RNase free water. Quantity and quality were checked with

Nanodrop and Total RNA NanoBioanalyzer kit (Agilent). 200 μg of each RNA sample (triplicates of both IgG controls and α-IGF2BP3 immunoprecipitations) were subjected to reverse transcription using the High-Capacity cDNA Reverse Transcription Kit (Thermo-Fisher Scientific) and subsequent qPCR using the Lightcycler® 480 (Roche). Statistical analyses were performed using GraphPad. (MYC; Forward 5'-GCTGCTTAGACGCTGGATTT-3', Reverse 5'-TAACGTTGAGGGGCATCG-3'.CDK6;Forward:5'-TGATCAACTAGGAAAAATCTTGGA,Reverse:5'GGCAACATCTCTAGGCCAGT-3'; 18s rRNA reference; Forward: 5'-CTTCCACAGGAGGCCTACAC-3',Reverse:5'-CGCAAAATATGCTGGAACTTT-3')

RNA sequencing experiments.

RNA was purified from both cytosolic fractions of IGF2BP3-depleted or control RS4;11 cells using TRI-Reagent LS (Sigma), converted to double stranded libraries using the NEXTflex™ Rapid Directional qRNA-Seq™ Library Prep Kit (Bioo Scientific Corp., Austin, TX), and sequenced using Illumina HiSeq2500 platform.

High-throughput RNA sequencing data generated by Illumina HighSeq 2500 and corresponded to ~30-115M reads per sample was mapped to hg19 build of the human genome (Feb. 2009 GRCh37, NCBI Build 37.1) using Bowtie and TopHat (Langmead et al., 2009; Trapnell et al., 2009). All data collection and parsing was performed with the Perl programming language, and statistical analyses with the R

statistical language, version 2.14.1. All external library packages used are available on CPAN or CRAN, respectively. Differentially expressed genes were identified using DEseq

Dissertation author contributions

J.M.H: Performed all iCLIP experiments in B-ALL cells. Performed RNA-seq experiment in control and knockdown IGF2BP3 in RS4;11 cells as well as analysis. Helped write manuscript.
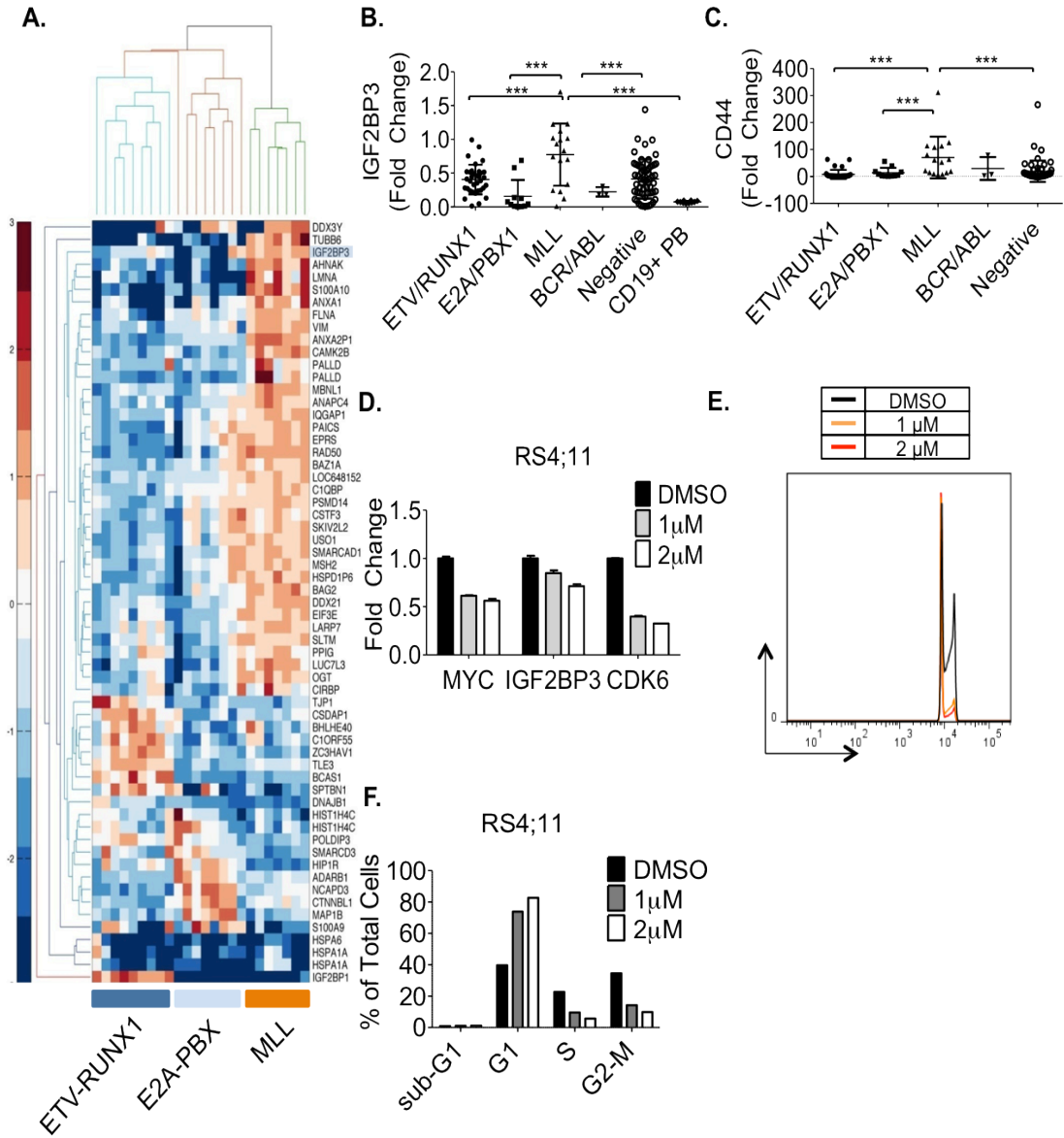
Study Approval

Written informed consent was obtained from all of the parents of the patients by the AIEOP (Italian Association of Pediatric Hematology and Oncology) and the BFM (Berlin-Frankfurt-Muenster) ALL-2000 trial. The local institutional review boards approved all procedures, and the study was considered exempt from review at UCLA. Peripheral blood mononuclear cells derived from anonymized donors were obtained from the Center for AIDS Research Virology Core Lab at UCLA. Written informed consent was obtained from all the donors in a manner approved by the UCLA Human Subject Protection Committee. Additional samples were obtained following diagnostic work from the UCLA Department of Pathology and Laboratory Medicine with IRB approval. All mouse experimental procedures were conducted

with the approval of the UCLA Institutional Animal Care and Use Committee, known

as the Chancellor's Animal Research Committee (ARC).

This page has been left intentionally blank for printing purposes.

Figure 1: IGF2BP3 is overexpressed in MLL-translocated B-ALL. (A) Heatmap from the microarray data showing differentially expressed RNA binding proteins between B-ALL. IGF2BP3 is highly expressed in MLL-rearranged B-ALL, (B and C) RT-qPCR based confirmation of overexpression of IGF2BP3 and its previously defined target, CD44, respectively, in MLL-rearranged B-ALL (Total n=134; t-test, p<0.0001 for all comparisons) (D-G) Treatment of RS4;11 (D-F) cell line with increasing doses of MLL inhibitor IBET151. (D) RT-qPCR of MYC, CDK6 and IGF2BP3 levels in RS4;11 cells shows a significant decrease in all three mRNA levels (T-test; all p-values <0.05). (E and F) Cell cycle analysis by propidium iodide staining after IBet151 treatment of RS4;11 cells shows G1-arrest secondary to CDK6 inhibition (F). RT-qPCR assays were normalized to Actin (B-C) and RNA Pol II (D). Error bars denote S.D. See also Figure S1.

A.

DDX3Y
TUBB6
IGF2BP3
AHNAK
LMNA
S100A10
ANXA1
FLNA
VIM
ANXA2P1
CAMK2B
PALLD
PALLD
MBNL1
ANAPC4
IQGAP1
PAICS
EPRS
RAD50
BAZ1A
LOC648152
C1QBP
PSMD14
CSTF3
SKIV2L2
USO1
SMARCAD1
MSH2
HSPD1P6
BAG2
DDX21
EIF3E
LARP7
SLTM
PPIG
LUC7L3
OGT
CIRBP
TJP1
CSDAP1
BHLHE40
C1ORF55
ZC3HAV1
TLE3
BCAS1
SPTBN1
DNAJB1
HIST1H4C
HIST1H4C
POLDIP3
SMARCD3
HIP1R
ADARB1
NCAPD3
CTNNBL1
MAP1B
S100A9
HSPA6
HSPA1A
HSPA1A
IGF2BP1

ETV-RUNX1 E2A-PBX MLL

B. IGF2BP3 (Fold Change)

ETV/RUNX1 E2A/PBX1 MLL BCR/ABL Negative CD19+ PB

C. CD44 (Fold Change)

ETV/RUNX1 E2A/PBX1 MLL BCR/ABL Negative

D. RS4;11

Fold Change

MYC IGF2BP3 CDK6

DMSO
1μM
2μM

E.

DMSO
1 μM
2 μM

F. RS4;11

% of Total Cells

sub-G1 G1 S G2-M

DMSO
1μM
2μM

Figure 2: IGF2BP3 knockdown leads to disruptions of cell growth and increased apoptosis. (A) IGF2BP3 expression in human B-ALL cell lines (B) Schematic of lentiviral vector used for IGF2BP3 knockdown (C) IGF2BP3 (denoted I3) knockdown, measured by RT-qPCR shown in RS4;11 cell line (t-test, p=0.0005) (D) Cell cycle analysis with propidium iodide staining (E) MTS assay showing significantly reduced cell proliferation with IGF2BP3 knockdown (F) Western blot showing IGF2BP3 expression after Cas9/CRISPR mediated targeting using the Cr1 or Cr2 constructs. Cr1-mediated targeting results in some residual protein. Beta-Actin is used as a loading control (G) MTS assay showing significantly reduced cell proliferation after Cr2 targeting (t-test, p≤0.01 for all marked comparisons) (H) Cell cycle analysis by propidium iodide staining showing increased cell death (sub-G1 peak) in Cr2 expressing cells. (I) Increased Annexin V staining in Cr2-targeted cells with IGF2BP3 knockout. Error bars denote S.D. See also Figure S2.
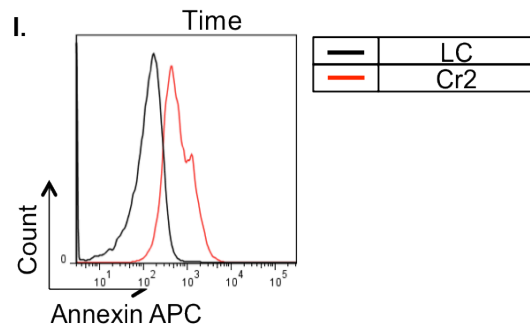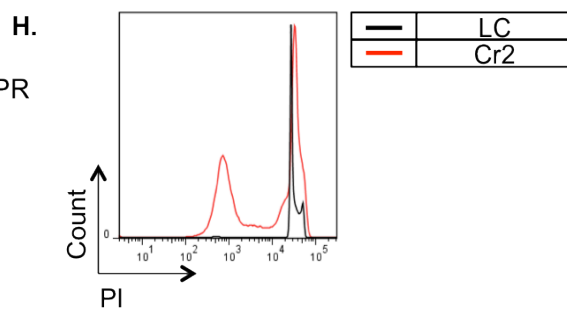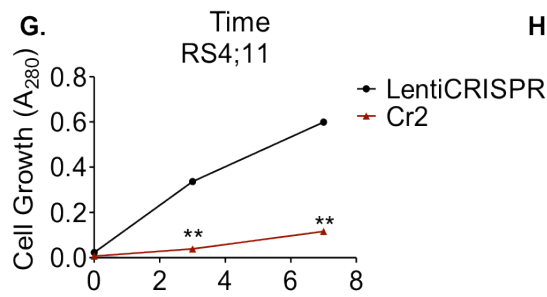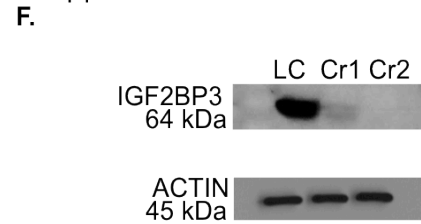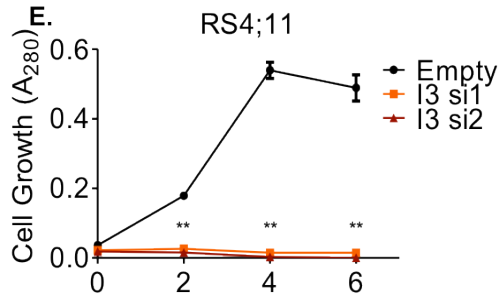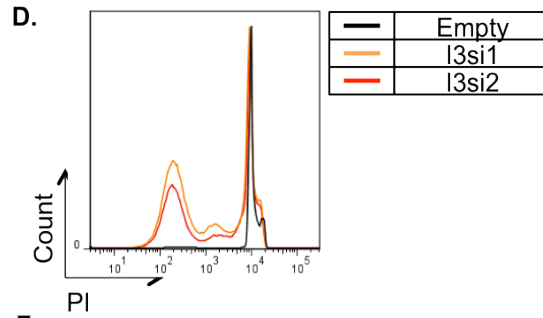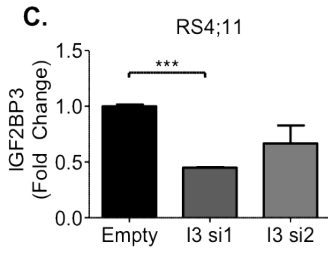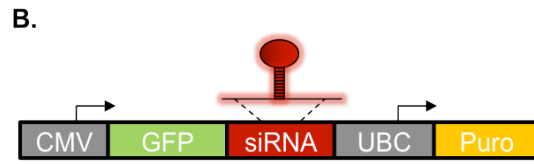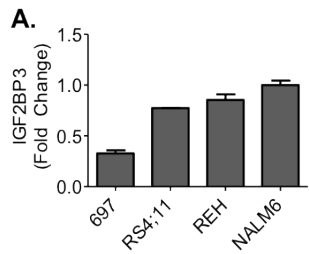
Figure 3: Enforced expression of IGF2BP3 leads to enhanced engraftment and skewing towards B-myeloid development. (A) Schematic diagram of the bicistronic vector used for enforced expression of IGF2BP3 (B) Western blot showing overexpression of IGF2BP3 in the murine pre-B cell line, 7OZ/3, and the human embryonic kidney cell line, 293T, at the protein level and (C) RT-qPCR showing overexpression in the 7OZ/3 cell line at the mRNA level (t-test, p=0.0013) (D) FACS analysis of peripheral bleed from mice 6 weeks after bone marrow transfer showing successful engraftment and transduction (GFP+) (E) FACS of peripheral blood done at 4 weeks after bone marrow transplantation, showing CD45.2 (donor marrow) and GFP positivity (indicating transduction) (t-test for hI3, p<0.0001; t-test for mI3, p=0.0002). (F) Quantitation of GFP expression in the peripheral blood between 4 and 16 weeks post-transplant shows that the effect is significant and sustained (G) Peripheral blood leukocyte counts at 16 weeks post-transplant show increased leukocytes in the peripheral blood (t-test, p<0.0001) (H) Significantly higher numbers of B220+ cells and (I) CD11b+ cells are also seen, determined by FACS staining of peripheral blood (t-test for B220+ cells p<0.0001; t-test for CD11b+ cells, p=0.0004). (J) FACS based enumeration of T-cells shows no significant change in circulating T-cells. (K-L) Enumeration of red blood cells and platelets by complete blood count show significant reductions (t-test for RBCs, p=0.0005; t-test for platelets, hI3, p=0.0322; t-test for platelets, mI3, p=0.0004). In this experiment, n=8 for all three groups (Abbreviations, hI3: hIGF2BP3, mI3: mIGF2BP3) Error bars denote S.D.

176

A.

LTR | IGF2BP3 | IRES | GFP

B.

7OZ/3
MIG hIGF2BP3

293T
MIG hIGF2BP3

IGF2BP3
64 kDa

ACTIN
45 kDa

C. 7OZ/3

IGF2BP3 (Fold Change)

MIG   hI3

D.

CD 45.2

GFP

E. Engraftment

% CD45.2+

MIG   hI3   mI3

F.

% CD45.2+ GFP+

Weeks

- MIG
- hI3
- mI3

G. WBC Count

Cell Number (K/uL)

MIG   hI3   mI3

H. B-cells Absolute Count

Cell Number (K/uL)

MIG   hI3   mI3

I. Myeloid Cells Absolute Count

Cell Number (K/ul)

MIG   hI3   mI3

J. T-cells Absolute Count

Cell Number (K/ul)

MIG   hI3   mI3

K. RBC Count

Cell Number (M/uL)

MIG   hI3   mI3

L. Platelet Count

Cell Number (K/uL)

MIG   hI3   mI3

177

Figure 4: Analysis of bone marrow progenitor populations from IGF2BP3 overexpressing mice. (A) Enumeration (left panel) and representative flow cytometry histograms to define HSCs from control vector (second from left), human IGF2BP3 (second from right) and murine IGF2BP3 (right panel) overexpressing mice (t-test for hI3, p=0.0057; t-test for mI3, p=0.0007). (B-C) Analysis for LMPPs and CLPs from mice noted as in A. Statistically significant differences were found in LMPPs and CLPs (t-test for LMPPs hI3 and mI3, p=0.0003 and p=0.0028, respectively; t-test for CLPs hI3 and mI3, p=0.0176 and p=0.0463, respectively). (D) Intracellular Ki67 staining and FACS based analyses, depicted in the same manner, with enumeration on the left hand side, within the LSK population enriched for hematopoietic stem cells. Significant differences in the high Ki67 expressing population were found (t-test for hI3, p=0.0103; t-test for mI3, p=0.0009). (E) Intracellular Ki67 staining and FACS analysis of proliferation in the LMPP population shows significant differences in the proliferative fraction (t-test for hI3, p=0.0003; t-test for mI3, p<0.0001). Error bars denote S.D. See also Figure S3.

A. HSC

B. LMPP

C. CLP

D. LSK Ki67 hi

E. LMPP Ki67 hi
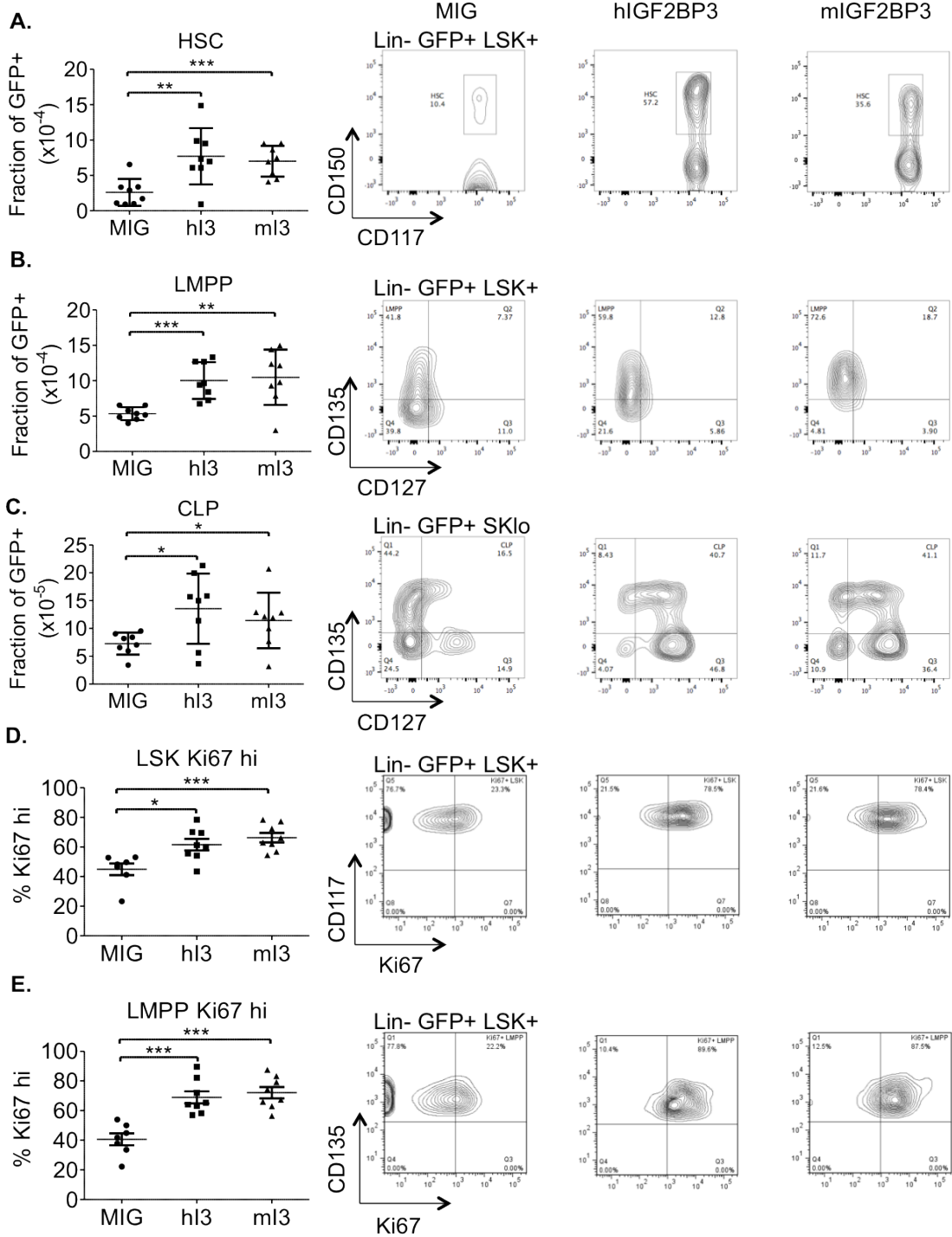
Figure 5: Analysis of thymic cellular composition from IGF2BP3 overexpressing mice. (A-C) Enumeration of CD45.2, GFP, and CD3ε+ cells, respectively, in thymi from control and human and murine IGF2BP3 overexpressing mice show no statistically significant differences. (D) Enumeration and representative FACS plots showing an increase in B220+ cells in the thymus of mice with enforced expression (t-test for hI3, p=0.0444; t-test for mI3, p=0.0145). (E) Histologic images of thymic sections from mice with enforced expression of IGF2BP3. Hematoxylin and eosin staining, scale bar, 40 μm. Error bars denote S.D. See also Figure S4.

A.

B.

C.

D.

MIG hIGF2BP3 mIGF2BP3

E.

MIG hIGF2BP3 mIGF2BP3
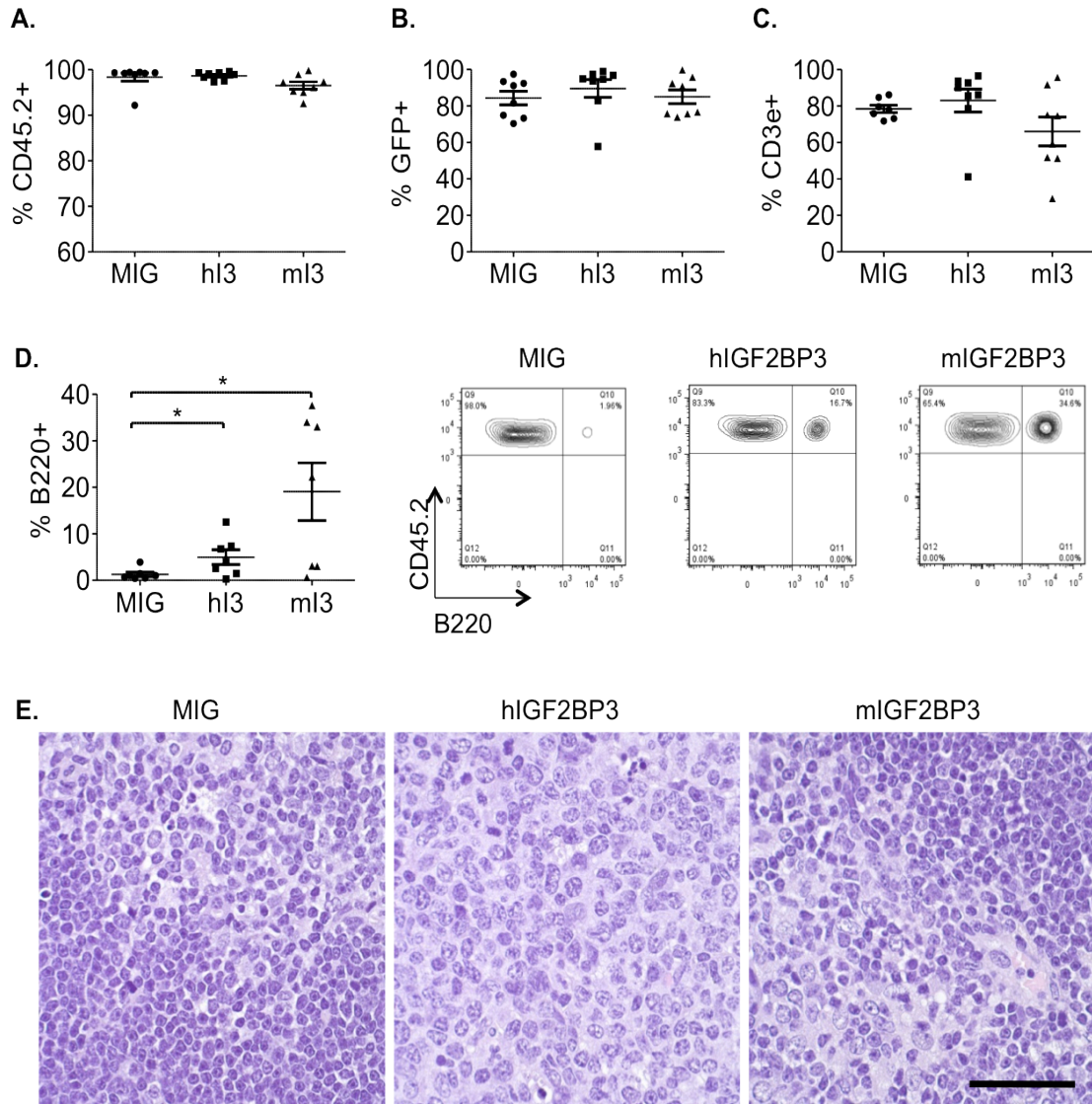
Figure 6. iCLIP analysis of IGF2BP3 in human leukemia cell lines. (A) Proportion of IGF2BP3 (REH and RS4;11 cells), hnRNPA1(HEK cells) and simulated (Genome) crosslinking sites observed in exons, introns or unannotated regions of the human genome. (B) Proportion of IGF2BP3 (REH and RS4;11 cells), hnRNPA1 (HEK cells) and simulated (mRNA background) binding sites in coding and noncoding exons. (C) IGF2BP3 (REH, RS4;11) and hnRNPA1 (HEK cells) crosslink site density relative to termination codons. (D) Tetramer sequence enrichment at IGF2BP3 crosslinking sites in RS4;11 and REH cells (upper and lower panel, respectively). (E) KEGG pathway enrichment in REH and RS4;11 cells (black and yellow, respectively). Genes identified by iCLIP that are associated with the CML pathway are shown. (F,G) UCSC Genome Browser snapshot of the CDK6 and MYC loci, respectively. Each panel shows the exon-intron structure of the gene, sequence conservation across vertebrate species and unique read coverage from two iCLIP replicates from each cell line. The maximum number of reads at each position is indicated to the left of each histogram. See also Figures S5 and S6.

A.

B.

C. Downstream (Stop Codon)

D. Best 4-mer Cluster

E.

F. CDK6

G. MYC
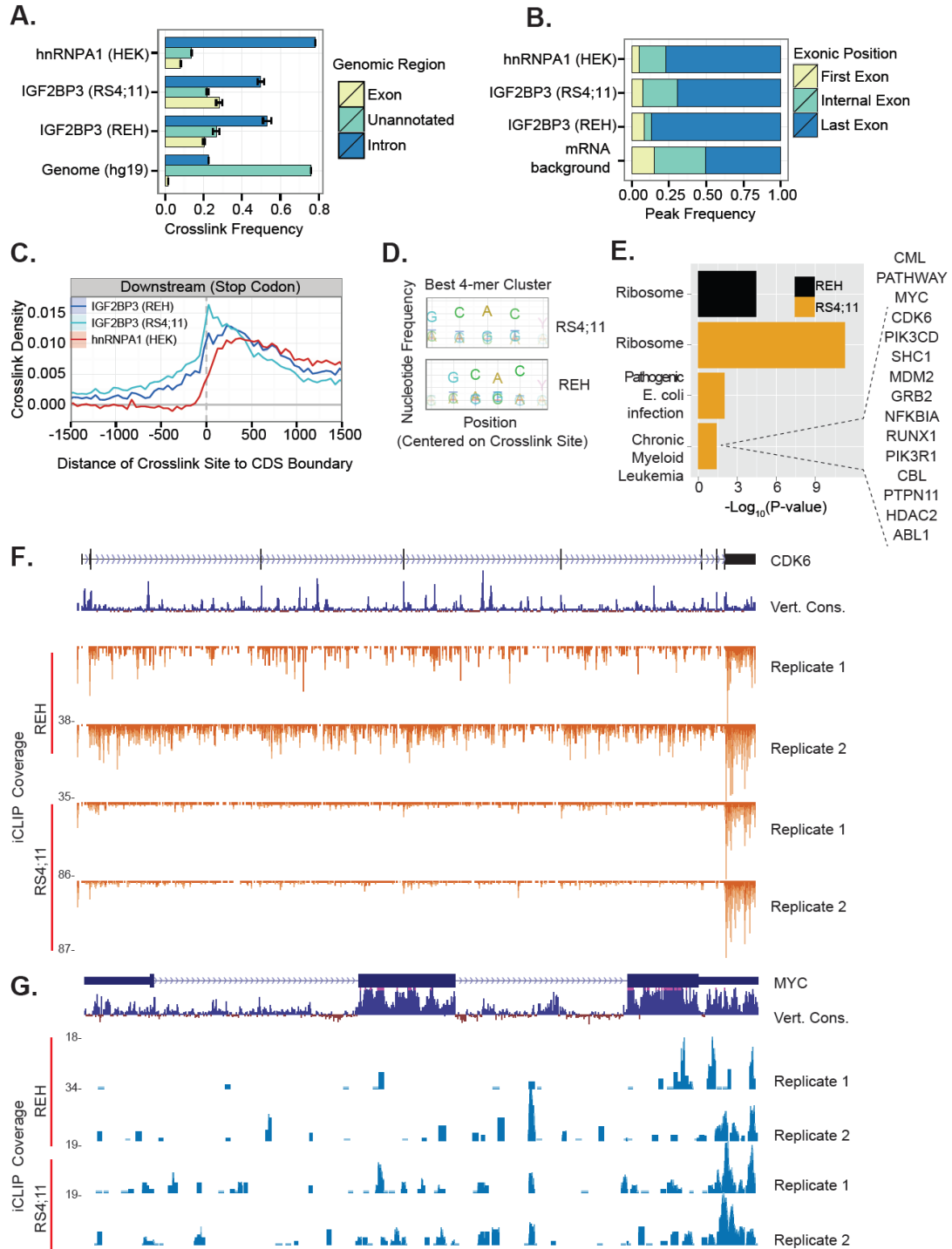
Figure 7: CDK6 and MYC are targeted by IGF2BP3 (A) Schematic of the luciferase assay used (B) Luciferase assay showing targeting of the CDK6 and MYC 3' UTRs by IGF2BP3 (C-D) CDK6 analysis of bone marrow progenitors shows a significantly increased amount of CDK6 protein in the GFP+ BM cells (C, t-test, p=0.0213) but not in the GFP- cells. Similarly intracellular staining for MYC reveals significantly increased levels in the GFP+ BM cells after IGF2BP3 enforced expression (E-F, t-test, p<0.0001) but not in the GFP- cells (CDK6 and MYC protein levels shown as the Mean Fluorescence Intensity, MFI) (n=8 for all three groups) (G and H) In the RS4;11 cell line, western blot confirmed knockdown of IGF2BP3 protein and reduced expression of CDK6 protein. After Igf2bp3 knockdown in 7OZ/3 cells (t-test, p-value=0.0007 and 0.0005 for IGF2BP3 si1 and si2 respectively), there is reduced expression of CDK6 and MYC protein. Error bars denote S.D. See also Figure S7.

A.

SV40 | Firefly | UTR | PGK | Renilla

+ MIG-Empty → Firefly/Renilla ratio equal

+ MIG-hIGF2BP3 → ↑↑ Firefly/Renilla ratio

B.

293T

Normalized Luciferase Activity

pMirGlo, CDK6-1, CDK6-2, CDK6-3, CDK6-4 *, CDK6-5 *, MYC *

C.

GFP+ CDK6

PE: MFI

MIG hl3 ml3

*

D.

GFP- CDK6

PE: MFI

MIG hl3 ml3

E.

GFP+ MYC

PE: MFI

MIG hl3 ml3

***

F.

GFP- MYC

PE: MFI

MIG hl3 ml3

G.

7OZ/3

IGF2BP3 (Fold Change)

MGP l3 si1 l3 si2

*** ***

H.

RS4;11   7OZ/3

CDK6   P6-E si1 si2   MGP si1 si2   CDK6
36 kDa   36 kDa
             1.00 0.35 0.74   1.00 0.52 0.62

IGF2BP3   MYC
64 kDa   48 kDa
             1.00 0.26 0.12   1.00 0.54 0.43

ACTIN   ACTIN
45 kDa   45 kDa

185

Figure S1: MLL-AF4 target gene expression by microarray. Scatter plots comparing the expression of (A) CDK6 (B) MYC (C) HOXA9 and (D) MEIS1 from the corresponding normalized fluorescence intensities on the microarray. (n=14 for ETV/RUNX1 samples; n=15 for E2A/PBX1 samples and n=15 for MLL rearranged samples) (t test, p< 0.0001 for all comparisons). Error bars denote S.D.
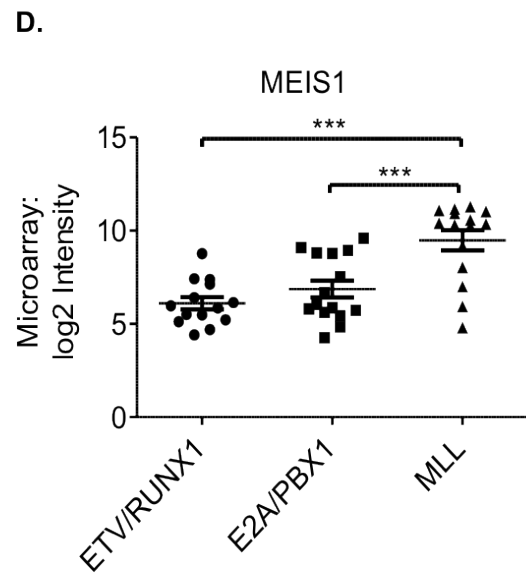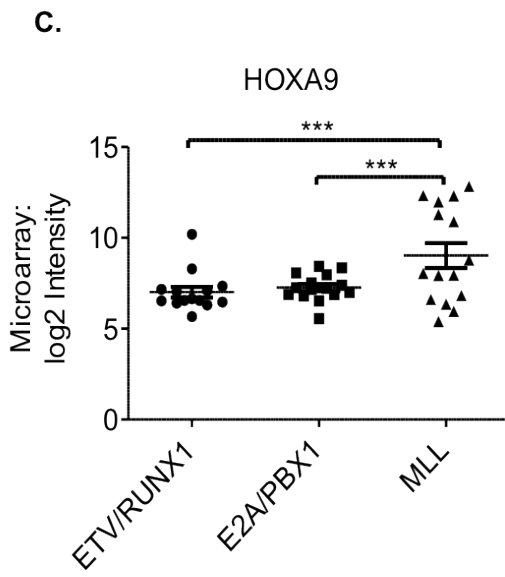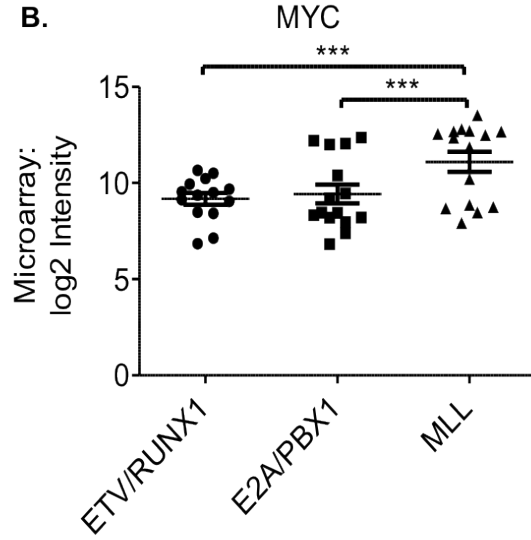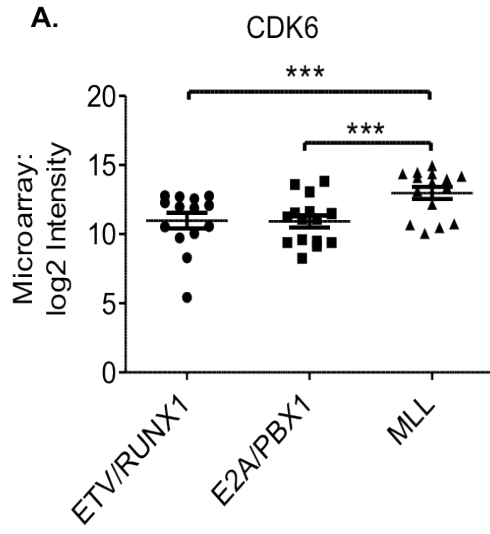
Figure S2: Knockdown of IGF2BP3 in B-ALL cell lines (A) Schematic of the lentiviral vector used for IGF2BP3 knockdown (B) IGF2BP3 knockdown by lentiviruses in NALM6 cell line (t test, p< 0.05) and (C) Reduced cell proliferation (MTS assay) in NALM6 after IGF2BP3 knockdown (t test, p<0.001 for all comparisons) (D) T7 Endonuclease assay of PCR done using primer pair F1 and R1 (top panel) and DNA extracted from RS4;11 cells stably integrated with LentiCRISPR, Cr1 or Cr2 (Lanes 1-3 and 5-7). On addition of T7 enzyme cleavage, is seen only in the Cr1 integrated cells (arrowheads). Bottom panel shows T7 assay done using the same samples and primer pair F2 and R2. Cleavage is seen only in the Cr2 integrated cells (arrowheads) (E) Human IGF2BP3 gene structure showing the location of exons (in orange) and targets of the CRISPR guide RNAs (Cr1 and Cr2); Primer pairs F1 and R1 flanking the target site of Cr1 as well as primer pairs F2 and R2 flanking the Cr2 target site are also shown. Error bars denote S.D.

A.

B.

NALM6

C.

NALM6

D.

Primer pair F1+R1

No T7 Enzyme   + T7 Enzyme

Primer pair F2+R2

No T7 Enzyme   + T7 Enzyme

E.

189
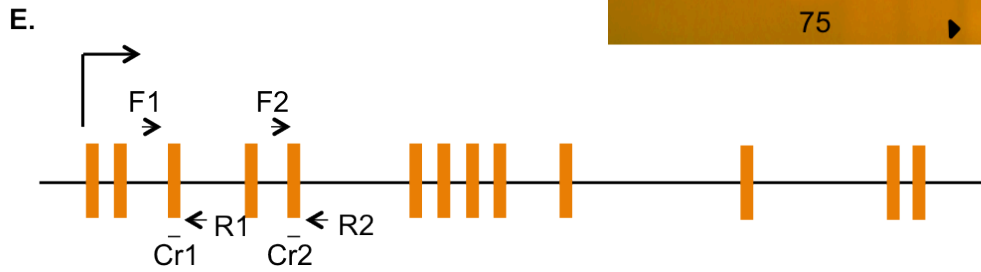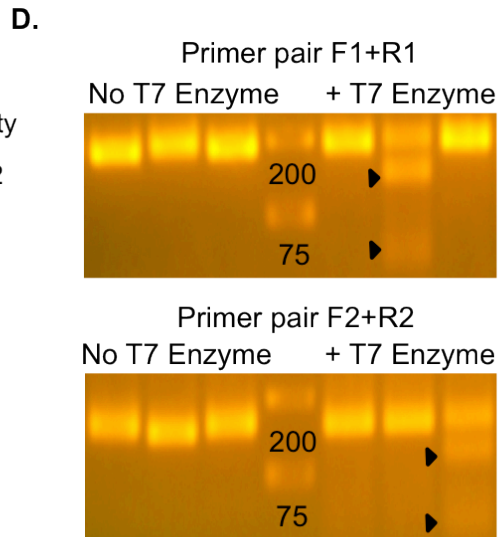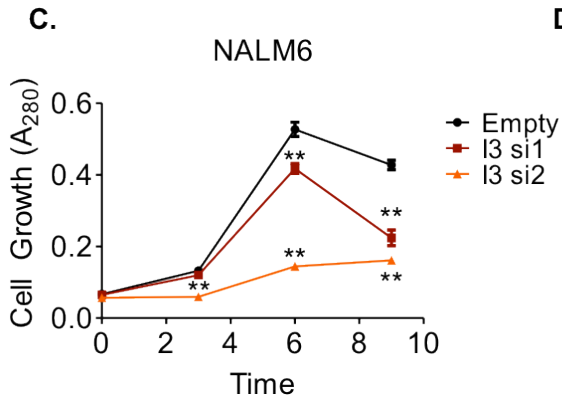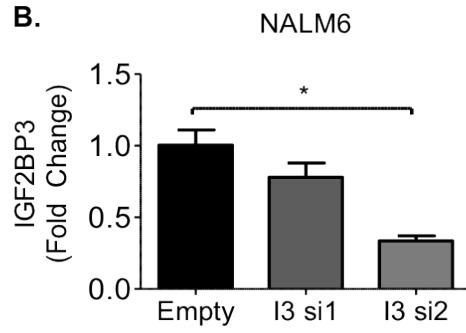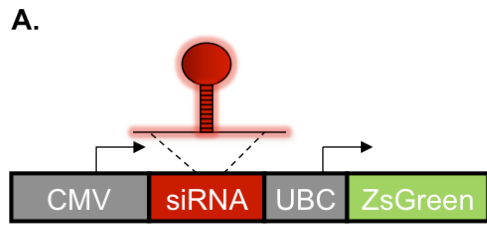
Figure S3: Bone marrow immunophenotypic analyses. Analysis at the end of 6 months after IGF2BP3 overexpression shows (A) similar engraftment, (B) increased GFP expressing cells (t test for hI3 and mI3, p=0.0022 and p=0.0045, respectively) and (C) similar numbers of myeloid and (D) B-cells (n=8 for all three groups). (E and F) Confirmation of over expression of human and mouse IGF2BP3 in mouse bone marrow by qPCR. Mouse Actin was used as internal control (G) Ki67 staining of CLPs showing no significant difference and (J) representative FACS plots. (H-I) Increased Hardy fractions A (t test for hI3 and mI3 , p=0.0115 and p=0.0003, respectively) and B (t test, p≤ 0.05 for both comparisons) after IGF2BP3 over expression with (K) representative plots. Error bars denote S.D.

191
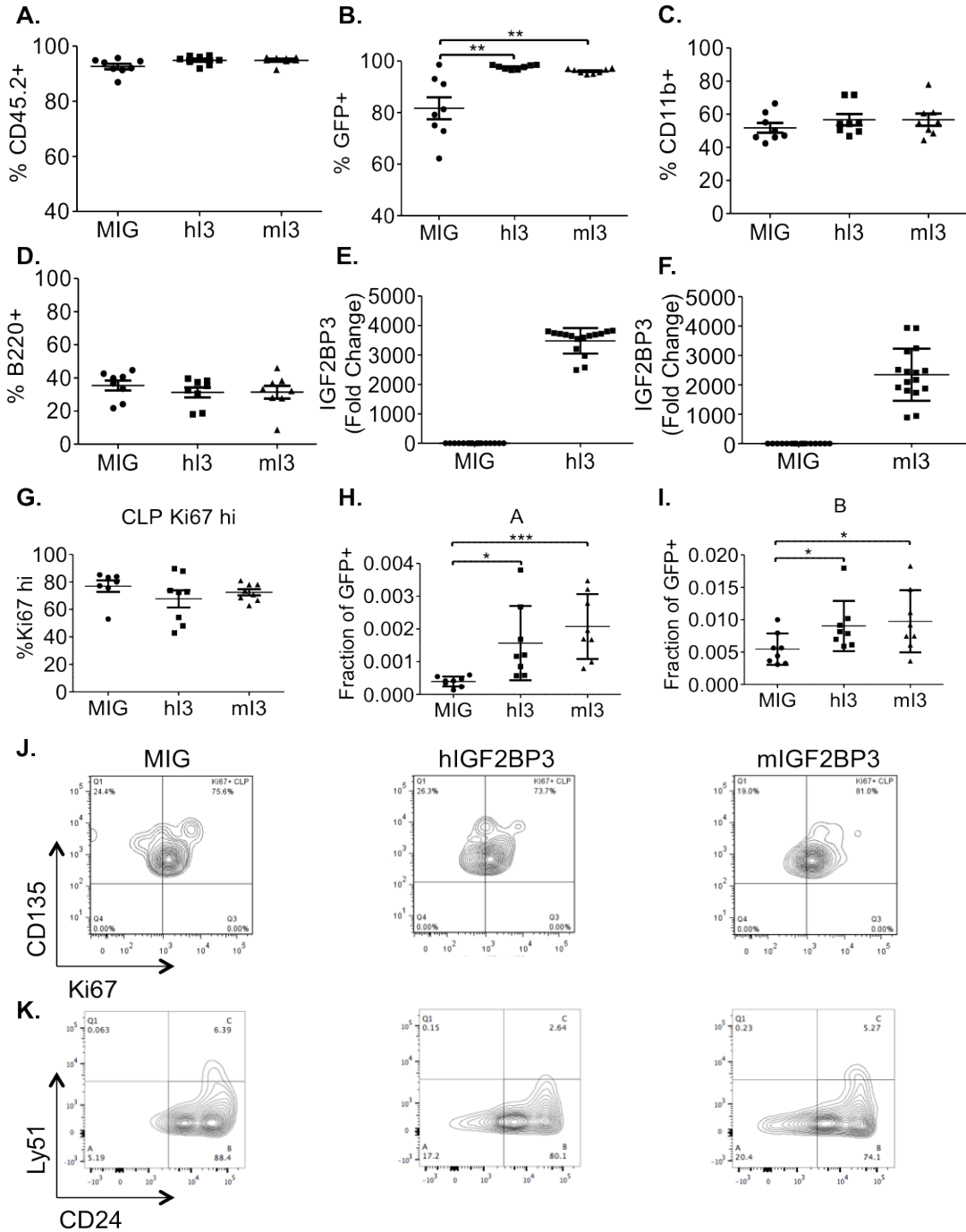
Figure S4: Immunophenotypic analyses of spleen with enforced expression of IGF2BP3 (A) Spleen analysis at the end of 27 weeks after IGF2BP3 overexpression shows significantly elevated spleen weights (t test, p=0.0021), (B) similar engraftment, (C) GFP expressing cells and (D) B-cells. (E) Decreased numbers of T-cells (t test, p≤ 0.05 for both comparisons) and (F) elevated myeloid cells are also seen in the spleen (t test, p≤ 0.05) (n=8 for all three groups). Error bars denote S.D.
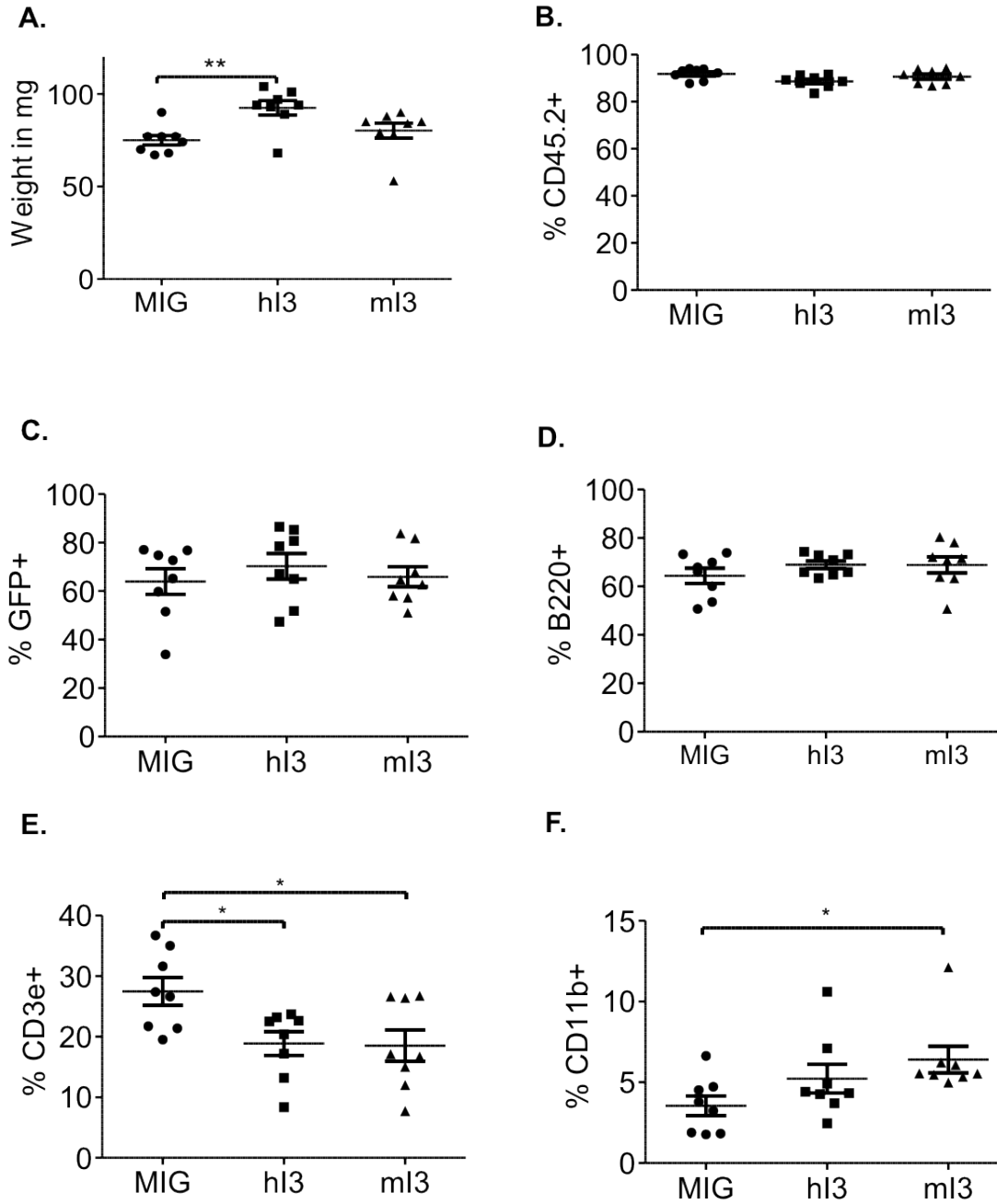
Figure S5: Purification of IGF2BP3-RNA complexes from human leukemia cell lines using crosslinking immunoprecipitation (iCLIP). (A) Western blot of RS4;11 whole cell lysate and anti-IGF2BP3 immunoprecipitate. Membrane is probed with anti-IGF2BP3. IgG heavy chain (HC) and a cross-reactive protein (*) are indicated on the blot. (B) Autoradiograph of protein-RNA complexes isolated by CLIP from RS4;11 cells. UV-dependent and nuclease-sensitive IGF2BP3-RNA complexes are indicated. (C and D). As in panels A and B, but using REH cell lysates. (E) Bioanalyzer trace of pooled amplicon libraries from REH and RS4;11 cells (lanes 2 and 3, respectively).

A.

RS411

| Input | IP |

IGF2BP3

HC

*

IP: α-IMP3 (MBL)
Blot: α-IMP3 (Bethyl)

B. RS411 B-Cell IGF2BP3 CLIP

[MNase]:	+	+	+++	+
Ab:	-	+	+	+
UV:	+	-	+	+

IGF2BP3-
RNA
Complex

1   2   3   4

C.

REH

| Input | IP |

IGF2BP3

HC

IP: α-IMP3 (MBL)
Blot: α-IMP3 (Bethyl)

D. REH B-Cell IGF2BP3 CLIP

[MNase]:	+	+	+++	+
Ab:	-	+	+	+
UV:	+	-	+	+

IGF2BP3-
RNA
Complex

1   2   3   4

E.

Cell Line

| Ladder | REH | RS411 |

200
150
100

50
35

IGF2BP3
iCLIP
Libraries

1   2   3

195
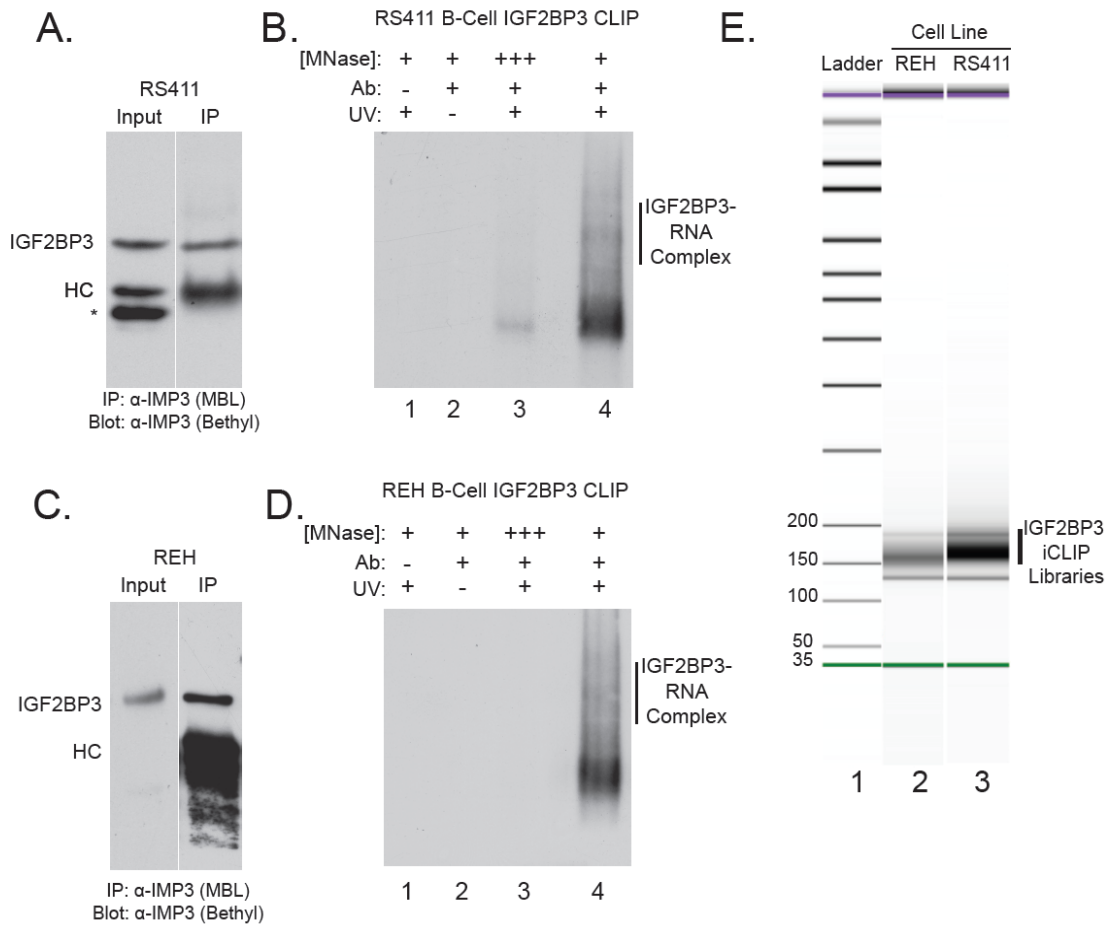
Figure S6: Annotation and overlap of IGF2BP3 binding sites identified in human leukemia cell lines. (A) Annotation of peaks called in RS4;11 cells. (B) Annotation of peaks called in REH cells.(C) Pie charts showing the count of significant CLIP peaks in different regions of protein coding genes. Venn Diagram of genes identified in the iCLIP analysis of IGF2BP3 in REH (red) and RS4;11 cells (Yellow). (D and E) Correlation of replicate IGF2BP3 iCLIP experiments (D) Comparison of iCLIP from RS4;11 cells and (E) REH cells. Spearman's rank correlation coefficient (R) and p-value are given for each comparison.
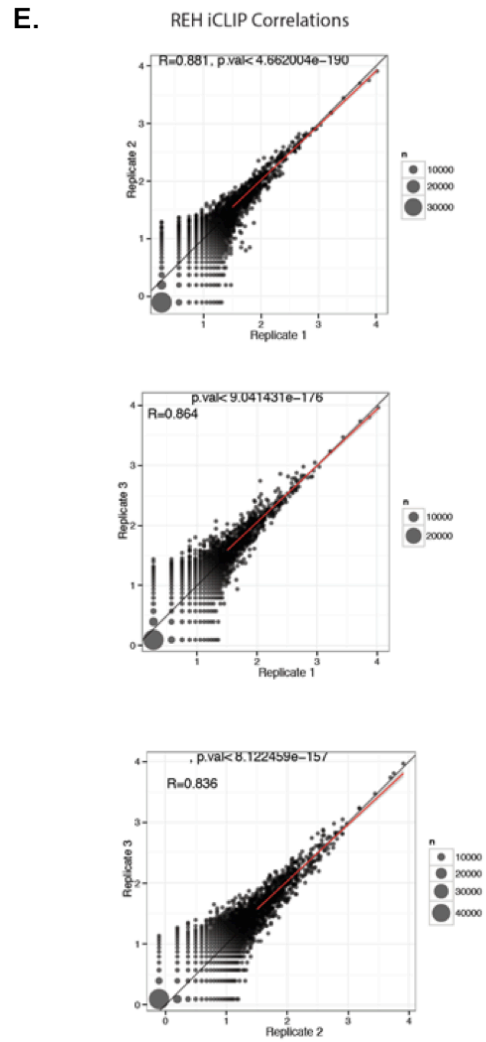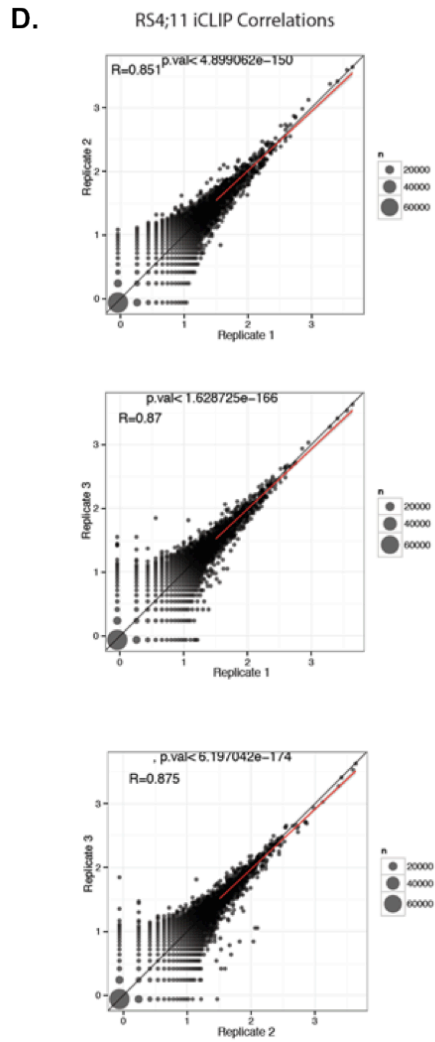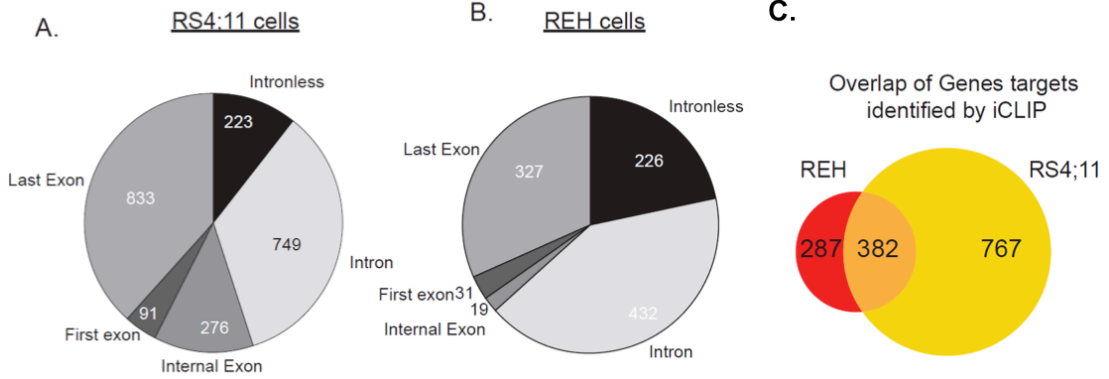
Figure S7: Expression of putative targets of IGF2BP3 (A and B) mRNA expression of Cdk6 and Myc in mouse bone marrow after IGF2BP3 over expression (t test, p=0.0009), (C) qPCR confirmation of overexpression of hIGF2BP3 and (D) mIGF2BP3 in mouse 7OZ/3 cells. (E) qPCR of Cdk6 in 7OZ/3 after IGF2BP3 overexpression shows no significant difference. Error bars denote S.D.
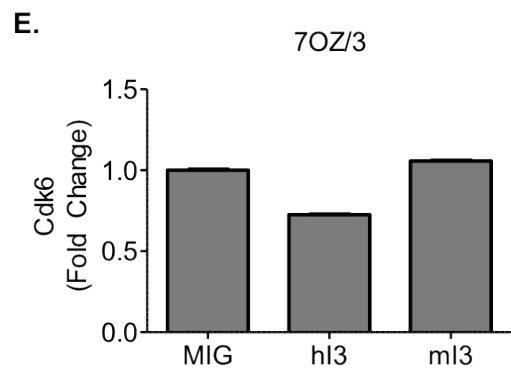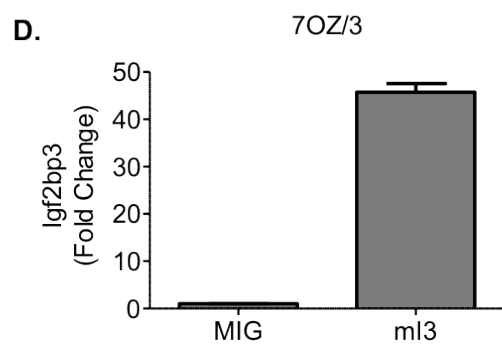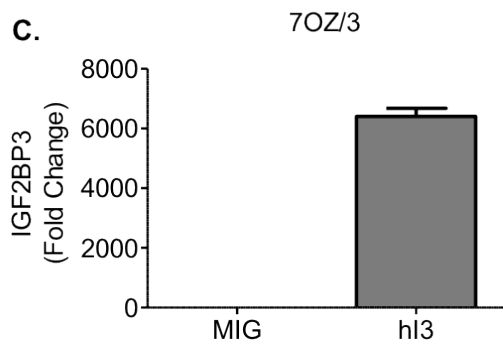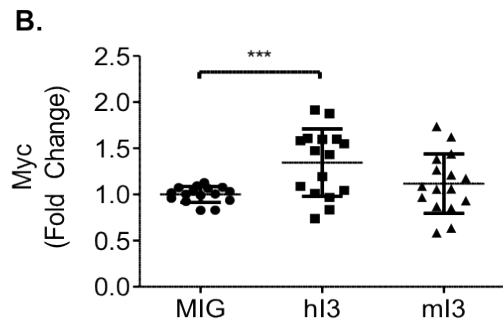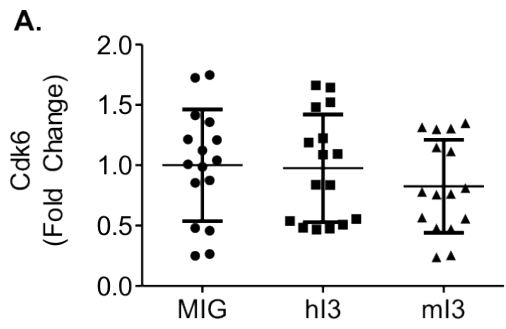
Figure S8: Cross-validation of IGF2BP3 iCLIP targets with IGF2BP3-sensitive differentially-expressed genes in RS4;11 B-ALL cells. (A) Volcano plot of differentially-expressed genes (blue dots) determined using DESeq analysis (Anders and Huber. Genome Biol. 2010) on RNA-seq samples from control and constitutive knock down for IGF2BP3 in RS4;11 cells (as described in Figure 2). Those differentially-expressed genes identified as IGF2BP3 targets by iCLIP have been highlighted (orange dots). Dots demarcated by black outlines are those genes determined to be associated with leukemogenesis by gene ontology analysis of OMIM(online Mendelian inheritance of Man)-associated disease pathways. Dotted lines represent cutoffs of -/+ 1.5 fold-change in expression (vertical lines) and p-value < 0.05 cutoff (horizontal line). (B,C) Gene Ontology analysis of gene subgroups showing increased expression (B) and decrease expression (C) with IGF2BP3 knockdown using Enrichr gene list enrichment analysis webtool (Chen et al, BMC Bioinformatics. 2013). Term lists used in this analysis were GO_Biological_Processes and KEGG (Kyoto Encyclopedia of Genes and Genomes) to determine enriched processes and pathways from our cross-validated list of 269 IGF2BP3-targeted and –sensitive genes. Vertical dotted lines represent p-value cutoff (p < 0.05).

A. Overlap of IGF2BP3 iCLIP targets and Differentially Expressed RNA transcripts in RS4;11 Cells

B. RNA-seq/iCLIP overlap in Genes showing increased expression w/ IGF2BP3 KD
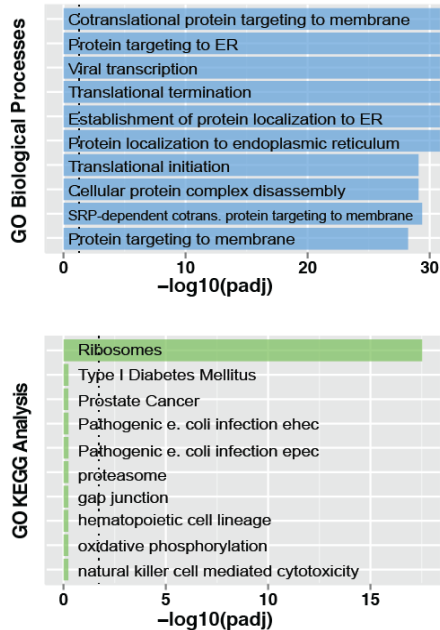
C. RNA-seq/iCLIP overlap in Genes showing decreased expression w/ IGF2BP3 KD
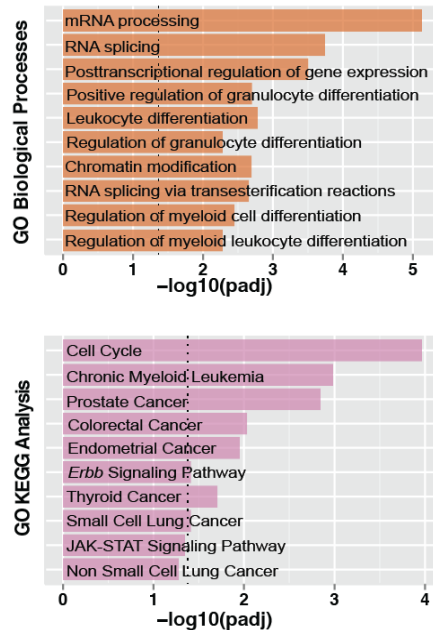
Figure S9: RNA immunoprecipitation (RIP) validation of MYC and CDK6 association with IGF2BP3 from RS4;11 cells. (A)Western blot of protein samples from IGF2BP3 RIP. Input refers to RS4;11 cell lysate used for immunprecipitation. FT is flowthrough of immunoprecipitation from either control (mouse IgG) or IGF2BP3 RIP. RIP is RNA immunoprecipitation from control (mouse IgG, Jackson Laboratories) or α-IGF2BP3 antibody (Santa Cruz Biotechnology, D-7). (B, C) Scatter bar plots comparing the fold-enrichment for MYC (I; n=4) and CDK6 (J; n=3) in control (mouse IgG) and α-IGF2BP3 antibody RNA immunoprecipitations. Levels of MYC and CDK6 are normalized to input levels from total RNA with 18s rRNA as reference. (t test, * = p< 0.05, ** = p < 0.01).

A.

B. MYC

C. CDK6
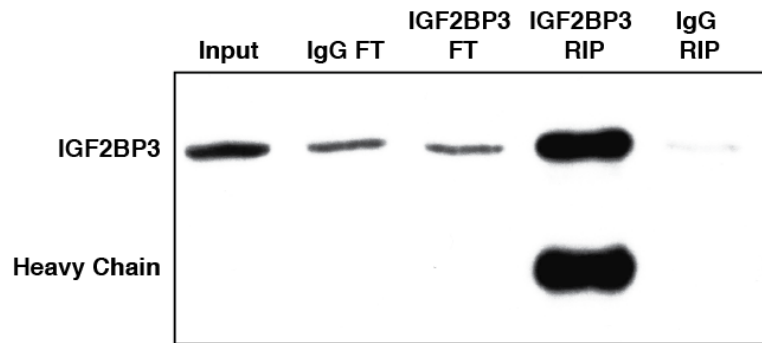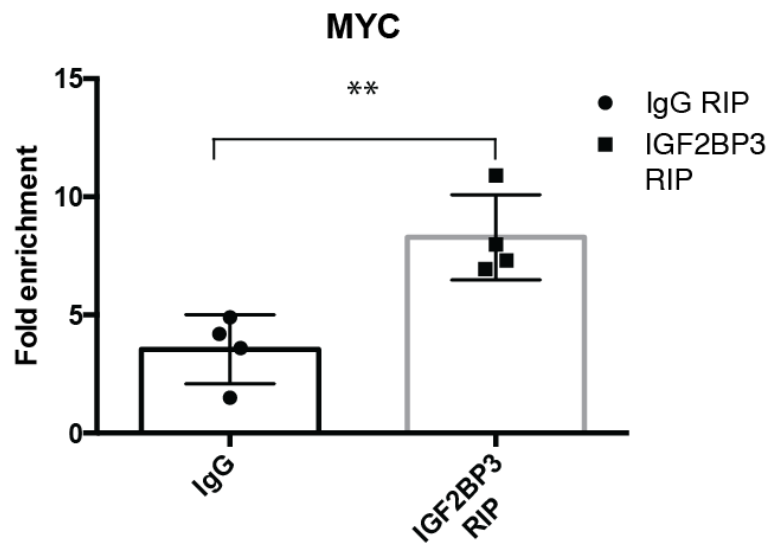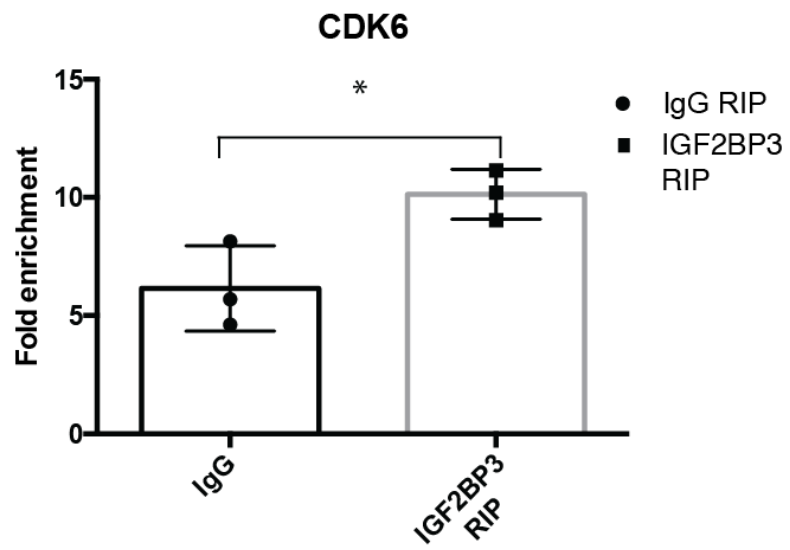
Chapter 6

Loss of exon identity is a common mechanism of human

inherited disease

Loss of exon identity is a common mechanism of human inherited disease

Timothy Sterne-Weiler,[1,2] Jonathan Howard,[1] Matthew Mort,[3] David N. Cooper,[3] and Jeremy R. Sanford[1,4]

[1]Department of Molecular, Cellular and Developmental Biology, University of California Santa Cruz, Santa Cruz, California 95064, USA; [2]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA; [3]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, United Kingdom

It is widely accepted that at least 10% of all mutations causing human inherited disease disrupt splice-site consensus sequences. In contrast to splice-site mutations, the role of auxiliary *cis*-acting elements such as exonic splicing enhancers (ESE) and exonic splicing silencers (ESS) in human inherited disease is still poorly understood. Here we use a top-down approach to determine rates of loss or gain of known human exonic splicing regulatory (ESR) sequences associated with either disease-causing mutations or putatively neutral single nucleotide polymorphisms (SNPs). We observe significant enrichment toward loss of ESEs and gain of ESSs among inherited disease-causing variants relative to neutral polymorphisms, indicating that exon skipping may play a prominent role in aberrant gene regulation. Both computational and biochemical approaches underscore the relevance of exonic splicing enhancer loss and silencer gain in inherited disease. Additionally, we provide direct evidence that both SRp20 (*SRSF3*) and possibly PTB (*PTBPI*) are involved in the function of a splicing silencer that is created de novo by a total of 83 different inherited disease mutations in 67 different disease genes. Taken together, we find that ~25% (7154/27,681) of known mis-sense and nonsense disease-causing mutations alter functional splicing signals within exons, suggesting a much more widespread role for aberrant mRNA processing in causing human inherited disease than has hitherto been appreciated.

[Supplemental material is available for this article.]

The sequences of mammalian exons perform at least two overlapping roles in gene expression. First, exons are encoded with the primary sequence determinants of proteins. This information is decoded by the ribosome and translated into functional polypeptides. Secondly, it has been understood for some time that exonic sequences also contribute to pre-mRNA splicing through both sequence and local structure (Watakabe et al. 1993; Wang and Cooper 2007; Warf and Berglund 2010). These latter observations are not surprising given the organization of mammalian genes, which typically contain small exons (~140 bp) flanked by thousands of base pairs of intronic DNA sequence. The large size of many mammalian genes and the apparent degeneracy of mammalian splice sites marking the 5′ and 3′ termini of introns are also suggestive of a requirement for auxiliary *cis*-acting elements in facilitating exon recognition (Keren et al. 2010).

Exonic sequences contain a staggering array of *cis*-acting elements that direct the activation or repression of splicing (Liu et al. 1998; Fairbrother et al. 2002; Cartegni et al. 2003; Wang et al. 2004; Yeo et al. 2004; Shi et al. 2005; Goren et al. 2006). Typically, these functional elements are classified as either exonic splicing enhancers (ESE) or exonic splicing silencers (ESS) based on their ability to stimulate or inhibit splicing, respectively. ESE and ESS elements, acting in concert with their cognate *trans*-acting RNA-binding proteins, represent important components in a splicing code that specifies how, where, and when mRNAs are assembled from their precursors (Barash et al. 2010). Two of the major players in establishing exon identity are the serine- and arginine-rich proteins (SR proteins) and the heterogeneous nuclear ribonucleoproteins (hnRNPs) (for review, see Wang and Burge 2008). SR proteins promote the initial stages of spliceosome assembly by binding to ESEs and recruiting basal splicing factors to adjacent splice sites or by antagonizing the effects of ESS elements (Kohtz et al. 1994; Graveley et al. 2001; Zhu et al. 2001). In contrast, hnRNPs mediate the repressive effects of silencers and can alter recruitment of the core splicing machinery (Wang et al. 2006; Yu et al. 2008). The interactions between silencers, enhancers, and their cognate binding proteins play a critical role in the fidelity and regulation of pre-mRNA splicing (Eperon et al. 2000; Zhu et al. 2001).

At least 10% of all mutations identified as causing human inherited disease are known to alter consensus 5′- or 3′-splice sites, thereby inducing aberrant pre-mRNA splicing (Krawczak et al. 2007). Nonetheless, the role(s) played by pre-mRNA splicing in human genetic disease remain enigmatic (Cooper et al. 2009). Although the mechanistic consequences of mutations on splice sites are fairly easy to interpret, evaluating precisely how inherited disease-causing mutations influence the loss or gain of ESE/ESS motifs is much more challenging (Cartegni and Krainer 2002; Pagani and Baralle 2004). This is due in part to the considerable functional overlap between protein-coding sequences and the *cis*-acting elements involved in splicing regulation. Hence, many mis-sense and nonsense mutations that alter pre-mRNA splicing may be incorrectly assumed to have an impact solely on protein structure–function relationships as a consequence of amino acid substitution or protein truncation, rather than on splicing changes per se (Liu et al. 2001; Pagani and Baralle 2004). It is also possible that the impact of a disease allele may be due to the combination of an aberrant splicing event and the presence of a normal-length mutation-bearing transcript. Such multifunc-

tional sites within coding regions have recently been identified by the intragenic mapping of common genetic variants known as single nucleotide polymorphisms (SNPs) (Majewski and Ott 2002; Fairbrother et al. 2004; Goren et al. 2008). As a consequence of purifying selection, SNPs appear somewhat depleted and synonymous codon bias restricted (GAA vs. GAG), revealing a silhouette of the "splicing code" that appears position-restricted relative to the edges of exons (Majewski and Ott 2002; Fairbrother et al. 2004; Chamary et al. 2006). Here we have investigated the relationship between coding sequence mutations and splicing regulation using a novel combination of bioinformatic and biochemical techniques.

Results

Disease-causing mutations overlap with the splicing code

We extracted 27,681 exonic disease-causing (mis-sense and nonsense) (see Table 1) mutations from the Human Gene Mutation Database (HGMD; http://www.hgmd.org), a proprietary, hand-curated database requiring one or more pieces of causal evidence for inclusion (e.g., absence from normal controls, cosegregation of lesion and phenotype through pedigree, independent occurrence in an unrelated patient, etc.) (Stenson et al. 2008). For common genetic variants, we extracted 8601 exonic single nucleotide polymorphisms from the 1000 Genomes Project (http://www.1000genomes.org) (see Table 2; Durbin et al. 2010). These exonic SNPs were selected for neutrality by filtering average heterozygosity to 30%–50%, corresponding to a Hardy-Weinberg minor allele frequency of at least ~0.18. In addition, we determined the ancestral allele (biallelic directionality) by comparison to the chimpanzee (*Pan troglodytes*) genome (Fairbrother et al. 2004; Karolchik et al. 2008).

We used a set of 238 hexameric sequences corresponding to the RESCUE-ESE data set and 176 hexameric sequences corresponding to the FAS-hex2 ESS data set (Fairbrother et al. 2002; Wang et al. 2004). Each set of hexanucleotides was experimentally validated to enhance or silence splicing of an alternative exon in a minigene context. We used the directionality of the substitutions, based on ancestral > variant for SNPs and wild-type > disease for HGMD mutations to calculated odds ratios (OR), expressing the relative likelihoods that either disease-causing mutations or the putatively neutral polymorphisms are associated with the loss or gain of ESEs or ESSs (see Methods). Whereas disruption of ESEs was found to be strongly associated with the mutations from the HGMD data set by comparison with neutral SNPs, there was substantially less evidence for the gain of ESEs in the disease mutation data set (Fig. 1A). In contrast, a strong association was noted between disease-causing mutations and the creation of ESS motifs (Fig. 1B). Taken together, these data suggest that exon skipping may play a key role in human inherited disease not only via the loss of exonic splicing enhancers but also via the gain of exonic splicing silencers.

Table 1. Summary of putative splicing–sensitive mutations, exons, and genes associated with genetic disease

	Disease-mutation count	Exon count	Gene count
3–72 bp from SS	27,681	7974	1760
ESR loss or gain	14,608	5743	1431
Statistically significant (5% FDR)	7154	3747	1055

Table 2. Summary of single nucleotide polymorphisms used in this study

	SNP count	Exon count	Gene count
>3 bp from SS	8438	7338	4529
ESR loss or gain	4248	3886	2866

Disease-associated alterations of the splicing code

To determine if specific *cis*-acting elements are more susceptible to disease-causing mutations than others relative to a background level ascertained by reference to putatively neutral SNPs, we calculated the binomial enrichment *P*-value for loss or gain of individual hexamer sequences (see Methods). We visualized the distribution of genomic variants across individual hexamers from the ESE and ESS data sets using Principal Component Analysis for optimal leaf ordering (Rajaram and Oono 2010). Figure 1, C and D, depicts $\log_2$ ratios of HGMD mutations versus SNPs for loss or gain of individual ESE and ESS hexamers with a binomial *P*-value significant at a 5% false discovery rate (FDR). Of the 238 ESE hexamers considered in this analysis, 106 showed no significant difference for either loss or gain by inherited disease-causing mutations relative to SNPs given the 5% FDR. Similarly, 67 out of 176 ESS hexamers were not significantly different between the HGMD and SNP data sets. For both ESEs and ESSs, the heat maps clearly demonstrate that HGMD mutations are not uniformly distributed across all hexamers but, rather, are enriched in select subsets corresponding to losses or gains. For ESEs, we not only observe clusters of hexamers that are both exclusively either ablated or created by disease-causing mutations (Fig. 1C, regions i and iii) but also a small subset of hexamers that are subject to a significant degree of both loss and gain by disease-causing mutations (Fig. 1C, region ii). In contrast to the ESEs, a much larger number of hexamers are significantly enriched for disease-causing mutations that create ESSs rather than abolish ESSs (Fig. 1D, cf. regions i and ii). An expanded version of Figure 1, C and D, containing the hexamer sequences, is presented in Supplemental Figures 1 and 2. We also examined loss or gain of each hexamer sequence in several different contexts including their proximity to the nearest 5′ or 3′ splice sites and their presence within alternative or constitutive exons (see Methods). Supplemental Figures 3–6 show that, although the general observations described in Figure 1, C and D hold true, there is inconclusive evidence for a bias of hexamer loss or gain relative to either splice site or between constitutive or alternative exons.

ESEs ablated by disease-causing mutations share hallmarks of functional splicing enhancers

Since evolutionary conservation usually implies functionality (Boffelli et al. 2003; Margulies et al. 2003; Siepel et al. 2005), we opted to determine whether there was a difference in average evolutionary conservation between those ESE hexamers lost as a consequence of disease-associated mutation and those lost as a result of the introduction of a neutral SNP allele. Because ESEs are more abundant in the vicinity of splice sites and the activity of splicing enhancers decreases with increasing distance from splice sites (Graveley and Maniatis 1998; Yeo and Burge 2004; Parmley et al. 2006), we evaluated average phyloP scores across alignments of 46 placental mammals (Pollard et al. 2009) for HGMD- or SNP-disrupted ESE hexamers relative to their positions within exons. Figure 2 shows
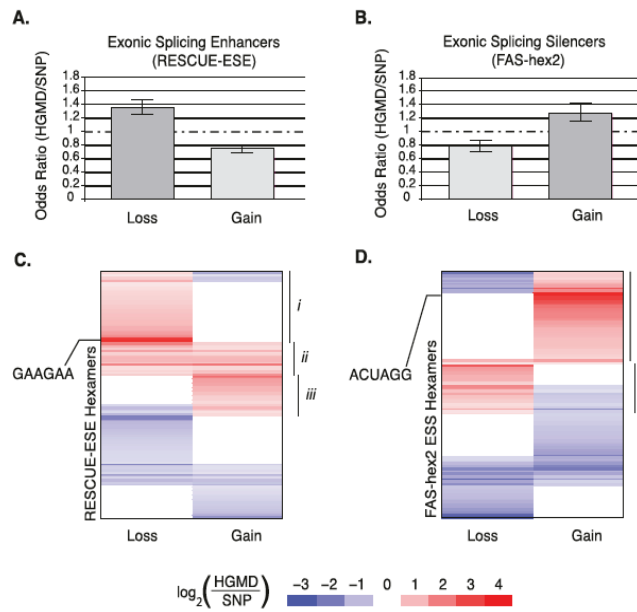
Figure 1. Patterns of exonic splicing regulator loss or gain among pathological mutations (HGMD) as compared to putatively neutral SNPs. (*A,B*) Bar height corresponds to the odds ratio (OR) of HGMD/SNPs for the loss or gain of enhancers and silencers, respectively. Each error bar represents a two-tailed 95% confidence interval for the bar height (see Methods). Directionality was expressed in the form of the ancestral state > variant for the SNPs and healthy > disease for the HGMD mutations. (*A*) Hexamers corresponding to exonic splicing enhancers were obtained from the RESCUE-ESE database. Each hexamer was scored for the loss or gain (de novo creation) of an ESE by the inherited disease-causing mutations (relative to the wild-type allele) or putatively neutral SNPs (relative to the ancestral allele). (*B*) Hexamers corresponding to exonic splicing silencers were obtained from the FAS-hex2 database and scored for loss or gain as described in *A*. (*C,D*) Principal component analysis (PCA) of normalized ratios of HGMD versus SNP substitution for loss or gain of ESE and ESS hexamers, respectively. Each row corresponds to a single ESE or ESS hexamer, whereas each column represents loss or gain of the hexamer by a genomic variant. Any hexamers that were not significant at the 5% level were omitted from the heat map. Each box depicts the log ratio for the counts of HGMD/SNP causing loss or gain of a specific hexamer. A positive log ratio in red corresponds to a hexamer in a certain context (column) that is significantly enriched in inherited disease. Alternatively, a blue value represents a hexamer that is polymorphic across human populations. White boxes correspond to non-significant *P*-values given a false discovery rate (FDR) of 5%. (*C*) Hexamer clusters corresponding to ESE-loss (region i), ESE-loss and ESE-gain (region ii), and ESE-gain (region iii). Hexamer clusters corresponding to ESS-gain (region i) and ESS-loss (region ii). The loss/gain of SRSF1-like binding sites is indicated by GAAGAA in C, whereas the ACUAGG hexamer is indicated in *D*.

the bivariate density distributions of ESE hexamers lost by disease-causing mutations and neutral SNPs (Fig. 2, left and right panels, respectively). In both density plots, the axes displaying phyloP scores are marked with red lines corresponding to the statistical threshold for evolutionary conservation (phyloP score >1.3, α = 0.05) and blue lines corresponding to the median phyloP scores for ESE hexamers. The median phyloP score corresponding to the distribution of putative ESE hexamers ablated by disease-causing mutations is 1.42 (Fig. 2, left panel, blue line), easily exceeding the statistical threshold for evolutionary conservation. In contrast, ESE hexamers abolished by putatively neutral SNPs have much lower distributed average phyloP scores, such that the median (1.02) is well below the statistical threshold to reject the null hypothesis of neutrally evolving sequence. Although the median distance of ESEs from splice sites is not significantly different for those disrupted by SNPs or HGMD mutations, the distribution of phyloP scores for ESEs lost by disease-causing mutations shifts toward higher values approaching splice sites. In

contrast, the distribution for ESEs disrupted by neutral SNPs is visibly shifted toward lower values near the edges of exons. Overall, the density plots in Figure 2 indicate that ESEs targeted by disease-causing mutations exhibit a bias not only toward higher conservation values, but also with respect to a location toward the edges of exons as compared to those ESEs targeted by neutral SNPs.

Disease-causing mutations often affect conserved regions of proteins (Kumar et al. 2009). To determine if the conservation levels observed for ESE hexamers ablated by disease-causing mutations could result as a byproduct of this bias, we sampled random hexamers that did not cause loss/gain of any ESR from HGMD mutation- or SNP-containing exons as well as another subset that encompassed HGMD mutations (Supplemental Fig. 7). As expected, both random hexamers sampled from HGMD exons and those containing HGMD mutations displayed lower distributed evolutionary conservation values than the ESEs lost by HGMD mutations (median average phyloP scores of 1.32 and 1.36 compared to 1.45, respectively; Welch test two-tailed *P*-value < 1.79×10^{-39} and 3.81×10^{-15}). Furthermore, the phyloP scores for ESEs targeted by neutral SNPs were distributed lower overall than the set of random hexamers sampled from SNP-targeted exons (median average phyloP scores of 1.09 and 1.15, respectively; Welch test two-tailed *P*-value < 0.039).

Functional validation of a splicing silencer mutationally linked to 67 different disease genes

To test the hypothesis that those exonic splicing silencers that harbor a preponderance of disease alleles could represent functionally repressive elements, we opted to validate the activity of one of the most significant hexamers identified by our comparison of disease-causing and neutral polymorphisms within ESSs (designated in Fig. 1D, ACUAGG, binomial *P*-value < 2.2×10^{-16}; Supplemental Table 4). This specific hexamer appears to have been created by a total of 83 different disease-causing mutations in 67 different genes. We searched this list of disease-causing mutations for sequences that were amenable for cloning into splicing reporter constructs and were present in exons of near average size and splice-site strength (Yeo and Burge 2004). Of the three different disease-causing mutations we selected—*OPA1*, *PYGM*, and *TFR2*—there was no a priori evidence for any effect on splicing from in vitro analysis (Bruno et al. 1999; Camaschella et al. 2000; Schimpf et al. 2008). However, aberrant splicing of *OPA1*, *PYGM*, and *TFR2* is observed in patients carrying coding and non-coding mutations at other positions in these genes (Schimpf et al. 2006; Biasiotto et al. 2008; Nogales-Gadea et al. 2008).
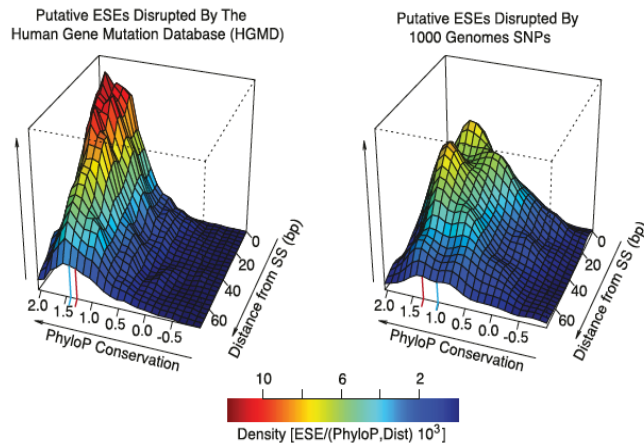
Figure 2. Conservation of exonic splicing enhancers ablated by genomic variants. The two-dimensional density distributions (relative values given in color scale) of ESEs containing associated average phyloP (Pollard et al. 2009) scores and distances to the nearest splice site (3–72 bp). The density distributions for ESEs targeted for loss by inherited disease-causing (HGMD) mutations (*left* panel) or neutral SNPs (*right* panel). In each panel the red line designates a phyloP score corresponding to a *P*-value of 0.05. The blue line designates the median phyloP score of each density distribution.

We created matched pairs of beta-hemoglobin-based (*HBB1*) splicing reporter gene constructs containing the wild-type or mutant exon plus 50 bp of adjacent intron sequence (Rothrock et al. 2003). To investigate the effects of the ACUAGG silencer on splicing of the reporter genes, HeLa cells were transiently transfected with both wild-type or mutant constructs. Because inclusion of all three of the test exons is predicted to induce nonsense-mediated decay (NMD) by inducing an in-frame premature termination codon (PTC) (see Supplemental Fig. 8), we assayed splicing in the presence of the translation inhibitor emetine dihydrochloride, a potent inhibitor of NMD in vivo (Noensie and Dietz 2001). After RNA isolation and conversion to cDNA, each sample was tested for reporter RNA splicing efficiency. Inhibition of NMD was confirmed by assaying the splicing of the *SRSF6* pre-mRNA, an endogenous PTC-containing gene known to undergo NMD (Lareau et al. 2007; Ni et al. 2007). The presence of the *SRSF6* poison exon-containing mRNA shows that NMD was, indeed, inhibited in the emetine-positive samples (Supplemental Fig. 9). As shown in Figure 3B, introduction of the ACUAGG hexamer resulted in a remarkable degree of exon skipping in the *OPA1*, *PYGM*, and *TFR2* reporters. Quantification of amplicons using an Agilent 2100 Bioanalyzer demonstrated that the ACUAGG silencer significantly decreased inclusion of the *OPA1*, *PYGM*, and *TFR2* test exons from 97% to 44% (*P*-value $< 2.39 \times 10^{-3}$), 62%–19% (*P*-value $< 6.83 \times 10^{-4}$), and 86%–49% (*P*-value $< 7.62 \times 10^{-4}$), respectively (Fig. 3B). These data suggest that it is possible to predict splicing-relevant mutations based on the statistical enrichment of hexamers in disease-associated mutation data sets.

Identification of *trans*-acting splicing silencers

The data presented in Figure 4 suggest that the ACUAGG motif functions as a strong splicing silencer. Splicing silencers have been shown to interact with *trans*-acting factors such as hnRNPs and to alter the kinetics of the non-rate-limiting steps of spliceosome assembly when two 5'-splice sites are in competition (Zhu et al. 2001; Yu et al. 2008). Given that we did not observe activation of cryptic

5'-splice sites in the mutant *OPA1* construct, we searched for potential *trans*-acting factors by RNA affinity chromatography using HeLa nuclear extracts. RNA-binding proteins captured by the wild-type and mutant RNA ligands at low and high stringency were identified by Multidimensional Protein Identification Technology (MudPIT) (Supplemental Fig. 10; Supplemental Tables 5–7). At both high and low concentrations of KCl, peptides corresponding to SRp20 (*SRSF3*), PTB (*PTBP1*), hnRNP D (*HNRNPD*), and hnRNP L (*HNRNPL*) were present on the mutant RNA ligand (Supplemental Fig. 10A,B; Supplemental Tables 5, 6), suggesting that these proteins may play a role in mediating the silencing activity of the ACUAGG hexamer.

To test the role of SRp20 and PTB in splicing silencing, we cotransfected HeLa cells with either the wild-type or mutant *OPA1* splicing reporter and siRNA targeting SRp20, PTB, or a non-targeting duplex. In cells transfected with non-targeting control siRNA, the mutant *OPA1* reporter was inefficiently spliced relative to the wild-type reporter (Fig. 4A, cf. lanes 1 and 4). In contrast, depletion of SRp20 and, to a lesser extent, PTB partially rescued inclusion of the mutant *OPA1* exon (Fig. 4A, cf. lane 4 with lanes 5 and 6). Quantification of the RT-PCR amplicons from duplicate experiments revealed that depletion of SRp20 and PTB restored inclusion of the mutant exon to ~50% and ~25% of the wild-type levels, respectively. Analysis of the exon inclusion ratios for the SRp20 and PTB depletion revealed that only knockdown of SRp20 resulted in statistically significant changes relative to the control (Fig. 4B). Depletion of both SRp20 and PTB was confirmed by fluorescent Western blot analysis of nuclear extracts prepared from transfected cells (Fig. 4B). Quantification of the Western blots revealed approximately twofold and 2.75-fold depletions of SRp20 and PTB, respectively, relative to cells transfected with non-targeting control duplex. Taken together, these data implicate SRp20 and PTB in both the recognition and function of the ACUAGG exonic splicing silencer motif. We did not test the role(s) of hnRNP D or hnRNP L in the function of this silencer motif.

Potential nonsense sequences are enriched in ESR hexamers

Nonsense-associated altered splicing (NAS) describes the phenomenon whereby exons encoding premature stop codons tend to be excluded from the mature RNA transcript during pre-mRNA splicing (Dietz et al. 1993; Li et al. 2002; Wachtel et al. 2004). Although the primary mechanism of NAS is still unknown, several different models have been proposed. These include a nuclear scanning model that invokes the action of a frame-sensitive mechanism in pre-mRNA splicing (Wang et al. 2002). Alternatively, NAS may be the direct result of ESE disruption as a means to abolish exon recognition (Shiga et al. 1997; Liu et al. 2001). In order for nonsense mutations to be specifically associated with the loss/gain of ESR sequences, there must be a sequence bias in the ESRs themselves. To investigate this hypothesis for ESR loss, we simulated mutations based on the transition/transversion rates observed for the 14,771 exonic HGMD mutations located near the

A.

B.
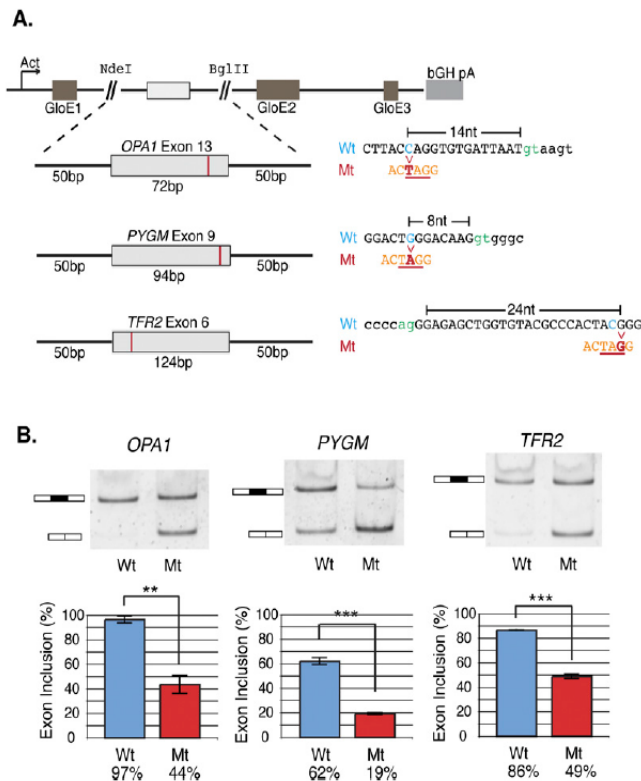
Figure 3. Validation of mutations creating the enriched silencer ACUAGG using the beta-globin splicing reporter. (*A*) Splicing reporter constructs created from matched pairs of wild-type (Wt) or mutant (Mt) alleles that give rise to a gain of the ACUAGG silencer in constitutive exons in three different disease genes: *OPA1*, *PYGM*, and *TFR2*. GloE1, GloE2, and GloE3 designate exons 1–3 of beta-globin. The polyadenylation signal from the bovine growth hormone 1 gene is indicated by bGH pA. (Blue) Wild-type allele; (red) the mutant; (orange) the silencer sequence created by the mutation. (*B*) HeLa cells were transiently transfected in triplicate with both wild-type (Wt) and mutant (Mt) alleles. Twenty-four hours after transfection, cells were treated with emetine to inhibit NMD, RNA was harvested, and the splicing efficiency was determined by RT-PCR and visualized using 6% non-denaturing (29:1) polyacrylamide gel electrophoresis (PAGE). The graphs depict mean exon inclusion quantified using an Agilent 2100 Bioanalyzer with standard error bars (see Methods). Statistical hypothesis testing on means was executed using a Welch *t*-test for normal data with unequal sample size and variance using α-values of (*) 0.05, (**) 0.01, and (***) 0.001.

abundance of enhancers within exons (Fairbrother et al. 2002) and their being subject to protein-coding restrictions. These data therefore support a model that involves the disruption of enhancers and the creation of silencers to yield nonsense-associated altered splicing (NAS). Consistent with this postulate, it would appear as though enhancer loss and silencer gain are specifically associated with potential-nonsense codons through the sequence bias of the ESRs.

Discussion

The results presented here demonstrate that nearly ~25% (7154/27,681) of exonic (i.e., mis-sense and nonsense) mutations that cause human inherited disease are likely to induce exon skipping either via the loss of evolutionarily conserved splicing enhancers or alternatively through the creation of potent splicing silencers (Table 2). Given that it has already been recognized that at least 10% of disease-causing mutations ablate 5'- or 3'-splice site consensus sequences (Krawczak et al. 2007), we conservatively estimate that approximately one-third of disease-causing mutations may induce aberrant splicing. Recently published work from another group using an independent strategy reached a similar estimate (22%) of splicing sensitive mis-sense/nonsense disease-causing mutations (Lim et al. 2011). Future studies that include mutations that affect intronic *cis*-elements may well increase this proportion. Overall, our results provide new insights into the underlying mechanisms that link mutation-induced aberrant splicing and human inherited disease. Understanding these mechanisms is a prerequisite for the optimization of treatment regimens as we enter the era of personalized medicine.

One surprising result from our study is that although genomic variants that create ESEs or abolish ESSs are more frequently associated with neutral SNPs (Fig. 1A,B), some individual ESE and ESS hexamers show a remarkable enrichment for disease-causing mutations when gained or lost, respectively. We believe that this class of mutations may induce aberrant splicing of adjacent exons as previously described for a polymorphism in the *MST1R1* gene (Ghigna et al. 2005). We find it interesting that specific ESR hexamers, based on their HGMD/SNP log ratios, appear to be disproportionately represented by disease-causing mutations (Fig. 1C, region i; Fig. 1D, region i). Within each of these clusters there are individual hexamers that appear to be mutated very frequently in genetic disease, suggesting that specific *trans*-acting factors may be associated with several genetic disorders. For example, one of the sequences in the enhancer loss-enriched cluster displays a remarkable degree of similarity to the canonical binding site for the splicing factor SF2/ASF (*SRSF1*) (Fig. 1C, GAAGAA; Tacke and

edges of exons (Supplemental Fig. 11). For the ESR gains, we simply evaluated the proportion of nonsense 3-mers (UAG, UAA, UGA) compared to all of the 3-mers within the corresponding hexamers. As a control for the experiment, we used the same algorithm to evaluate the "loss" or "gain" of all 3-mers (excluding the first 3 bp) in a previously used set of 206,029 human internal exons (Fig. 5; Fairbrother et al. 2004). Using these data, we compared the nonsense potential of exon retention to exon skipping with respect to that of our control. For all ESR hexamers in our data sets, we observed at least an approximately twofold increase in nonsense potential, consistent with silencer gains (*P*-value < 5.50×10^{-21}) and enhancer losses (*P*-value < 1.27×10^{-14}) when compared to controls (χ^2 goodness-of-fit test) (Fig. 5). As expected, minimal values of nonsense potential were observed for silencer loss (*P*-value < 0.19) and enhancer gain (*P*-value < 0.86), consistent with the nonsense potential seen for the respective "loss" and "gain" of all human exonic 3-mers (χ^2 goodness-of-fit test) (Fig. 5). The lack of nonsense potential for enhancer gain is not surprising given the
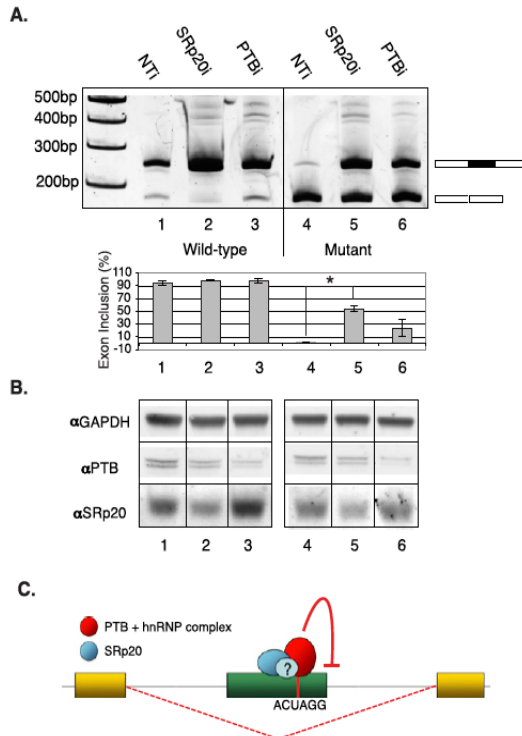
Figure 4. Identification of *trans*-acting factors implicated in skipping of the ACUAGG-containing *OPA1* allele. (*A*) RT-PCR analysis of *OPA1* splicing reporters from HeLa cells cotransfected with non-targeting siRNA (NTi), *SRSF3* siRNA (SRp20i), *PTBP1* siRNA (PTBi). Lanes *1–3* and *4–6*, wild-type and mutant reporters, respectively. Statistical hypothesis testing on means was executed using a Welch *t*-test for normal data with unequal sample size and variance using α-values of (*) 0.05, (**) 0.01, and (***) 0.001. (*B*) Western blot showing relative depletion of SRp20 and PTB as compared to the GAPDH loading control. (*C*) Model for aberrant splicing by "ACUAGG" ESS. A point mutation creating the sequence ACUAGG results in recruitment of a silencer complex that may contain SRp20 and members of the hnRNP protein family, either directly or indirectly bound to the RNA sequence. The complex is involved in deterring inclusion of the mutant exon via mechanism(s) that still remain to be determined.

Manley 1995; Sanford et al. 2009). Indeed, the presence of an SF2/ASF consensus motif in this cluster supports previous evidence for the loss of ESEs as an important cause of human inherited disease (Sanford et al. 2009). Finally, there are a striking number of ESEs and ESSs that are relatively untouched by disease-causing mutations but appear to be more polymorphic in different human populations. These data suggest at least two non-mutually exclusive possibilities. The first is that those hexamers that are overrepresented in the SNP data set may be redundant with the function of other hexamers and hence more prone to variation across human populations. Alternatively, the polymorphic ESRs identified here may be associated with allele-specific alternative splicing that confers a gain of fitness rather than a disease phenotype (Fraser and Xie 2009). Each of these hypotheses will require further testing.

As described above, we investigated the function of a specific ESS hexamer, ACUAGG (Fig. 2D), that has been created de novo by no fewer than 83 different single nucleotide substitutions (both mis-sense and nonsense) in 67 different genes as a cause of hu-

man inherited disease (see Supplemental Table 4). Reporter constructs derived from three different disease genes—*OPA1*, *TFR2*, and *PYGM*—all demonstrated that ACUAGG promotes skipping of test exons derived from the mutant alleles of each gene. We also prepared a reporter construct corresponding to an ACUAGG introduction near the 3′ss of exon 13 from *MYH7*. This mutation failed to induce appreciable skipping of the test exon (data not shown). For the case of ACUAGG insertion in *OPA1*, the effects of this ectopic silencer element on exon skipping appear to be mediated, at least in part, by SRp20 and possibly PTB. This is somewhat surprising since SR proteins are typically thought to promote exon inclusion by binding to splicing enhancers (Ram and Ast 2007). It is possible that interactions between SRp20, PTB, and other hnRNPs create an exon silencing complex that promotes exon skipping. A comprehensive analysis of PTB–RNA interactions identified many examples of alternative cassette exons that are skipped through the action of PTB-binding sites located near the 5′-splice site (Xue et al. 2009). Our data suggest that the ACUAGG hexamer is a potent splicing silencer that functions at both 5′- and 3′-splice sites.

The impact of premature termination (nonsense) codons on gene expression remains an important consideration in the elucidation of the pathogenic basis of disease-causing mutations. It is now well established that the NMD pathway plays a central role in preventing the accumulation and translation of nonsense-containing mRNA isoforms (Maquat 2004; McGlincy and Smith 2008). However, PTCs are also suggested to directly influence alternative splicing decisions (Wang et al. 2002; Wachtel et al. 2004). The most plausible model is that PTCs disrupt ESEs and induce exon skipping (Liu et al. 2001; Cartegni et al. 2002; Pagani et al. 2003; Zatkova et al. 2004). Our results presented in Figure 5 extend this model by suggesting that such a surveillance mechanism
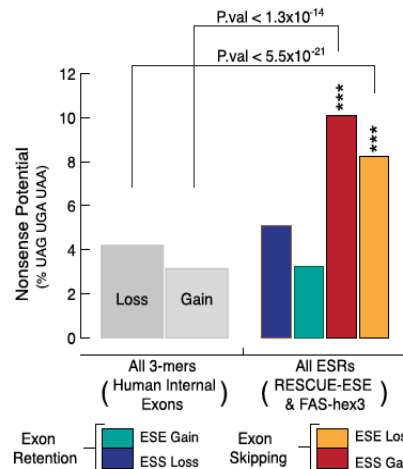
Figure 5. An overview of the nonsense codon sequence bias in exonic splicing regulators. Bars correspond to the nonsense-coding potential of ESR loss or gain, the proportion (expressed as a percentage) of 3-mers matching UAG, UGA, or UAA out of total 3-mers. For ESR loss, this was calculated via simulated mutation based on HGMD transition/transversion probabilities (Supplemental Fig. 2). For all human internal exonic 3-mers, the nonsense-coding potential was calculated using the same algorithm as the ESRs, except using a set of all human internal exonic sequences instead of ESR hexamers. The frequencies were normalized, and the values for the data given for ESR loss or gain were analyzed statistically (P-values from χ^2 goodness-of-fit test) using an α-value of (***) 0.001.

might evolve via the acquisition of ESR sequences to counteract PTC-containing exons associated with a greater likelihood of skipping. The apparent bias of ESR sequences toward potential nonsense codons would appear to be the most logical explanation for nonsense-associated altered splicing (Valentine and Heflich 1997). To test this postulate, we examined the very first observation of NAS where exon skipping was observed in the fibrillin (*FBN1*) gene due to a nonsense-causing T > G transversion 26 bp from a constitutive 3'-splice site (Dietz et al. 1993). Consistent with our model, the mutation appears to create a disease-enriched silencer CU<u>UAG</u>G (Supplemental Table 4, binomial *P*-value < 2.2 × 10^{-16}), with the core of the motif containing the previously observed nonsense codon, UAG. We suspect that many NAS observations may be consistent with this model due to ESR sequence bias or else attributable to PCR amplification artifacts after NMD (Cartegni et al. 2002). It remains to be determined if such a mechanism might arise as an attempt to preserve the transcript at the expense of a single exon or as a hammer to ensure that NMD is successfully elicited by the PTC.

Methods

Data set preparations

Mutations from the Human Gene Mutation Database (HGMD; http://www.hgmd.org) and SNPs (30%–50% heterozygosity) from the 1000 Genomes Project (http://www.1000genomes.org) were extracted and mapped to hg19 internal exons as annotated by the UCSC Known Gene track (Karolchik et al. 2008). Intersecting alleles found in both the HGMD and SNP data sets were removed from the SNP data set. Biallelic SNPs whose ancestral allele could not be determined were also removed from the SNP data set. Alleles mapping to within the first 3 bp from a splice site were removed from the SNP data set due to possible splice-site consensus sequence overlap. HGMD mutations were divided into subsets corresponding to the nearest splice site (5' or 3') and according to whether the mutation mapped to constitutive or alternative exons using the UCSC Alt Events track. As a quality control measure, HGMD mutations mapping past half the average HGMD exon length (144 bp) from a splice site were also removed, leaving only mutations within 3 to 72 bp from the nearest splice site.

Odds ratios and binomial estimation

ESE loss was defined as an event involving a directional change from one allele to another that served to convert an ESE to neutral or an ESS. ESE gain was defined as an event involving a change from either neutral or ESS to an ESE. We counted the numbers of HGMD mutations or SNPs causing ESE loss/gain. An odds ratio (OR), was calculated given

$$OR = \frac{\{P(\text{event}|\text{ HGMD})/[1 - P(\text{event}|\text{ HGMD})]\}}{\{P(\text{event}|\text{ SNP})/[1 - P(\text{event}|\text{ SNP})]\}},$$

where the event can be loss or gain of a given ESE hexamer and $P(\text{loss}|\text{data set}) = 1 - P(\text{gain}|\text{data set})$. Odds ratios are plotted as bars, with 95% confidence intervals (two-tailed) as error bars calculated using standard methods (Pagano and Gauvreau 2000). The same assumptions and calculations were used when considering loss or gain of ESS hexamers in neutral SNP and HGMD data sets.

To assess the significance of the enrichment of individual ESE hexamers in the HGMD data set as compared to the neutral SNP data set, we used the binomial distribution. For each hexamer i, the probability of $P(\text{HGMD}[i] = k)$ is distributed such that $\text{HGMD}[i] \sim \text{Bin}(n, p[i])$, where $\text{HGMD}[i]$ is a random variable corresponding to

the number of mutations causing loss of a particular ESE, $p[i]$ is the neutral background probability for the particular hexamer to be targeted for loss, and n is the total number of HGMD mutations causing loss of any ESE. The neutral background probability ($p[i]$) for each ESE hexamer i is calculated as the number of times a neutral SNP causes loss of i plus a pseudocount normalized over the total number of SNPs causing loss to ESEs ($\text{SNP}[i] + 0.5/\sum_i (\text{SNP}[i] + 0.5)$). Based on large values for n and exceptionally low values for $p[i]$, binomial *P*-values are approximated using the Poisson distribution such that $\lambda[i] \sim np[i]$. We also applied this to ESE gain and both ESS loss and ESS gain, each with their own set of neutral background probabilities. For any mutation that alters multiple ESR hexamers, only the hexamer with the lowest binomial *P*-value is used in statistical tests. The significance of each *P*-value is determined for multiple hypotheses using a Benjamini-Hochberg false discovery rate (FDR) of 5% (Benjamini and Hochberg 1995).

Conservation of ESE loss hexamers

To assess the evolutionary conservation of lost ESE motifs, we calculated the average phyloP score from multiple orthologous alignments of 46 placental mammals (Karolchik et al. 2008; Pollard et al. 2009) for each ESE hexamer ablated by a directional allele. Typically, phyloP scores are used to determine the conservation of individual sequence alignment columns between species, given a null model of neutral evolution at single nucleotide resolution. These scores in the human genome range from values as low as −13.79 to 2.94 representing the −$\log_{10}$(*P*-value) to reject the null hypothesis. In this study, the average phyloP score is used to determine the relative conservation of hexamers rather than as a strict statistical test. Using each ESE phyloP score and its corresponding distance to the nearest exon–intron boundary, we performed two-dimensional (2D) Gaussian kernel density estimation and plotted the three-dimensional (3D) density using R. To compare the distribution of phyloP scores for ESEs disrupted by HGMD mutations to that due to chance alone, we randomly sampled 13,000 hexamers from both the HGMD- and SNP-targeted exons that did not match a hexamer in the ESE data set. We also sampled an equal size of random hexamers containing HGMD mutations that did not cause loss or gain of known ESRs. Statistical hypothesis testing on means was executed using a Welch *t*-test for normal data with unequal sample size and variance using α-values of (*) 0.05, (**) 0.01, and (***) 0.001. Given such large sample sizes, normality assumptions are approximately satisfied through the asymptotic relationship to the normal distribution provided by the central limit theorem. Additionally, we performed a non-parametric alternative, the Wilcoxon rank-sum test, which provided similarly significant *P*-values for each test shown.

Splicing reporter assay and RNAi

To assess the functional relevance of non-synonymous HGMD mutations to splicing, DNA inserts containing the entire exon plus 50 bp of flanking intron sequence for both the matched wild-type and mutant versions of selected mutations were created using Custom Gene Synthesis from IDT (http://www.idtdna.com) flanked by NdeI and BglII restriction sites. Test alleles were subcloned from the pSMART vector using NdeI and BglII restriction sites into the pSC14mw vector. All constructs were validated by sequencing. Splicing reporters were transiently transfected into HeLa and 293T cells in six-well plates using Lipofectamine 2000 (http://www.invitrogen.com) following the manufacturer's instructions. Cells were harvested 24 h post-transfection, and cytoplasmic RNA was subsequently isolated using Tri-Reagent LS (Sigma-Aldrich). RNA samples were converted to cDNA using GoScript (Promega). One

hundred nanograms of cDNA was used as templates for PCR using Bulls Eye rTaq (Midwest Scientific). The sequences of the PCR primers used in this study are the following: Reporter Forward, CAAACAGACACCATGGTGCACC; Reporter Reverse, AACAGCATC AGGAGTGGACAGATC; *SRSF6* Forward, TACGGCTTCGTGGAGT TCGAG; *SRSF6* Reverse, TCTTGCCAACTGCACCGACTAG. Following PCR, the amplicons were purified using Purelink microcentrifuge columns (Invitrogen). Amplicons corresponding to alternative mRNA isoforms were separated with 6% (29:1) polyacrylamide gel electrophoresis and visualized using syberSAFE staining (Invitrogen). Bands corresponding to exon inclusion and exclusion were cut out and validated by DNA sequencing (data not shown). The linearity of the PCR reaction was confirmed by assaying splicing at increasing PCR cycles (Supplemental Fig. 12). For the experiments described above, quantification was performed following 29 cycles of PCR. Ratios corresponding to splicing efficiencies (% exon inclusion) were used to assay the effects of single nucleotide substitutions between samples rather than the absolute amount of each product. Molar ratios of mRNA isoforms were quantified using peak integration on a DNA1000 chip using an Agilent 2100 Bioanalyzer. To assay for activity of nonsense-mediated decay, we treated cells with 100 μg/mL emetine dihydrocholoride hydrate (Fluka) 10 h before harvesting.

For the RNA interference assay, HeLa cells were transiently cotransfected with both the construct and appropriate siRNA (NTi, SRp20i, or PTBi) using DharmaFECT Duo (http://www.thermoscientific.com) according to the manufacturer's instructions and harvested at 48 h post-transfection. Nuclear protein was resolved on Novex 10% bis tris polyacryamide gels and transferred to Immobilon FL (Millipore) using a Genie Blotter (Idea Scientific). Antibodies corresponding to PTB (mAb BB7), SRp20 (Sigma-Aldrich), and GAPDH (Calbiochem) were visualized with fluorescent-labeled secondary antibodies (GE) using the Fluoro-Chem Q system (Cell Bioscience). Following purification of cytoplasmic RNA using Tri-reagent LS, amplicons were generated using Onestep RT-PCR (Invitrogen), and the following cycling program: 30 min at 55°C; 2 min at 94°C; and 30 cycles of 30 sec at 94°C, 30 sec at 59°C, and 60 sec at 72°C.

RNA affinity chromatography

RNA affinity chromatography was performed as previously described (Caputi and Zahler 2001) with the following modifications: For the *OPA1* ligands, we selected a region 35 nt upstream of and 25 nt downstream from the 5'-splice site from exon 12 of the *OPA1* gene. Both the wild-type and mutant alleles were sequenced using IDT Custom Gene Synthesis. RNA was transcribed in vitro using T7 RNA polymerase (Ambion) and gel-purified from 6% (19:1) polyacrylamide gels. Fifteen hundred picomoles of purified RNA was oxidized by metaperiodate treatment and coupled to adipic acid dihydrazide agarose beads (Sigma-Aldrich). 1.5 mg of HeLa nuclear extract was incubated with the beads coupled to wild-type or mutant RNA bait, washed, and eluted with increasing concentrations of KCl. One-half of the sample was resolved by 10% Novex Nupage gel electrophoresis and silver-stained (Silver SNAP; BD Bioscience). The remaining half was precipitated with 20% TCA and washed in acetone. The protein pellet was analyzed by MudPIT Mass spectrometry at the Vincent J. Coates Proteomic Laboratory at University of California, Berkeley. Differences in peptide coverage between the wild-type and mutant eluates were quantified using MASCOT (Perkins et al. 1999), and peptide spectra from each sample were compared using CONTRAST (Tabb et al. 2002). A complete table of all peptides identified in both eluates can be found in Supplemental Tables 5 and 6. Criteria settings for CONTRAST can be found in Supplemental Table 7.

References

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. 2010. Deciphering the splicing code. *Nature* **465:** 53–59.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* **57:** 289–300.

Biasiotto G, Camaschella C, Forni GL, Polotti A, Zecchina G, Arosio P. 2008. New *TFR2* mutations in young Italian patients with hemochromatosis. *Haematologica* **93:** 309–310.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299:** 1391–1394.

Bruno C, Tamburino L, Kawashima N, Andreu AL, Shanske S, Hadjigeorgiou GM, Kawashima A, DiMauro S. 1999. A nonsense mutation in the myophosphorylase gene in a Japanese family with McArdle's disease. *Neuromuscul Disord* **9:** 34–37.

Camaschella C, Roetto A, Cali A, De Gobbi M, Garozzo G, Carella M, Majorano N, Totaro A, Gasparini P. 2000. The gene *TFR2* is mutated in a new type of haemochromatosis mapping to 7q22. *Nat Genet* **25:** 14–15.

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

Cartegni L, Krainer AR. 2002. Disruption of an SF2/ASF-dependent exonic splicing enhancer in *SMN2* causes spinal muscular atrophy in the absence of *SMN1*. *Nat Genet* **30:** 377–384.

Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* **3:** 285–298.

Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. 2003. ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res* **31:** 3568–3571.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* **7:** 98–108.

Cooper TA, Wan L, Dreyfuss G. 2009. RNA and disease. *Cell* **136:** 777–793.

Dietz HC, Valle D, Francomano CA, Kendzior RJ Jr., Pyeritz RE, Cutting GR. 1993. The skipping of constitutive exons in vivo induced by nonsense mutations. *Science* **259:** 680–683.

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073.

Eperon IC, Makarova OV, Mayeda A, Munroe SH, Caceres JF, Hayward DG, Krainer AR. 2000. Selection of alternative 5' splice sites: Role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. *Mol Cell Biol* **20:** 8303–8318.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* **297:** 1007–1013.

Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism–based validation of exonic splicing enhancers. *PLoS Biol* **2:** e268. doi: 10.1371/journal.pbio.0020268.

Fraser HB, Xie X. 2009. Common polymorphic transcript variation in human disease. *Genome Res* **19:** 567–575.

Ghigna C, Giordano S, Shen H, Benvenuto F, Castiglioni F, Comoglio PM, Green MR, Riva S, Biamonti G. 2005. Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. *Mol Cell* **20:** 881–890.

Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G. 2006. Comparative analysis identifies exonic splicing regulatory sequences—the complex definition of enhancers and silencers. *Mol Cell* **22:** 769–781.

Goren A, Kim E, Amit M, Bochner R, Lev-Maor G, Ahituv N, Ast G. 2008. Alternative approach to a heavy weight problem. *Genome Res* **18:** 214–220.

Graveley BR, Maniatis T. 1998. Arginine/serine-rich domains of SR proteins can function as activators of pre-mRNA splicing. *Mol Cell* **1:** 765–771.

Graveley BR, Hertel KJ, Maniatis T. 2001. The role of U2AF35 and U2AF65 in enhancer-dependent splicing. *RNA* **7:** 806–818.

Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36:** D773–D779.

Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* **11:** 345–355.

Kohtz JD, Jamison SF, Will CL, Zuo P, Luhrmann R, Garcia-Blanco MA, Manley JL. 1994. Protein–protein interactions and 5′-splice-site recognition in mammalian mRNA precursors. *Nature* **368:** 119–124.

Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, Cooper DN. 2007. Single base-pair substitutions in exon–intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum Mutat* **28:** 150–158.

Kumar S, Suleski MP, Markov GJ, Lawrence S, Marco A, Filipski AJ. 2009. Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. *Genome Res* **19:** 1562–1569.

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* **446:** 926–929.

Li B, Wachtel C, Miriami E, Yahalom G, Friedlander G, Sharon G, Sperling R, Sperling J. 2002. Stop codons affect 5′ splice site selection by surveillance of splicing. *Proc Natl Acad Sci* **99:** 5277–5282.

Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. 2011. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci* **108:** 11093–11098.

Liu HX, Zhang M, Krainer AR. 1998. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* **12:** 1998–2012.

Liu HX, Cartegni L, Zhang MQ, Krainer AR. 2001. A mechanism for exon skipping caused by nonsense or missense mutations in *BRCA1* and other genes. *Nat Genet* **27:** 55–58.

Majewski J, Ott J. 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res* **12:** 1827–1836.

Maquat LE. 2004. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol* **5:** 89–99.

Margulies EH, Blanchette M, Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res* **13:** 2507–2518.

McGlincy NJ, Smith CW. 2008. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem Sci* **33:** 385–393.

Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, Shiue L, Clark TA, Blume JE, 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.

Noensie EN, Dietz HC. 2001. A strategy for disease gene identification through nonsense-mediated mRNA decay inhibition. *Nat Biotechnol* **19:** 434–439.

Nogales-Gadea G, Rubio JC, Fernandez-Cadenas I, Garcia-Consuegra I, Lucia A, Cabello A, Garcia-Arumi E, Arenas J, Andreu AL, Martin MA. 2008. Expression of the muscle glycogen phosphorylase gene in patients with McArdle disease: The role of nonsense-mediated mRNA decay. *Hum Mutat* **29:** 277–283.

Pagani F, Baralle FE. 2004. Genomic variants in exons and introns: identifying the splicing spoilers. *Nat Rev Genet* **5:** 389–396.

Pagani F, Buratti E, Stuani C, Baralle FE. 2003. Missense, nonsense, and neutral mutations define juxtaposed regulatory elements of splicing in cystic fibrosis transmembrane regulator exon 9. *J Biol Chem* **278:** 26580–26588.

Pagano M, Gauvreau K. 2000. *Principles of biostatistics*, 2nd ed. Duxbury Press, Pacific Grove, CA.

Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol* **23:** 301–309.

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20:** 3551–3567.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2009. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20:** 110–121.

Rajaram S, Oono Y. 2010. NeatMap—non-clustering heat map alternatives in R. *BMC Bioinformatics* **11:** 45.

Ram O, Ast G. 2007. SR proteins: a foot on the exon before the transition from intron to exon definition. *Trends Genet* **23:** 5–7.

Rothrock C, Cannon B, Hahm B, Lynch KW. 2003. A conserved signal-responsive sequence mediates activation-induced alternative splicing of CD45. *Mol Cell* **12:** 1317–1324.

Sanford JR, Wang X, Mort M, Vanduyn N, Cooper DN, Mooney SD, Edenberg HJ, Liu Y. 2009. Splicing factor *SFRS1* recognizes a functionally diverse landscape of RNA transcripts. *Genome Res* **19:** 381–394.

Schimpf S, Schaich S, Wissinger B. 2006. Activation of cryptic splice sites is a frequent splicing defect mechanism caused by mutations in exon and intron sequences of the *OPA1* gene. *Hum Genet* **118:** 767–771.

Schimpf S, Fuhrmann N, Schaich S, Wissinger B. 2008. Comprehensive cDNA study and quantitative transcript analysis of mutant *OPA1* transcripts containing premature termination codons. *Hum Mutat* **29:** 106–112.

Shi FD, Zhang JY, Liu D, Rearden A, Elliot M, Nachtsheim D, Daniels T, Casiano CA, Heeb MJ, Chan EK, et al. 2005. Preferential humoral immune response in prostate cancer to cellular proteins p90 and p62 in a panel of tumor-associated antigens. *Prostate* **63:** 252–258.

Shiga N, Takeshima Y, Sakamoto H, Inoue K, Yokota Y, Yokoyama M, Matsuo M. 1997. Disruption of the splicing enhancer sequence within exon 27 of the dystrophin gene by a nonsense mutation induces partial skipping of the exon and is responsible for Becker muscular dystrophy. *J Clin Invest* **100:** 2204–2210.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050.

Stenson PD, Ball E, Howells K, Phillips A, Mort M, Cooper DN. 2008. Human Gene Mutation Database: towards a comprehensive central mutation database. *J Med Genet* **45:** 124–126.

Tabb DL, McDonald WH, Yates JR III. 2002. DTASelect and Contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* **1:** 21–26.

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

Valentine CR, Heflich RH. 1997. The association of nonsense mutation with exon-skipping in hprt mRNA of Chinese hamster ovary cells results from an artifact of RT-PCR. *RNA* **3:** 660–676.

Wachtel C, Li B, Sperling J, Sperling R. 2004. Stop codon-mediated suppression of splicing is a novel nuclear scanning mechanism not affected by elements of protein synthesis and NMD. *RNA* **10:** 1740–1750.

Wang GS, Cooper TA. 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8:** 749–761.

Wang Z, Burge CB. 2008. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA* **14:** 802–813.

Wang J, Chang YF, Hamilton JI, Wilkinson MF. 2002. Nonsense-associated altered splicing: A frame-dependent response distinct from nonsense-mediated decay. *Mol Cell* **10:** 951–957.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. *Cell* **119:** 831–845.

Wang Y, Xiao X, Van Nostrand E, Burge CB. 2006. General and specific functions of exonic splicing silencers in splicing control. *Mol Cell* **23:** 61–70.

Warf MB, Berglund JA. 2010. Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci* **35:** 169–178.

Watakabe A, Tanaka K, Shimura Y. 1993. The role of exon sequences in splice site selection. *Genes Dev* **7:** 407–418.

Xue Y, Zhou Y, Wu T, Zhu T, Ji X, Kwon YS, Zhang C, Yeo G, Black DL, Sun H, et al. 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, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11:** 377–394.

Yeo G, Holste D, Kreiman G, Burge CB. 2004. Variation in alternative splicing across human tissues. *Genome Biol* **5:** R74. doi: 10.1186/gb-2004-5-10-r74.

Yu Y, Maroney PA, Denker JA, Zhang XH, Dybkov O, Luhrmann R, Jankowsky E, Chasin LA, Nilsen TW. 2008. Dynamic regulation of alternative splicing by silencers that modulate 5′ splice site competition. *Cell* **135:** 1224–1236.

Zatkova A, Messiaen L, Vandenbroucke I, Wieser R, Fonatsch C, Krainer AR, Wimmer K. 2004. Disruption of exonic splicing enhancer elements is the principal cause of exon skipping associated with seven nonsense or missense alleles of *NF1*. *Hum Mutat* **24:** 491–501.

Zhu J, Mayeda A, Krainer AR. 2001. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol Cell* **8:** 1351–1361.

Dissertation Author Contributions

J.H: Performed RNA affinity chromatography and prepares samples for MudPIT; contributed to writing of manuscript

Chapter 4

Concluding Remarks and Future Directions

Here I have discussed the data and results of two projects that have employed use of iCLIP. I studied the mechanisms by which exon identity is determined during human pre-mRNA splicing, as well as the interplay between splicing factors hnRNP A1, SRSF1, and U2AF2 that lead to choices alternative splicing during gene expression. I have also used iCLIP to elucidated mRNA targets of an orphan RNA binding protein IGF2BP3 in both pancreatic and B-cell leukemia cell models, which not only led to a greater understanding in how IGF2BP3 interacts with RNA in a transcriptome-wide context, but insights into a testable mechanism by which IGF2BP3 appears to regulate gene expression, that being possible interactions near miRNA seed to potentially both promote and inhibit binding of the Argonatue 2-containing RISC complex. I have also show work where I have observed changes in protein occupancy of exons with regard to functional human mutations associated with disease. Overall, I have identified the functions of multiple RNA binding proteins and related these functions to human disease.

The coming years will undoubtedly see an explosion in data utilizing high-throughput assays (e.g. HITS-CLIP, iCLIP, etc.) to determine the transcriptome-wide RNA-interactions networks of a variety of RNA binding proteins. These studies will provide a more general overview as to what RBP-RNA interactome, such as how these proteins associate with their RNA targets and regulate them. This will certainly solidify the notion that RNA metabolism and gene expression is arguably the most highly regulated and complex of cellular processes. The challenge for future work is to begin to determine how fluctuations in the levels of RNA binding proteins

influence the binding specificity of other proteins and splicing factors globally (Baltz et al., 2012; Pandit et al., 2013; Schueler et al., 2014; Zarnack et al., 2013). These types of experiments will elucidate the context-specific interactions that determine how exon identity is established in living cells as well as other RBP "codes" as they function in downstream steps of gene expression. Overall, the near future holds a greater understanding of how RNA binding proteins govern the RNA world.

APPENDIX

List of known CLIP experiments in the literature

Protein	CLIP Methodology	Model System/Cell Type	Reference
Nova	CLIP	Mouse Brain	Ule et al.
Nova2	HITS-CLIP	Mouse Brain	Licatalosi et al.
SRSF1	CLIP-seq	HEK293T	Sanford et al..
	CLIP-seq	MEFs	Pandit et al..
Fox2	CLIP-seq	hESCs	Yeo et al..
Argonaute	HITS-CLIP	Mouse Brain	Chi et al..
	CLIP-seq	C. elegans	Zisoulis et al.
	CLIP-seq	mESCs	Leung AK et al.
	HITS-CLIP	Human Pancreatic Islets	Kameswaran et al.
	HITS-CLIP	Human Brain	Boudreau et al.
	HITS-CLIP	Marmoset T-Cells	Guo et al..
		MCF7	Pillai et al.
		BT474	Pillai et al.
		MDA-231	Pillai et al.
	HITS-CLIP	HCV-infected cells	Luna et al..
	CLIP-seq	293S cells	Karginov et al..
	CLIP-seq	C2C12 cells	Zhang et al..

	PAR-CLIP	FlpIn T-Rex HEK 293	Hafner et al.
	PAR-CLIP	HIV-1 infected TZM-bl cells	Whisnant et al..
	PAR-CLIP	Human B-Cells	Jin et al..
	iCLIP	C. elegans	Broughton and Pasquinelli
	iCLIP	hESCs	Bosson et al.
	iCLIP	mESCs	Bosson et al.
FMRP	HITS-CLIP	Mouse Brain	Darnell et al.
EBV	HITS-CLIP	Human B-Cells	Riley KJ
	PAR-CLIP	B95-8-infected lymphoblastoid cell lines	Skalsky et al..
DGCR8	HITS-CLIP	HEK293T	Macias S
FUS	HITS-CLIP	mouse cerebrum	Ishigaki S
	HITS-CLIP	Human brain	Nakaya et al.
	HITS-CLIP	Differentiated Mouse neurons	Nakaya et al.
	CLIP-seq	HeLa	Zhuo et al..
	PAR-CLIP	FlpInTrex HEK293	Hoell et al.
	iCLIP	Mouse Brain	Rogelj et al..
SRSF2	CLIP-seq	MEFs	Pandit et al..
Mili	HITS-CLIP	Mouse Testis Tissue	Vourekas et al.

Miwi	HITS-CLIP	Mouse Testis Tissue	Vourekas et al.
Mbnl1	HITS-CLIP	MEFs	Batra R et al..
	CLIP-seq	Mouse Brain	Wang et al.
	CLIP-seq	Mouse Heart	Wang et al.
	CLIP-seq	Mouse Muscle	Wang et al.
	CLIP-seq	C2C12 cells	Wang et al.
Mbnl2	HITS-CLIP	Mouse Brain	Charizanis et al..
	HITS-CLIP	MEFs	Batra R et al..
Mbnl3	HITS-CLIP	MEFs	Batra R et al..
RBP42	HITS-CLIP	T. brucei	Das et al..
nElavl	HITS-CLIP	Mouse Brain	Ince-Dunn et al..
RbFox1	HITS-CLIP	Mouse Brain	Weyn-Vanhentenryck et al.
	CLIP-seq	Mouse Brain	Lovci et al..
DeaD	HITS-CLIP	E. coli	Vakulskas et al..
RBM47	HITS-CLIP	231-BrM2	Vanharanta et al.
APC	HITS-CLIP	Mouse Brain	Preitner et al..
KSHV ORF57	HITS-CLIP	PEL cels	Sei et al..
lin-28	HITS-CLIP	C. elegans	Stefani et al..
	CLIP-seq	hESCs	Wilbert et al.
	CLIP-seq	Flp-In 293 cells	Wilbert et al.
	CLIP-seq	mESCs	Cho et al..

	CLIP-seq	Caco-2 cells	Madison et al.
	CLIP-seq	DLD1 cells	Madison et al.
	CLIP-seq	Lovo cells	Madison et al.
	PAR-CLIP	HEK293T	Hafner et al..
	individual domain PAR-CLIP	HEK293T	Graf et al..
hnRNP A1	CLIP-seq	HEK293T	Huelga et al.
hnRNP A2/B1	CLIP-seq	HEK293T	Huelga et al.
hnRNP F	CLIP-seq	HEK293T	Huelga et al.
hnRNP H1	CLIP-seq	HEK293T	Huelga et al.
	CLIP-seq	HEK293T	Katz et al.
hnRNP M	CLIP-seq	HEK293T	Huelga et al.
hnRNP U	CLIP-seq	HEK293T	Huelga et al.
	CLIP-seq	HeLa	Xiao et al..
hnRNP L	CLIP-seq	human CD4(+) T cells	Shankarling et al.
	CLIP-seq	Jurkat T Cells	Shankarling et al.
eIF4AIII	CLIP-seq	HeLa	Saulière et al.
UPF1	CLIP-seq	mESCs	Hurt et al..
PTB	CLIP-seq	HeLa cells	Xue et al..
EWSR1	PAR-CLIP	FlpInTRex HEK 293	Hoell et al.
	CLIP-seq	HeLa cells	Paronetto et al.

TAF15	PAR-CLIP	FlpIn T-Rex HEK 293	Hoell et al.
RBM20	CLIP-seq	Rat Cardiomyocytes	Maatz et al.
	PAR-CLIP	293T cells	Maatz et al.
DDX17	CLIP-seq	U2OS cells	Moy et al..
ADAR1	CLIP-seq	U87MG cells	Bahn et al..
	iCLIP	hESC2	Chen et al..
CELF1	CLIP-seq	Mouse Heart	Wang et al.
	CLIP-seq	Mouse Muscle	Wang et al.
	CLIP-seq	C2C12 cells	Wang et al.
PUM2	PAR-CLIP	FlpIn T-Rex HEK 293	Hafner et al.
QKI	PAR-CLIP	FlpIn T-Rex HEK 293	Hafner et al.
IGF2BP1	PAR-CLIP	FlpIn T-Rex HEK 293	Hafner et al.
IGF2BP2	PAR-CLIP	FlpIn T-Rex HEK 293	Hafner et al.
IGF2BP3	PAR-CLIP	FlpIn T-Rex HEK 293	Hafner et al.
AGO1	PAR-CLIP	FlpIn T-Rex HEK 293	Hafner et al.
AGO3	PAR-CLIP	FlpIn T-Rex HEK 293	Hafner et al.

AGO4	PAR-CLIP	FlpIn T-Rex HEK 293	Hafner et al.
TNRC6A	PAR-CLIP	FlpIn T-Rex HEK 293	Hafner et al.
TNRC6B	PAR-CLIP	FlpIn T-Rex HEK 293	Hafner et al.
TNRC6C	PAR-CLIP	FlpIn T-Rex HEK 293	Hafner et al.
HuR	PAR-CLIP	FlpIn T-Rex HEK 293	Mukherjee et al.
	PAR-CLIP	HeLa	Lebedeva et al..
	iCLIP	Mouse B cells	Diaz-Muñoz et al..
GLD-1	*in vivo* PAR-CLIP	C. elegans	Jungkamp et al..
KSHV	PAR-CLIP	BC-1 cells	Gottwein et al.
RBM4	PAR-CLIP	Human renal proximal tubular epithelial cells	Uniacke et al..
MOV10	PAR-CLIP	FlpIn T-Rex HEK 293	Sievers et al..
	PAR-CLIP	FlpIn T-Rex HEK 293	Gregersen et al..
	iCLIP	HEK293T	Kenny et al..
FBL	PAR-CLIP	FlpIn T-Rex HEK 293	Kishore et al.
DKC1	PAR-CLIP	FlpIn T-Rex HEK	Kishore et al.

		293	
NOP58	PAR-CLIP	FlpIn T-Rex HEK 293	Kishore et al.
Cirbp	PAR-CLIP	MEFs	Liu et al..
Rbm3	PAR-CLIP	MEFs	Liu et al..
RBM10	PAR-CLIP	FlpIn T-Rex HEK 293	Wang et al..
WTAP	PAR-CLIP	HEK293T	Ping et al..
METTL3	PAR-CLIP	HEK293T	Ping et al..
CTCF	PAR-CLIP	U2OS cells	Saldana-Meyer et
RBPMS	PAR-CLIP	HEK293T	Farazi et al..
RBPMS2	PAR-CLIP	HEK293T	Farazi et al..
Ataxin-2	PAR-CLIP	HEK293T	Yokoshi et al..
RbFox3	PAR-CLIP	P19 cells	Kim et al..
	PAR-CLIP	Mouse Brain	Kim et al..
WDR33	PAR-CLIP	HEK293T	Schonemann et al.
	iCLIP	HEK293T	Chan et al..
CPSF-160	PAR-CLIP	HEK293T	Martin et al..
CPSF-100	PAR-CLIP	HEK293T	Martin et al..
CPSF-73	PAR-CLIP	HEK293T	Martin et al..
CPSF-30	PAR-CLIP	HEK293T	Martin et al..

	iCLIP	HEK293T	Chan et al..
Fip1	PAR-CLIP	HEK293T	Martin et al..
CstF-64	PAR-CLIP	HEK293T	Martin et al..
CstF-64τ	PAR-CLIP	HEK293T	Martin et al..
CF I(m)25	PAR-CLIP	HEK293T	Martin et al..
CF I(m)59	PAR-CLIP	HEK293T	Martin et al..
CF I(m)68	PAR-CLIP	HEK293T	Martin et al..
AUF1	PAR-CLIP	HEK293T	Yoon et al..
hnRNP C	iCLIP	HeLa	Konic et al..
	PAR-CLIP	HEK293T	Liu et al..
hnRNPLL	PAR-CLIP	MPC11 plasma cell	Chang et al..
eIF3	PAR-CLIP	HEK293T	Lee et al..
TIA1	iCLIP	HeLa	Wang et al..
TIAL1	iCLIP	HeLa	Wang et al..
TDP-43	iCLIP	SH-SY5Y neuroblastoma	Tollervey et al..
	iCLIP	hESCs	Tollervey et al..
	CLIP-seq	Mouse Brain	Polymenidou et al..
SRSF3	iCLIP	P19 cells	Änkö et al..
SRSF4	iCLIP	P19 cells	Änkö et al..
U2AF2	CLIP-seq	HeLa	Shao et al..
	iCLIP	HeLa	Zarnack et al..

	iCLIP	Flp-In HEK293 cell	Schor et al..
NSun2	methylation iCLIP	Cos7	Hossain et al..
	methylation iCLIP	HEK293T	Hossain et al..
hnRNP L	iCLIP	HeLa	Rossbach et al..
RbFox2	iCLIP	mESCs	Jengi et al..
	CLIP-seq	Mouse Brain	Lovci et al..
B52	iCLIP	S2 cells	Bradley et al..
Rbp1	iCLIP	S2 cells	Bradley et al..
Srp54	iCLIP	S2 cells	Bradley et al..
XL6	iCLIP	S2 cells	Bradley et al..
U1C	iCLIP	T. brucei	Preußer et al..
U1-70K	iCLIP	T. brucei	Preußer et al..
Coilin	iCLIP	HeLa	Machyna et al..
	iCLIP	P19 cells	Machyna et al..
APOBEC3G	iCLIP	CEM-SS T cells infected with HIV	Apolonia et al.
APOBEC3F	iCLIP	CEM-SS T cells infected with HIV	Apolonia et al..
Matrin3	iCLIP	HeLa	Coelho et al..
PTB	iCLIP	HeLa	Coelho et al..
Celf4	iCLIP	Mouse Brain	Wagnon et al..
RBM38	iCLIP	HeLa	Léveillé et al..

Rrm4	iCLIP	Ustilago maydis	Koepke et al. .

Bibliography

Aifantis, I., Raetz, E., and Buonamici, S. (2008). Molecular pathogenesis of T-cell leukaemia and lymphoma. Nature reviews Immunology *8*, 380-390.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology *11*, R106.

Baltz, A.G., Munschauer, M., Schwanhausser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M.*, et al*. (2012). The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Molecular cell *46*, 674-690.

Barrandon, C., Bonnet, F., Nguyen, V.T., Labas, V., and Bensaude, O. (2007). The transcription-dependent dissociation of P-TEFb-HEXIM1-7SK RNA relies upon formation of hnRNP-7SK RNA complexes. Molecular and cellular biology *27*, 6996-7006.

Bekenstein, U., and Soreq, H. (2013). Heterogeneous nuclear ribonucleoprotein A1 in health and neurodegenerative disease: from structural insights to post-transcriptional regulatory roles. Molecular and cellular neurosciences *56*, 436-446.

Bell, J.L., Wachter, K., Muhleck, B., Pazaitis, N., Kohn, M., Lederer, M., and Huttelmaier, S. (2013). Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression? Cell Mol Life Sci *70*, 2657-2675.

Berkhout, B., and Klaver, B. (1993). In vivo selection of randomly mutated retroviral genomes. Nucleic acids research *21*, 5020-5024.

Bernt, K.M., and Armstrong, S.A. (2011). A role for DOT1L in MLL-rearranged leukemias. Epigenomics *3*, 667-670.

Beyer, A.L., Christensen, M.E., Walker, B.W., and LeStourgeon, W.M. (1977). Identification and characterization of the packaging proteins of core 40S hnRNP particles. Cell *11*, 127-138.

Blackinton, J.G., and Keene, J.D. (2014). Post-transcriptional RNA regulons affecting cell cycle and proliferation. Semin Cell Dev Biol *34*, 44-54.

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. The EMBO journal *18*, 1939-1952.

Boylan, K.L., Mische, S., Li, M., Marques, G., Morin, X., Chia, W., and Hays, T.S. (2008). Motility screen identifies Drosophila IGF-II mRNA-binding protein--zipcode-binding protein acting in oogenesis and synaptogenesis. PLoS genetics *4*, e36.

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. The EMBO journal *13*, 1197-1204.

Burke, D.H., and Gold, L. (1997). RNA aptamers to the adenosine moiety of S-adenosyl methionine: structural inferences from variations on a theme and the reproducibility of SELEX. Nucleic acids research *25*, 2020-2024.

Buvoli, M., Cobianchi, F., and Riva, S. (1992). Interaction of hnRNP A1 with snRNPs and pre-mRNAs: evidence for a possible role of A1 RNA annealing activity in the first steps of spliceosome assembly. Nucleic acids research *20*, 5017-5025.

Caceres, J.F., Stamm, S., Helfman, D.M., and Krainer, A.R. (1994). Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. Science *265*, 1706-1709.

Campillos, M., Lamas, J.R., Garcia, M.A., Bullido, M.J., Valdivieso, F., and Vazquez, J. (2003). Specific interaction of heterogeneous nuclear ribonucleoprotein A1 with the -219T allelic form modulates APOE promoter activity. Nucleic acids research *31*, 3063-3070.

Cartegni, L., and Krainer, A.R. (2002). Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. Nature genetics *30*, 377-384.

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

Castello, A., Fischer, B., Hentze, M.W., and Preiss, T. (2013). RNA-binding proteins in Mendelian disease. Trends in genetics : TIG *29*, 318-327.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics *14*, 128.

Chen, L.L., and Carmichael, G.G. (2009). Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. Molecular cell *35*, 467-478.

Chiou, N.T., Shankarling, G., and Lynch, K.W. (2013). hnRNP L and hnRNP A1 induce extended U1 snRNA interactions with an exon to repress spliceosome assembly. Molecular cell *49*, 972-982.

Cho, S., Moon, H., Loh, T.J., Jang, H.N., Liu, Y., Zhou, J., Ohn, T., Zheng, X., and Shen, H. (2015). Splicing inhibition of U2AF65 leads to alternative exon skipping. Proceedings of the National Academy of Sciences of the United States of America *112*, 9926-9931.

Copley, M.R., Babovic, S., Benz, C., Knapp, D.J., Beer, P.A., Kent, D.G., Wohrer, S., Treloar, D.Q., Day, C., Rowe, K., *et al*. (2013). The Lin28b-let-7-Hmga2 axis determines the higher self-renewal potential of fetal haematopoietic stem cells. Nature cell biology *15*, 916-925.

Dassi, E., Zuccotti, P., Leo, S., Provenzani, A., Assfalg, M., D'Onofrio, M., Riva, P., and Quattrone, A. (2013). Hyper conserved elements in vertebrate mRNA 3'-UTRs reveal a translational network of RNA-binding proteins controlled by HuR. Nucleic acids research *41*, 3201-3216.

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.

Dawson, M.A., Prinjha, R.K., Dittmann, A., Giotopoulos, G., Bantscheff, M., Chan, W.I., Robson, S.C., Chung, C.W., Hopf, C., Savitski, M.M., *et al*. (2011). Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. Nature *478*, 529-533.

Deer, E.L., Gonzalez-Hernandez, J., Coursen, J.D., Shea, J.E., Ngatia, J., Scaife, C.L., Firpo, M.A., and Mulvihill, S.J. (2010). Phenotype and genotype of pancreatic cancer cell lines. Pancreas *39*, 425-435.

Dirksen, W.P., Li, X., Mayeda, A., Krainer, A.R., and Rottman, F.M. (2000). Mapping the SF2/ASF binding sites in the bovine growth hormone exonic splicing enhancer. The Journal of biological chemistry *275*, 29170-29177.

Ellington, A.D., and Szostak, J.W. (1990). In vitro selection of RNA molecules that bind specific ligands. Nature *346*, 818-822.

Eperon, I.C., Makarova, O.V., Mayeda, A., Munroe, S.H., Caceres, J.F., Hayward, D.G., and Krainer, A.R. (2000). Selection of alternative 5' splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. Molecular and cellular biology *20*, 8303-8318.

Ernst, P., Mabon, M., Davidson, A.J., Zon, L.I., and Korsmeyer, S.J. (2004). An Mll-dependent Hox program drives hematopoietic progenitor expansion. Current biology : CB *14*, 2063-2069.

Esposito, I., Konukiewitz, B., Schlitter, A.M., and Kloppel, G. (2014). Pathology of pancreatic ductal adenocarcinoma: facts, challenges and future developments. World J Gastroenterol *20*, 13833-13841.

Fadare, O., Liang, S.X., Crispens, M.A., Jones, H.W., 3rd, Khabele, D., Gwin, K., Zheng, W., Mohammed, K., Parkash, V., Hecht, J.L., *et al*. (2013). Expression of the oncofetal protein IGF2BP3 in endometrial clear cell carcinoma: assessment of frequency and significance. Hum Pathol *44*, 1508-1515.

Fairbrother, W.G., and Chasin, L.A. (2000). Human genomic sequences that inhibit splicing. Molecular and cellular biology *20*, 6816-6825.

Findeis-Hosey, J.J., and Xu, H. (2011). The use of insulin like-growth factor II messenger RNA binding protein-3 in diagnostic pathology. Hum Pathol *42*, 303-314.

Flynn, R.A., Martin, L., Spitale, R.C., Do, B.T., Sagan, S.M., Zarnegar, B., Qu, K., Khavari, P.A., Quake, S.R., Sarnow, P.*, et al*. (2015). Dissecting noncoding and pathogen RNA-protein interactomes. Rna *21*, 135-143.

Friedersdorf, M.B., and Keene, J.D. (2014). Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. Genome biology *15*, R2.

Gal-Mark, N., Schwartz, S., Ram, O., Eyras, E., and Ast, G. (2009). The pivotal roles of TIA proteins in 5' splice-site selection of alu exons and across evolution. PLoS genetics *5*, e1000717.

Ge, H., and Manley, J.L. (1990). A protein factor, ASF, controls cell-specific alternative splicing of SV40 early pre-mRNA in vitro. Cell *62*, 25-34.

Gerstberger, S., Hafner, M., Ascano, M., and Tuschl, T. (2014a). Evolutionary conservation and expression of human RNA-binding proteins and their role in human genetic disease. Advances in experimental medicine and biology *825*, 1-55.

Gerstberger, S., Hafner, M., and Tuschl, T. (2014b). A census of human RNA-binding proteins. Nature reviews Genetics *15*, 829-845.

Gong, C., and Maquat, L.E. (2011). lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. Nature *470*, 284-288.

Gu, W., Katz, Z., Wu, B., Park, H.Y., Li, D., Lin, S., Wells, A.L., and Singer, R.H. (2012). Regulation of local expression of cell adhesion and motility-related mRNAs in breast cancer cells by IMP1/ZBP1. J Cell Sci *125*, 81-91.

Guenther, M.G., Lawton, L.N., Rozovskaia, T., Frampton, G.M., Levine, S.S., Volkert, T.L., Croce, C.M., Nakamura, T., Canaani, E., and Young, R.A. (2008). Aberrant chromatin at genes encoding stem cell regulators in human mixed-lineage leukemia. Genes & development *22*, 3403-3408.

Guil, S., and Caceres, J.F. (2007). The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. Nature structural & molecular biology *14*, 591-596.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., *et al*. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223-227.

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

Hamilton, B.J., Burns, C.M., Nichols, R.C., and Rigby, W.F. (1997). Modulation of AUUUA response element binding by heterogeneous nuclear ribonucleoprotein A1 in human T lymphocytes. The roles of cytoplasmic location, transcription, and phosphorylation. The Journal of biological chemistry *272*, 28732-28741.

Hamilton, B.J., Nagy, E., Malter, J.S., Arrick, B.A., and Rigby, W.F. (1993). Association of heterogeneous nuclear ribonucleoprotein A1 and C proteins with reiterated AUUUA sequences. The Journal of biological chemistry *268*, 8881-8887.

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

Hansen, T.V., Hammer, N.A., Nielsen, J., Madsen, M., Dalbaeck, C., Wewer, U.M., Christiansen, J., and Nielsen, F.C. (2004). Dwarfism and impaired gut development in insulin-like growth factor II mRNA-binding protein 1-deficient mice. Molecular and cellular biology *24*, 4448-4464.

Hardy, R.R., and Shinton, S.A. (2004). Characterization of B lymphopoiesis in mouse bone marrow and spleen. Methods Mol Biol *271*, 1-24.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., *et al*. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome research *22*, 1760-1774.

Hartmann, E.M., Bea, S., Navarro, A., Trapp, V., Campo, E., Ott, G., and Rosenwald, A. (2012). Increased tumor cell proliferation in mantle cell lymphoma is associated with elevated insulin-like growth factor 2 mRNA-binding protein 3 expression. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc *25*, 1227-1235.

Hasler, J., and Strub, K. (2006). Alu elements as regulators of gene expression. Nucleic acids research *34*, 5491-5497.

He, Y., and Smith, R. (2009). Nuclear functions of heterogeneous nuclear ribonucleoproteins A/B. Cell Mol Life Sci *66*, 1239-1256.

Ho, J.J., and Marsden, P.A. (2014). Competition and collaboration between RNA-binding proteins and microRNAs. Wiley interdisciplinary reviews RNA *5*, 69-86.

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.

Huizenga, D.E., and Szostak, J.W. (1995). A DNA aptamer that binds adenosine and ATP. Biochemistry *34*, 656-665.

Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y., Tajnik, M., Konig, J., and Ule, J. (2014). iCLIP: protein-RNA interactions at nucleotide resolution. Methods *65*, 274-287.

Imamura, T., Morimoto, A., Takanashi, M., Hibi, S., Sugimoto, T., Ishii, E., and Imashuku, S. (2002). Frequent co-expression of HoxA9 and Meis1 genes in infant acute lymphoblastic leukaemia with MLL rearrangement. British journal of haematology *119*, 119-121.

Jean-Philippe, J., Paz, S., and Caputi, M. (2013). hnRNP A1: the Swiss army knife of gene expression. International journal of molecular sciences *14*, 18999-19024.

Jens, M., and Rajewsky, N. (2015). Competition between target sites of regulators shapes post-transcriptional gene regulation. Nature reviews Genetics *16*, 113-126.

Jiang, P., Singh, M., and Coller, H.A. (2013). Computational assessment of the cooperativity between RNA binding proteins and MicroRNAs in Transcript Decay. PLoS Comput Biol *9*, e1003075.

Jiang, X., Huang, H., Li, Z., Li, Y., Wang, X., Gurbuxani, S., Chen, P., He, C., You, D., Zhang, S., *et al.* (2012). Blockade of miR-150 maturation by MLL-fusion/MYC/LIN-28 is required for MLL-associated leukemia. Cancer cell *22*, 524-535.

Jo, O.D., Martin, J., Bernath, A., Masri, J., Lichtenstein, A., and Gera, J. (2008). Heterogeneous nuclear ribonucleoprotein A1 regulates cyclin D1 and c-myc internal ribosome entry site function through Akt signaling. The Journal of biological chemistry *283*, 23274-23287.

Jonson, L., Christiansen, J., Hansen, T.V., Vikesa, J., Yamamoto, Y., and Nielsen, F.C. (2014). IMP3 RNP safe houses prevent miRNA-directed HMGA2 mRNA decay in cancer and development. Cell reports *7*, 539-551.

Jude, C.D., Climer, L., Xu, D., Artinger, E., Fisher, J.K., and Ernst, P. (2007). Unique and independent roles for MLL in adult hematopoietic stem cells and progenitors. Cell stem cell *1*, 324-337.

Jurica, M.S., Licklider, L.J., Gygi, S.R., Grigorieff, N., and Moore, M.J. (2002). Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis. Rna *8*, 426-439.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic acids research *32*, D493-496.

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

Kedde, M., Strasser, M.J., Boldajipour, B., Oude Vrielink, J.A., Slanchev, K., le Sage, C., Nagel, R., Voorhoeve, P.M., van Duijse, J., Orom, U.A., *et al.* (2007). RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. Cell *131*, 1273-1286.

Kedde, M., van Kouwenhove, M., Zwart, W., Oude Vrielink, J.A., Elkon, R., and Agami, R. (2010). A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility. Nat Cell Biol *12*, 1014-1020.

Keene, J.D. (2007). RNA regulons: coordination of post-transcriptional events. Nature reviews Genetics *8*, 533-543.

Kellenberger, E., Stier, G., and Sattler, M. (2002). Induced folding of the U2AF35 RRM upon binding to U2AF65. FEBS letters *528*, 171-176.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome research *12*, 996-1006.

Kharas, M.G., Lengner, C.J., Al-Shahrour, F., Bullinger, L., Ball, B., Zaidi, S., Morgan, K., Tam, W., Paktinat, M., Okabe, R., *et al*. (2010). Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. Nature medicine *16*, 903-908.

Kielkopf, C.L., Lucke, S., and Green, M.R. (2004). U2AF homology motifs: protein recognition in the RRM world. Genes & development *18*, 1513-1526.

Kielkopf, C.L., Rodionova, N.A., Green, M.R., and Burley, S.K. (2001). A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. Cell *106*, 595-605.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology *14*, R36.

Kim, H.H., Kuwano, Y., Srikantan, S., Lee, E.K., Martindale, J.L., and Gorospe, M. (2009). HuR recruits let-7/RISC to repress c-Myc expression. Genes & development *23*, 1743-1748.

King, R.L., Pasha, T., Roullet, M.R., Zhang, P.J., and Bagg, A. (2009). IMP-3 is differentially expressed in normal and neoplastic lymphoid tissue. Hum Pathol *40*, 1699-1705.

Kistler, A.L., and Guthrie, C. (2001). Deletion of MUD2, the yeast homolog of U2AF65, can bypass the requirement for sub2, an essential spliceosomal ATPase. Genes & development *15*, 42-49.

Klipper-Aurbach, Y., Wasserman, M., Braunspiegel-Weintrob, N., Borstein, D., Peleg, S., Assa, S., Karp, M., Benjamini, Y., Hochberg, Y., and Laron, Z. (1995). Mathematical formulae for the prediction of the residual beta cell function during the first two years of disease in children and adolescents with insulin-dependent diabetes mellitus. Med Hypotheses *45*, 486-490.

Kobel, M., Xu, H., Bourne, P.A., Spaulding, B.O., Shih Ie, M., Mao, T.L., Soslow, R.A., Ewanowich, C.A., Kalloger, S.E., Mehl, E., *et al*. (2009). IGF2BP3 (IMP3) expression is a marker of unfavorable prognosis in ovarian carcinoma of clear cell subtype. Mod Pathol *22*, 469-475.

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. Nature structural & molecular biology *17*, 909-915.

Krainer, A.R., Conway, G.C., and Kozak, D. (1990). The essential pre-mRNA splicing factor SF2 influences 5' splice site selection by activating proximal sites. Cell *62*, 35-42.

Krivtsov, A.V., and Armstrong, S.A. (2007). MLL translocations, histone modifications and leukaemia stem-cell development. Nature reviews Cancer *7*, 823-833.

Krivtsov, A.V., Feng, Z., Lemieux, M.E., Faber, J., Vempati, S., Sinha, A.U., Xia, X., Jesneck, J., Bracken, A.P., Silverman, L.B.*, et al*. (2008). H3K79 methylation profiles define murine and human MLL-AF4 leukemias. Cancer Cell *14*, 355-368.

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.

Lederer, M., Bley, N., Schleifer, C., and Huttelmaier, S. (2014). The role of the oncofetal IGF2 mRNA-binding protein 3 (IGF2BP3) in cancer. Seminars in cancer biology.

Lev-Maor, G., Ram, O., Kim, E., Sela, N., Goren, A., Levanon, E.Y., and Ast, G. (2008). Intronic Alus influence alternative splicing. PLoS genetics *4*, e1000204.

Li, B., Wan, X., Zhu, Q., Li, L., Zeng, Y., Hu, D., Qian, Y., Lu, L., Wang, X., and Meng, X. (2013). Net expression inhibits the growth of pancreatic ductal adenocarcinoma cell PL45 in vitro and in vivo. PloS one *8*, e57818.

Li, D., Ren, W., Wang, X., Wang, F., Gao, Y., Ning, Q., Han, Y., Song, T., and Lu, S. (2009). A modified method using TRIzol reagent and liquid nitrogen produces high-quality RNA from rat pancreas. Applied biochemistry and biotechnology *158*, 253-261.

Li, J., Kim, T., Nutiu, R., Ray, D., Hughes, T.R., and Zhang, Z. (2014a). Identifying mRNA sequence elements for target recognition by human Argonaute proteins. Genome research *24*, 775-785.

Li, W., Liu, D., Chang, W., Lu, X., Wang, Y.L., Wang, H., Zhu, C., Lin, H.Y., Zhang, Y., Zhou, J., *et al*. (2014b). Role of IGF2BP3 in trophoblast cell invasion and migration. Cell Death Dis *5*, e1025.

Liao, B., Hu, Y., Herrick, D.J., and Brewer, G. (2005). The RNA-binding protein IMP-3 is a translational activator of insulin-like growth factor II leader-3 mRNA during proliferation of human K562 leukemia cells. The Journal of biological chemistry *280*, 18517-18524.

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., *et al*. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464-469.

Lochhead, P., Imamura, Y., Morikawa, T., Kuchiba, A., Yamauchi, M., Liao, X., Qian, Z.R., Nishihara, R., Wu, K., Meyerhardt, J.A., *et al*. (2012). Insulin-like growth factor 2 messenger RNA binding protein 3 (IGF2BP3) is a marker of unfavourable prognosis in colorectal cancer. European journal of cancer *48*, 3405-3413.

Lovci, M.T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T.Y., Stark, T.J., Gehman, L.T., Hoon, S., *et al*. (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nature structural & molecular biology *20*, 1434-1442.

Lu, D., Yang, X., Jiang, N.Y., Woda, B.A., Liu, Q., Dresser, K., Mercurio, A.M., Rock, K.L., and Jiang, Z. (2011). IMP3, a new biomarker to predict progression of cervical intraepithelial neoplasia into invasive cancer. The American journal of surgical pathology *35*, 1638-1645.

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

Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nature reviews Molecular cell biology *8*, 479-490.

Luo, J., Solimini, N.L., and Elledge, S.J. (2009). Principles of cancer therapy: oncogene and non-oncogene addiction. Cell *136*, 823-837.

Mabon, S.A., and Misteli, T. (2005). Differential recruitment of pre-mRNA splicing factors to alternatively spliced transcripts in vivo. PLoS Biol *3*, e374.

Mackereth, C.D., Madl, T., Bonnal, S., Simon, B., Zanier, K., Gasch, A., Rybin, V., Valcarcel, J., and Sattler, M. (2011). Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. Nature *475*, 408-411.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., *et al*. (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res *41*, D64-69.

Michlewski, G., and Caceres, J.F. (2010). Antagonistic role of hnRNP A1 and KSRP in the regulation of let-7a biogenesis. Nature structural & molecular biology *17*, 1011-1018.

Michlewski, G., Guil, S., and Caceres, J.F. (2011). Stimulation of pri-miR-18a Processing by hnRNP A1. Advances in experimental medicine and biology *700*, 28-35.

Mili, S., Shu, H.J., Zhao, Y., and Pinol-Roma, S. (2001). Distinct RNP complexes of shuttling hnRNP proteins with pre-mRNA and mRNA: candidate intermediates in formation and export of mRNA. Molecular and cellular biology *21*, 7307-7319.

Mogilyansky, E., and Rigoutsos, I. (2013). The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. Cell Death Differ *20*, 1603-1614.

Moore, M.J. (2005). From birth to death: the complex lives of eukaryotic mRNAs. Science *309*, 1514-1518.

Mount, S.M., Pettersson, I., Hinterberger, M., Karmas, A., and Steitz, J.A. (1983). The U1 small nuclear RNA-protein complex selectively binds a 5' splice site in vitro. Cell *33*, 509-518.

Mueller-Pillasch, F., Pohl, B., Wilda, M., Lacher, U., Beil, M., Wallrapp, C., Hameister, H., Knochel, W., Adler, G., and Gress, T.M. (1999). Expression of the highly conserved RNA binding protein KOC in embryogenesis. Mech Dev *88*, 95-99.

Mullighan, C.G. (2012). Molecular genetics of B-precursor acute lymphoblastic leukemia. J Clin Invest *122*, 3407-3415.

Mullighan, C.G., and Downing, J.R. (2009). Genome-wide profiling of genetic alterations in acute lymphoblastic leukemia: recent insights and future directions. Leukemia *23*, 1209-1218.

Nielsen, F.C., Nielsen, J., and Christiansen, J. (2001). A family of IGF-II mRNA binding proteins (IMP) involved in RNA trafficking. Scandinavian journal of clinical and laboratory investigation Supplementum *234*, 93-99.

Nielsen, J., Christiansen, J., Lykke-Andersen, J., Johnsen, A.H., Wewer, U.M., and Nielsen, F.C. (1999). A family of insulin-like growth factor II mRNA-binding

proteins represses translation in late development. Molecular and cellular biology *19*, 1262-1270.

O'Connell, R.M., Balazs, A.B., Rao, D.S., Kivork, C., Yang, L., and Baltimore, D. (2010). Lentiviral vector delivery of human interleukin-7 (hIL-7) to human immune system (HIS) mice expands T lymphocyte populations. PLoS One *5*, e12009.

Okunola, H.L., and Krainer, A.R. (2009). Cooperative-binding and splicing-repressive properties of hnRNP A1. Molecular and cellular biology *29*, 5620-5631.

Palanichamy, J.K., and Rao, D.S. (2014). miRNA dysregulation in cancer: towards a mechanistic understanding. Frontiers in genetics *5*, 54.

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. Molecular cell *50*, 223-235.

Pastor, T., and Pagani, F. (2011). Interaction of hnRNPA1/A2 and DAZAP1 with an Alu-derived intronic splicing enhancer regulates ATM aberrant splicing. PloS one *6*, e23349.

Pino, I., Pio, R., Toledo, G., Zabalegui, N., Vicent, S., Rey, N., Lozano, M.D., Torre, W., Garcia-Foncillas, J., and Montuenga, L.M. (2003). Altered patterns of expression of members of the heterogeneous nuclear ribonucleoprotein (hnRNP) family in lung cancer. Lung Cancer *41*, 131-143.

Placke, T., Faber, K., Nonami, A., Putwain, S.L., Salih, H.R., Heidel, F.H., Kramer, A., Root, D.E., Barbie, D.A., Krivtsov, A.V.*, et al*. (2014). Requirement for CDK6 in MLL-rearranged acute myeloid leukemia. Blood *124*, 13-23.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M.*, et al*. (2014). RefSeq: an update on mammalian reference sequences. Nucleic acids research *42*, D756-763.

Pruitt, K.D., Harrow, J., Harte, R.A., Wallin, C., Diekhans, M., Maglott, D.R., Searle, S., Farrell, C.M., Loveland, J.E., Ruef, B.J.*, et al*. (2009). The consensus coding

sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome research *19*, 1316-1323.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

R Development Core Team (2008). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).

Rain, J.C., Rafi, Z., Rhani, Z., Legrain, P., and Kramer, A. (1998). Conservation of functional domains involved in RNA binding and protein-protein interactions in human and Saccharomyces cerevisiae pre-mRNA splicing factor SF1. Rna *4*, 551-565.

Rao, D.S., O'Connell, R.M., Chaudhuri, A.A., Garcia-Flores, Y., Geiger, T.L., and Baltimore, D. (2010). MicroRNA-34a perturbs B lymphocyte development by repressing the forkhead box transcription factor Foxp1. Immunity *33*, 48-59.

Ray, D., Kazan, H., Chan, E.T., Pena Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q., and Hughes, T.R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nature biotechnology *27*, 667-670.

Reid, D.C., Chang, B.L., Gunderson, S.I., Alpert, L., Thompson, W.A., and Fairbrother, W.G. (2009). Next-generation SELEX identifies sequence and structural determinants of splicing factor binding in human pre-mRNA sequence. Rna *15*, 2385-2397.

Robinson, B.W., Behling, K.C., Gupta, M., Zhang, A.Y., Moore, J.S., Bantly, A.D., Willman, C.L., Carroll, A.J., Adamson, P.C., Barrett, J.S., *et al*. (2008). Abundant anti-apoptotic BCL-2 is a molecular target in leukaemias with t(4;11) translocation. British journal of haematology *141*, 827-839.

Rooke, N., Markovtsov, V., Cagavi, E., and Black, D.L. (2003). Roles for SR proteins and hnRNP A1 in the regulation of c-src exon N1. Molecular and cellular biology *23*, 1874-1884.

Rozovskaia, T., Feinstein, E., Mor, O., Foa, R., Blechman, J., Nakamura, T., Croce, C.M., Cimino, G., and Canaani, E. (2001). Upregulation of Meis1 and HoxA9 in acute lymphocytic leukemias with the t(4 : 11) abnormality. Oncogene *20*, 874-878.

Ruskin, B., Zamore, P.D., and Green, M.R. (1988). A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. Cell *52*, 207-219.

Schaeffer, D.F., Owen, D.R., Lim, H.J., Buczkowski, A.K., Chung, S.W., Scudamore, C.H., Huntsman, D.G., Ng, S.S., and Owen, D.A. (2010). Insulin-like growth factor 2 mRNA binding protein 3 (IGF2BP3) overexpression in pancreatic ductal adenocarcinoma correlates with poor survival. BMC Cancer *10*, 59.

Schueler, M., Munschauer, M., Gregersen, L.H., Finzel, A., Loewer, A., Chen, W., Landthaler, M., and Dieterich, C. (2014). Differential protein occupancy profiling of the mRNA transcriptome. Genome biology *15*, R15.

Schwartz, S., Gal-Mark, N., Kfir, N., Oren, R., Kim, E., and Ast, G. (2009). Alu exonization events reveal features required for precise recognition of exons by the splicing machinery. PLoS computational biology *5*, e1000300.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G.*, et al*. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. Science *343*, 84-87.

Shao, C., Yang, B., Wu, T., Huang, J., Tang, P., Zhou, Y., Zhou, J., Qiu, J., Jiang, L., Li, H.*, et al*. (2014). Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome. Nature structural & molecular biology *21*, 997-1005.

Shorter, J., and Taylor, J.P. (2013). Disease mutations in the prion-like domains of hnRNPA1 and hnRNPA2/B1 introduce potent steric zippers that drive excess RNP granule assembly. Rare diseases *1*, e25200.

Smith, A.D., Chung, W.Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z., and Zhang, M.Q. (2009). Updates to the RMAP short-read mapping software. Bioinformatics *25*, 2841-2842.

Smith, A.D., Xuan, Z., and Zhang, M.Q. (2008). Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC bioinformatics *9*, 128.

Smith, G.K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical Applications in Genetics and Molecular Biology *3*.

Somervaille, T.C., Matheny, C.J., Spencer, G.J., Iwasaki, M., Rinn, J.L., Witten, D.M., Chang, H.Y., Shurtleff, S.A., Downing, J.R., and Cleary, M.L. (2009). Hierarchical maintenance of MLL myeloid leukemia stem cells employs a transcriptional program shared with embryonic rather than adult stem cells. Cell Stem Cell *4*, 129-140.

Sorek, R., Ast, G., and Graur, D. (2002). Alu-containing exons are alternatively spliced. Genome research *12*, 1060-1067.

Stoskus, M., Gineikiene, E., Valceckiene, V., Valatkaite, B., Pileckyte, R., and Griskevicius, L. (2011). Identification of characteristic IGF2BP expression patterns in distinct B-ALL entities. Blood Cells Mol Dis *46*, 321-326.

Sun, Q., Mayeda, A., Hampson, R.K., Krainer, A.R., and Rottman, F.M. (1993). General splicing factor SF2/ASF promotes alternative splicing by binding to an exonic splicing enhancer. Genes & development *7*, 2598-2608.

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PloS one *6*, e21800.

Suvasini, R., Shruti, B., Thota, B., Shinde, S.V., Friedmann-Morvinski, D., Nawaz, Z., Prasanna, K.V., Thennarasu, K., Hegde, A.S., Arivazhagan, A., *et al*. (2011). Insulin growth factor-2 binding protein 3 (IGF2BP3) is a glioblastoma-specific marker that activates phosphatidylinositol 3-kinase/mitogen-activated protein kinase (PI3K/MAPK) pathways by modulating IGF-2. The Journal of biological chemistry *286*, 25882-25890.

Tang, H., Wei, Q., Ge, J., Jian, W., Liu, J., Zhong, L., Fu, B., and Zhao, T. (2013). IMP3 as a supplemental diagnostic marker for Hodgkin lymphoma. Hum Pathol *44*, 2167-2172.

Taniuchi, K., Furihata, M., Hanazaki, K., Saito, M., and Saibara, T. (2014). IGF2BP3-mediated translation in cell protrusions promotes cell invasiveness and metastasis of pancreatic cancer. Oncotarget *5*, 6832-6845.

Tavanez, J.P., Madl, T., Kooshapur, H., Sattler, M., and Valcarcel, J. (2012). hnRNP A1 proofreads 3' splice site recognition by U2AF. Molecular cell *45*, 314-329.

Toledano, H., D'Alterio, C., Czech, B., Levine, E., and Jones, D.L. (2012). The let-7-Imp axis regulates ageing of the Drosophila testis stem-cell niche. Nature *485*, 605-610.

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

Tsutsumi, S., Taketani, T., Nishimura, K., Ge, X., Taki, T., Sugita, K., Ishii, E., Hanada, R., Ohki, M., Aburatani, H.*, et al*. (2003). Two distinct gene expression signatures in pediatric acute lymphoblastic leukemia with MLL rearrangements. Cancer research *63*, 4882-4887.

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.

Uren, P.J., Bahrami-Samani, E., Burns, S.C., Qiao, M., Karginov, F.V., Hodges, E., Hannon, G.J., Sanford, J.R., Penalva, L.O., and Smith, A.D. (2012). Site identification in high-throughput RNA-protein interaction data. Bioinformatics *28*, 3013-3020.

Uren, P.J., Burns, S.C., Ruan, J., Singh, K.K., Smith, A.D., and Penalva, L.O. (2011). Genomic analyses of the RNA-binding protein Hu antigen R (HuR) identify a complex network of target genes and novel characteristics of its binding sites. The Journal of biological chemistry *286*, 37063-37066.

Ushigome, M., Ubagai, T., Fukuda, H., Tsuchiya, N., Sugimura, T., Takatsuka, J., and Nakagama, H. (2005). Up-regulation of hnRNP A1 gene in sporadic human colorectal cancers. Int J Oncol *26*, 635-640.

Valcarcel, J., Gaur, R.K., Singh, R., and Green, M.R. (1996). Interaction of U2AF65 RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA [corrected]. Science *273*, 1706-1709.

Valcarcel, J., Singh, R., Zamore, P.D., and Green, M.R. (1993). The protein Sex-lethal antagonizes the splicing factor U2AF to regulate alternative splicing of transformer pre-mRNA. Nature *362*, 171-175.

Venables, J.P., Bourgeois, C.F., Dalgliesh, C., Kister, L., Stevenin, J., and Elliott, D.J. (2005). Up-regulation of the ubiquitous alternative splicing factor Tra2beta causes inclusion of a germ cell-specific exon. Human molecular genetics *14*, 2289-2303.

Vikesaa, J., Hansen, T.V., Jonson, L., Borup, R., Wewer, U.M., Christiansen, J., and Nielsen, F.C. (2006). RNA-binding IMPs promote cell adhesion and invadopodia formation. The EMBO journal *25*, 1456-1468.

Wachter, K., Kohn, M., Stohr, N., and Huttelmaier, S. (2013). Subcellular localization and RNP formation of IGF2BPs (IGF2 mRNA-binding proteins) is modulated by distinct RNA-binding domains. Biol Chem *394*, 1077-1090.

Wagner, M., Kunsch, S., Duerschmied, D., Beil, M., Adler, G., Mueller, F., and Gress, T.M. (2003). Transgenic overexpression of the oncofetal RNA binding protein KOC leads to remodeling of the exocrine pancreas. Gastroenterology *124*, 1901-1914.

Webb, C.J., and Wise, J.A. (2004). The splicing factor U2AF small subunit is functionally conserved between fission yeast and humans. Molecular and cellular biology *24*, 4229-4240.

Wu, J.Y., and Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. Cell *75*, 1061-1070.

Wu, T., and Fu, X.D. (2015). Genomic functions of U2AF in constitutive and regulated splicing. RNA biology *12*, 479-485.

Xia, H. (2005). Regulation of gamma-fibrinogen chain expression by heterogeneous nuclear ribonucleoprotein A1. The Journal of biological chemistry *280*, 13171-13178.

Xiao, S.H., and Manley, J.L. (1997). Phosphorylation of the ASF/SF2 RS domain affects both protein-protein and protein-RNA interactions and is necessary for splicing. Genes & development *11*, 334-344.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338-345.

Yaniv, K., Fainsod, A., Kalcheim, C., and Yisraeli, J.K. (2003). The RNA-binding protein Vg1 RBP is required for cell migration during early neural development. Development *130*, 5649-5661.

Yaniv, K., and Yisraeli, J.K. (2002). The involvement of a conserved family of RNA binding proteins in embryonic development and carcinogenesis. Gene *287*, 49-54.

Yisraeli, J.K. (2005). VICKZ proteins: a multi-talented family of regulatory RNA-binding proteins. Biology of the cell / under the auspices of the European Cell Biology Organization *97*, 87-96.

Yu, J., Ohuchida, K., Mizumoto, K., Fujita, H., Nakata, K., and Tanaka, M. (2010). MicroRNA miR-17-5p is overexpressed in pancreatic cancer, associated with a poor prognosis, and involved in cancer cell proliferation and invasion. Cancer Biol Ther *10*, 748-757.

Zahler, A.M., Damgaard, C.K., Kjems, J., and Caputi, M. (2004). SC35 and heterogeneous nuclear ribonucleoprotein A/B proteins bind to a juxtaposed exonic splicing enhancer/exonic splicing silencer element to regulate HIV-1 tat exon 2 splicing. The Journal of biological chemistry *279*, 10077-10084.

Zarnack, K., Konig, J., Tajnik, M., Martincorena, I., Eustermann, S., Stevant, I., Reyes, A., Anders, S., Luscombe, N.M., and Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. Cell *152*, 453-466.

Zhang, Q.S., Manche, L., Xu, R.M., and Krainer, A.R. (2006). hnRNP A1 associates with telomere ends and stimulates telomerase activity. Rna *12*, 1116-1128.

Zhou, Z., Licklider, L.J., Gygi, S.P., and Reed, R. (2002). Comprehensive proteomic analysis of the human spliceosome. Nature *419*, 182-185.

Zhou, Z.J., Dai, Z., Zhou, S.L., Fu, X.T., Zhao, Y.M., Shi, Y.H., Zhou, J., and Fan, J. (2013). Overexpression of HnRNP A1 promotes tumor invasion through regulating CD44v6 and indicates poor prognosis for hepatocellular carcinoma. International journal of cancer Journal international du cancer *132*, 1080-1089.

Zhu, J., Mayeda, A., and Krainer, A.R. (2001). Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. Molecular cell *8*, 1351-1361.