UC Santa Cruz

UC Santa Cruz Electronic Theses and Dissertations

Title

Global analysis of post-transcriptional gene regulatory mechanisms

Permalink

https://escholarship.org/uc/item/2qn8p733

Author

Philipp, Julia

Publication Date

2021

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

GLOBAL ANALYSIS OF POST-TRANSCRIPTIONAL GENE REGULATORY MECHANISMS

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

DOCTOR OF PHILOSOPHY

in

MOLECULAR, CELL AND DEVELOPMENTAL BIOLOGY

by

Julia Philipp

June 2021

The Dissertation of Julia Philipp is approved:

_____
Professor Jeremy Sanford, Chair

_____
Professor Angela Brooks

_____
Professor Joshua Arribere

_____
Quentin Williams
Interim Vice Provost and Dean of Graduate Studies

Copyright © by

Julia Philipp

June 2021

Table of Contents

List of Figures

Abstract

Global analysis of post-transcriptional gene regulatory mechanisms
by Julia Philipp

Post-transcriptional regulation of gene expression is complicated and multi-leveled and essential for the final phenotype of a cell or tissue in disease and in health. RNA binding proteins (RBPs) are associated with mRNA transcripts throughout their life cycles and regulate many post-transcriptional processes including, but not limited to, mRNA stability, localization, and translation and have the ability to couple multiple of these processes. Aberrant regulation by RBPs can frequently result in malignancies like cancer.

Alternative splicing, a co- and post-transcriptional process, has diverse impacts on different levels of regulation of gene expression including the diversity of protein sequence, mRNA stability, and subcellular transcript localization as well.

Previously, the Sanford Lab and others discovered that alternative splicing may also influence translation by identifying mRNA transcripts from the same gene with distinct polyribosome association patterns in human cell lines. Here, I present data that shows that the coupling of alternative splicing with translational control is a conserved mechanism of gene regulation in higher primates and that specific mRNA sequences altered through alternative splicing seem responsible for this regulatory coupling.

Subsequently, I explored the changes in cis- and transregulatory landscape of these mRNAs that could be connected to the isoform-specific polysome association and translation: I investigated the effect of single nucleotide variants on RNA binding protein (RBP) binding, mRNA secondary structure, and codon optimality in relation to isoform-specific polysome association. While I found multiple candidate RBPs worth investigating, I only found a modest correlation between mRNA secondary structure or

codon optimality and isoform specific translation.

In addition, I explored the interaction of RNA binding proteins with microRNAs and was able to find a strong binding preference for RBPs binding to (which area) of microRNAs.

Further, I investigated the role of a specific RBP and a known marker for malignancy, IGF2BP3 in multiple oncogenic environments. In the context of B-cell acute lymphoblastic leukemia, I was able to identify high confidence targets of IGF2BP3 involved in leukemogenesis.

Taken together, these projects elucidate the complexity of protein/RNA interactions and their multifaceted abilities to regulate post-transcriptional processes both in healthy and disease phenotypes.

Finally, I am presenting BiocSwirl, a novel platform for teaching R based bioinformatics to students and scientists of all levels of computational understanding. The courses are interactive with live feedback and provide a rich and new learning experience in bioinformatics.

Acknowledgements

I would like to express my utmost gratitude to Dr. Jeremy Sanford for the incredible opportunity to pursue my PhD in this lab and at this university, and his continued mentorship through all the ups and downs. A huge 'Thank you' to my thesis committee, including Dr. Jeremy Sanford, Dr. Angela Brooks, and Dr. Joshua Arribere for their support, mentorship, and guidance. Thanks to every member for the Sanford Lab, past or present, for constant collaboration and scientific discussions, for Taco Tuesdays and Wine Walks. Thank you to my parents, Thore and Beatrice Philipp, extended family, and friends, near and far, for their unwavering support and grounding perspective. I would not have gotten this far without them. Thanks to the Seabright LadiesTM for their 'sage advice'. Finally, thanks to all the furry friends for their emotional support and reminding me that there is more to life than grad school (e.g. snacks and naps in sunny spots).

The text of this dissertation includes reprints of the following previously published material:

Tran T., Philipp J., Bassi J., Nibber N., Draper J., Lin T., Palanichamy J.K., Kumar A., Paing M., King J., Katzman S, Sanford J.R., Rao D.S. 2020. Post-transcriptional gene regulation by the RNA binding protein IGF2BP3 is critical for MLL-AF4 mediated leukemogenesis. bioRxiv doi: 10.1101/2020.12.20.423624

Dargytė M., Philipp J., Palka C., Howard J., Katzman S., Stone M., Sanford J. 2020. Splicing Factor SRSF1 expands the regulatory logic of microRNA expression. bioRxiv doi: 10.1101/2020.05.12.092270

The co-authors listed in these publications directed and supervised the research which forms the basis for the dissertation.

To my parents

1. Chapter 1: Global protein-RNA interactions

1.1 The central dogma of molecular biology, transcriptional and post-transcriptional regulation

The central dogma of molecular biology describes the flow of information from DNA to RNA to protein (Figure 1.1)(Crick, 1970). The two main processes within the Central Dogma are the transcription of DNA into RNA by RNA polymerase and the translation of RNA into proteins by ribosomes. However, aside from these central steps, multiple layers of regulation affect the steady-state protein levels and, therefore, the cell or tissue's phenotype.

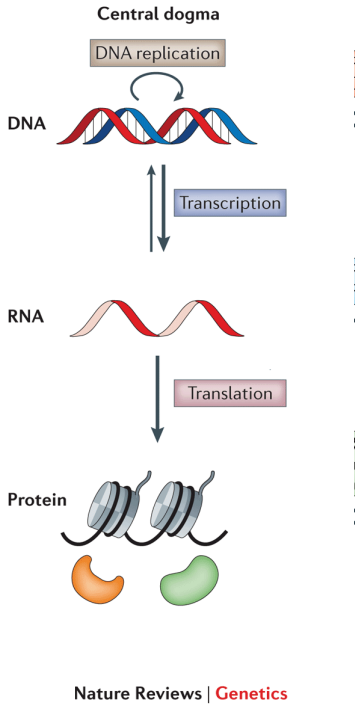

Figure 1.1 The Central Dogma of Molecular Biology.
Genetic information is stored in the DNA, located in the nucleus of the cell and replicates itself before cell division. DNA is transcribed into messenger RNA which is then transported out of the nucleus, where it can be translated into proteins by the ribosomes. Reverse transcription, the process of converting RNA into DNA, occurs only in exceptional cases, such as retroviruses or retrotransposons. Adapted from Fu et al., 2014, Nature Reviews Genetics..

During and after transcription, multiple mechanisms, referred to as co- and post-transcriptional regulatory mechanisms, result in various and diverse transcript isoforms. Differential regulation of transcription of some human genes results in as many as 80 different transcripts (Floor and Doudna, 2016), with an average of 6.3 alternatively spliced and 3.6 protein coding transcripts (ENCODE Project Consortium, 2012; Tung et al., 2020). The huge transcript diversity is a result of the combination of alternative transcriptional start site choice (Davuluri et al., 2008; (dgt) and The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014; Shiraki et al., 2003), alternative splicing (Xiong et al., 2015), and alternative polyadenylation (Di Giammartino et al., 2011; Hoque et al., 2012). Many of these processes are tissue or cell-type specific (Barbosa-Morais et al., 2012; Melé et al., 2015; Merkin et al., 2012) and contribute to a cell or tissue's identity via the regulation of gene expression.

While we have a decent understanding of the processes of post-transcriptional regulation as isolated mechanisms, the regulatory code and how these individual layers of regulation come together is yet to be understood in detail. The lack of detailed understanding becomes apparent when predicting protein expression levels based on mRNA abundance in a cell (or a system). The consensus is that mRNA and protein levels correlate to a very limited degree, the correlation coefficient (Pearson correlation) is about 0.35 to 0.40 (Kosti et al., 2016; Vogel and Marcotte, 2012). In humans, the mRNA concentration only explains approximately 27% of the protein abundance (Vogel et al., 2010). The discrepancy can be explained by the multiple levels of post-transcriptional regulation that modify the outcome of protein synthesis. Vogel et al. built predictive

models that included known regulators, such as the lengths of UTRs and coding sequences, putative regulatory elements, miRNA binding sites, and predicted secondary structures. Vogel et al.'s models were able to increase prediction levels to 67% leaving about 33% of protein abundance still unexplained (Vogel et al., 2010). While more recent reanalyses of these datasets suggest an initial underestimation of predictability, some uncertainty and variation in these models remain (Li et al., 2014a). This uncertainty indicates that there are still levels of interaction and regulation that remain to be uncovered in order to reliably explain steady-state protein expression levels.

1.2 RNA binding proteins

Many of the gene expression regulatory steps are governed by a family of proteins called RNA binding proteins (RBPs). RBPs are strongly evolutionarily conserved, are the most common family of proteins found in mammalian cells, and are almost ubiquitously expressed (Gerstberger et al., 2014). Current estimates range from 500-700 RBPs in mammalian cells (Anantharaman et al., 2002; Galante et al., 2009), of which only approximately two percent are estimated to be expressed in a tissue-specific manner (Dezső et al., 2008; Gerstberger et al., 2014; Ramsköld et al., 2009).

Complexes of multiple RBPs bind to RNA to form heterogeneous nuclear ribonucleoproteins (hnRNPs) or, if binding specifically to mRNA, messenger ribonuclear proteins (mRNPs). These mRNPs are involved in the regulation of diverse processes such as alternative splicing, mRNA stability and decay, mRNA localization, and translation (Gehring et al., 2017; Gerstberger et al., 2014). They are thought to work in complex networks and synergy and competition between different RBPs can lead to many differential and combinatorial actions (Gehring et al., 2017; Gerstberger et al., 2014; Keene, 2007; Lunde et al., 2007). Understanding RBPs and delineating their combinatorial ac-

tions will be vital to understanding RNA processing and regulation.

1.2.1 RNA binding domains

RRBPs interact with their target RNAs via a variety of different binding domains. Most RBPs bind to single-stranded RNA, typically in a sequence-dependent matter, however, multiple RBPs are also known to interact with double-stranded RNA secondary structures such as stem-loops depending on the nature of their RNA binding domains (RBD). Proteins mainly interact with RNA via hydrogen bonds (HB) and Van-der-Waals (VdW) interactions. While the estimates of HB to VdW ratios in protein-RNA interactions vary, VdW interactions are thought to be the predominant force (Corley et al., 2020). Further stabilizing forces in protein-RNA interactions can be provided by hydrophobic π interactions (Corley et al., 2020).

The sequence-specific RNA-recognition motif (RRM) is the RNA binding domain most commonly used in protein/RNA interactions (Änkö and Neugebauer, 2012). The RRM consists of 80-90 amino acids arranged into four antiparallel beta sheets and two alpha-helices, organized into the order βαββαβ. The binding specificity of RRMs is conferred by two highly conserved ribonucleoprotein domain (RNP) motifs RNP1 and RNP2 in the two central beta-sheets. Comparative analyses of structure-function relationships of different RRMs also identified the two framing beta sheets, the loops and the C- and N-termini of the domain as relevant for the RNA-binding specificity of the RRM domain (Maris et al., 2005). IGF2BP3, a protein of interest in Chapter four of this thesis, contains two RRMs. The K homology domain (KH domain), is the second most widely used RBD in humans (Änkö and Neugebauer, 2012) and is often found in splicing regulators as well as our previously mentioned protein of interest IMP3. The KH domain is approximately 70 amino acids long, folded into three beta-sheets and

two alpha-helices, organized into either βααββα (predominantly found in eukaryotes) or αββααβ (found in prokaryotes). The main contributor to the binding specificity of the KH domain is a cleft formed by the first two alpha helices and the second beta-sheet together with a variable loop. This cleft can typically bind four unpaired bases (Valverde et al., 2008).

Other RNA binding domains include DEAD motifs, often found in helicases and named after the sequence of their binding motif (Jarmoskaite and Russell, 2011), zinc finger domains (Hall, 2005), which are very small and the protein's binding specificity usually depends on the number of domains, e.g. in U2AF, SWAP domains (suppressor of white apricot homolog splicing factor), and PIWI domains (as found in the miRNA processing factor Ago1)(Parker et al., 2006). More than fifty percent of all RBPs are estimated to contain at least one of these popular RBDs (Gerstberger et al., 2014).

A substantial proportion of RBP binding is less sequence and more structure-dependent (Seemann et al., 2017; Stefl et al., 2005). Some RBPs with similar binding sequence motifs have been shown to differ in their structural binding preferences (Dominguez et al., 2018). Structure-dependent binding commonly happens through the double-stranded RNA binding domain (dsRBM), a 70-75aa long domain with conserved topology (St Johnston et al., 1992). This type of structure-dependent binding has been observed hundreds of times in eukaryotes, dozens of which were observed in humans (SMART database, data from (Stefl et al., 2005)). Some of the earliest and still prominent examples are the *D. melanogaster* protein Staufen (Änkö and Neugebauer, 2012; Ramos et al., 2000), the Xenopus laevis RNA-binding protein A (Ryter and Schultz, 1998), and an RNase III homolog from budding yeast (Stefl et al., 2005). Several RBPs have also been reported to bind to RNA guanine-quadruplex (rG4) structures (Herdy et al., 2018; Huang et al., 2018) capable of regulating gene expression (Wolfe et al., 2014). Their interactions, however, are in more need of functional characterization

(Komatsu et al., 2020). On a more global scale, it has been reported that the overall secondary structure of mRNAs positively correlates with the number of RBPs bound to the transcript (Sanchez de Groot et al., 2019).

1.3 Protein-RNA interactions in post-transcriptional regulatio of gene expression

As previously mentioned, RBPs engage in the regulation of many post-transcriptional mechanisms. Alternative splicing, the post-transcriptional excision of introns and ligation of exon sequences by the spliceosome, is a highly prevalent processing mechanism affecting as many as 60 to 70 percent of genes. This high prevalence indicates the importance of alternative splicing in expanding the cell's transcriptome. Alternative splicing is highly regulated by RBPs or so-called splicing factors. RBPs are part of the splicing code, aiding in exon recognition and intron/exon definition, binding to both splicing enhancer and silencer sequences, and stabilizing or destabilizing the interactions between mRNAs and the spliceosome. RBPs, specifically splicing factors, constitute the majority of the exon junction complexes (EJC). These EJC's are localized about 20 nucleotides upstream of the exon junctions and are vital for the recognition of exon junctions and the regulation of alternative splicing. The concentration of RBPs can therefore have an effect on the regulation of splicing (Kalnina et al., 2005).

Many RBPs are also required for the proper subcellular localization of transcripts. A subset of SR proteins, named after their prominent serine-arginine-rich (SR) domain, are responsible for shuttling mature mRNAs out of the nucleus into the cytoplasm for translation (Cáceres et al., 1997; Caceres et al., 1998). The EJCs label mature transcripts as successfully spliced and ready for nuclear export by SR proteins. EJCs might further be responsible for the localization of transcripts to other cellular compartments. For

example, the subcellular localization of the Oskar mRNA in *D. melanogaster* is dependent on two nuclear shuttling proteins, the human homologs of which are core components of the EJC (Hachet and Ephrussi, 2004). Mammalian equivalent proteins of the *D. melanogaster* EJC components (Magoh and Y14) have been found in the dendrites of neurons, suggesting that this mechanism is relevant for mammalian cells and mRNA regulation as well (Glanzer et al., 2005).

RBPs can regulate the stability and decay of mRNA transcripts through multiple pathways. AU-rich elements (ARE) are cis-regulatory elements in the 3' UTR known to affect the stability of mRNA transcripts (García-Mauriño et al., 2017). Different RBPs can recognize and bind to these AREs and either up- or down-regulate transcript stability. For example, the binding of Tristetraprolin (TTP) (Deleault et al., 2008) or AUF1 (Zhang et al., 1993) have destabilizing effects, while HuR appears to be stabilizing transcripts. RBPs, specifically those contributing to the EJC, are further involved in the recognition of premature termination codons and consequently truncated open reading frames, tagging mRNAs for their subsequent nonsense-mediated decay (NMD) pathway (Gehring et al., 2005). mRNA decay is typically triggered by the recruitment of decapping and/or de-polyadenylation factors and the transport of the target mRNA to the exosome (Wang and Kiledjian, 2001).

Finally, many RBPs play central roles in the regulation of mRNA translation. For example, translation initiation is the rate-limiting step of translation and is dependent on the RBP complex eIF4F (Sonenberg et al., 1979) (Tahara et al., 1981). eIF4F binds to the 5'-cap of an mRNA and is in turn bound by eIF4G and eIF4A forming the eIF4F complex. 4E-BP can bind to eIF4E, prevent the formation of the eiF4F complex, and prevent cap-dependent translation. If phosphorylated, 4E-BP is released, and the eiF4F

complex can form, and regular cap-dependent translation can proceed. The phosphorylation and, therefore, initiation repressing activity can be regulated by mTOR, which can be recruited to mRNA transcripts by SRSF1 (Kroczynska et al., 2009; Maslon et al., 2014). Another popular example is the translational control of ferritin by iron. The presence of iron in the cytoplasm will disrupt the translation-inhibiting binding of IRP-1 and IRP-2 to the iron response element (IRE) in the 5'UTR of the ferritin mRNA and therefore allow the translation of ferritin (Menotti et al., 1998). Further, the SR protein SRSF1 has been found to regulate the translation of select mRNAs via direct binding as well, coupling alternative splicing, shuttling, and translational control (Sterne-Weiler et al., 2013).

Observing the coupling of multiple post-transcriptional processes by SRSF1 introduced the idea of RBPs in complex and multifunctional regulatory roles. This is possible due to multifunctional domains and the RBP sticking with an mRNA throughout multiple stages of its life cycle (Figure 1.2). The prominent example for such multifunctional proteins is the previously mentioned SR proteins. SR proteins are serine-arginine rich, originally called splicing factors for their initially discovered function, and are involved in multiple post-transcriptional processes. They aid in recognizing exon-intron boundaries, mRNA shuttling out of the nucleus, and translation (Änkö, 2014; Sanford et al., 2004, 2005, 2009). We expect to identify more multifunctional RBPs, specifically splicing factors, involved in additional regulatory roles.

In addition to multifunctional RBPs, understanding post-transcriptional regulation is further complicated by the combinatorial actions of the sum of RBPs bound to an mRNA transcript. RBPs involved in similar regulatory processes often colocalize in the same region of the mRNA (Mukherjee et al., 2019), e.g., proteins affecting the transcript

stability bind the 3'UTR predominantly. It can therefore be expected that these neighboring RBPs interact with each other to finetune the expression of a transcript. Keene et al. first mentioned the idea of so-called RNA regulons (Keene, 2007), introducing the concept of large combinatorial networks of RBPs in a sometimes hierarchical structure that can have both cooperative or antagonizing effects on mRNA stability or translation (Gerstberger et al., 2014). RBPs with similar or overlapping binding sites could antagonize each other due to direct steric hindrance (Sternburg and Karginov, 2020). RBPs with nearby binding sites could also act cooperatively by enabling co-recruitment through potentially direct binding between two RBPs (Sternburg and Karginov, 2020). Another proposed mechanism of RBP binding affecting other RBPs is that RBP interactions can change the secondary structure of an mRNA transcript, affecting further RBP binding by masking or opening up binding sites for other proteins (Sternburg and Karginov, 2020).

Current and central questions in RNA biology revolve around the synergy and competition of individual RBPs and how that contributes to overall regulation and the target specificity of multifunctional RNA binding proteins involved in post-transcriptional regulation. To address these open questions and investigate protein-RNA interactions, the following *in vitro* and *in vivo* methods are commonly used.

1.4 Experimental Approaches to RNA protein interactions, in vitro

The **electrophoretic mobility shift assay (EMSA)** (Hellman and Fried, 2007) is an *in vitro*, low throughput, and qualitative assay to access protein/RNA interactions. This approach was first published by Fried and Crothers and Garner and Revzin (Fried and Crothers, 1981; Garner and Revzin, 1981). EMSA is based on the fact that protein-RNA complexes have lower electrophoretic mobility than unbound RNA. Lower

electrophoretic mobility means that protein/RNA complexes move slower through a polyacrylamide or agarose gel when subjected to gel electrophoresis. This assay is primarily qualitative. It can determine the presence or absence of protein/RBP complexes but not determine the number or type of RBPs bound. However, it can be used for quantitative analyses, e.g., stoichiometry, affinities, and kinetics (Fried, 1989; Fried and Garner, 1998; Schuck, 2007). Further, EMSA can be combined with western blotting or mass spectroscopy to identify the proteins in complex with the assayed RNA in a so-called "supershift" assay (Hellman and Fried, 2007). Briefly, 32P-labelled nucleic acids in complex with proteins (as the result of purification or crude cell extracts) are subjected to gel electrophoresis under native conditions on either polyacrylamide or agarose gel, and the results are visualized using autoradiography. This assay's limitations are, prominently, that the samples are not at chemical equilibrium, which could lead to rapid dissociation and, therefore, failure to detect specific protein/RNA complexes. Additionally, other factors such as the size of the proteins or the RNA can further affect the samples' mobility.

RNA affinity chromatography (Sharma, 2008) is another *in vitro* assay to investigate protein RNA interactions. In this assay, the goal is to isolate and identify interacting proteins. While it is not a global approach, it can help to understand the function of non-coding RNAs and examine the trans-acting landscape of individual RNA transcripts. Briefly, the assay consists of crosslinking *in vivo*, cell lysis and hybridization with biotin-labeled DNA probes, the capture of the complex with streptavidin-coated beads, disruption of the complex, elution of proteins or RNA, and subsequently either western blotting or mass spectroscopy to identify the proteins or sequencing, PCR, or Northern blotting to identify the RNA.

SELEX, tthe systematic evolution of ligands by exponential enrichment (Stoltenburg et al., 2007; Tuerk and Gold, 1990; Zhuo et al., 2017), is an *in vitro* assay for determining high-affinity sequence ligands for RBPs by relying on mechanisms of evolution and selection. SELEX is very good at identifying consensus motifs and is prominently used for aptamer selection in targeted drug therapies. However, it does not provide any quantitative results nor information about *in vivo* binding specificities (Änkö and Neugebauer, 2012) and has been criticized for its bias towards high-affinity motifs (Lambert et al., 2014). For the assay, a pool of RNAs is randomized at specific positions and then selected for binding by RBPs on nitrocellulose. The selected RNAs are amplified as double-stranded DNA, which is competent for *in vitro* transcription. The resulting RNA will be enriched for the preferred binding motifs. This RNA pool will be subjected to the selection and amplification cycle again. Finally, the enriched sequences can be clonally isolated and characterized, and used to determine consensus sequences. The SELEX assay has since been further developed to select single-stranded DNA successfully (Stoltenburg et al., 2007). Next-generation SELEX aims to evaluate the RBP binding to secondary structures (Änkö and Neugebauer, 2012; Reid et al., 2009). SELEX coupled to high-throughput sequencing is much faster than traditional SELEX (Nguyen Quang et al., 2016). Finally, multiple variations of vivo SELEX allow for the investigation of binding by native proteins (Sola et al., 2020).

RNAcompete is another *in vitro* approach to investigating protein-RNA interactions, both sequence and structure-dependent (Ray et al., 2009, 2013, 2017)(Figure 1.3). This approach is non-iterative and, therefore, faster than SELEX. The assay involves generating an RNA pool, initially, with all possible 10-base sequences and all possible 7 and 8 base stem-loops, with the help of specialized microarrays, followed by an incubation of the tagged (glutathione S-transferase (GST)) protein of interested with

excess RNA from the generated pool. The excess RNAs compete for binding with the protein, hence the assay's name, leading to the detection of high-affinity protein/RNA interactions. A GST pulldown is performed, the recovered RNA labeled with Cy5, some of the initial RNA pool labeled with Cy3, and both are co-hybridized to a microarray. Computationally, the enrichment of pulled-down RNA to the original RNA pool can be determined. The signal is measured as the log-ratio between bound and ingoing RNA, which is used to measure binding affinity and sequence preference (Ray et al., 2009). This assay's limitations can be the low temperatures at which it is performed, limiting the RNA secondary structures that can be tested this way (Lambert et al., 2014). A recent expansion of the assay with a single-step *in vitro* selection, called RNAcompete-S (Cook et al., 2017), allows for global detection of longer binding motifs (> 12 nt) as well. A large proportion of RNAcompete experiments have been made publicly available by the Hughes Lab (Ray et al., 2013). A list of available RNAcompete experiments can be found in the Appendix. RNAcompete data sets were used for the prediction of cis-regulatory sequences in Chapter Three.

Finally, **RNA Bind-n-Seq (RBNS)** (RBNS) (Lambert et al., 2014) is an *in vitro* assay that aims to overcome the majority of limitations of previously mentioned *in vitro* RBP-RNA binding assays. The recombinantly expressed and purified protein of interest is incubated with a pool of randomized RNAs, typically about 40nt long plus short primers (for subsequent adapter addition). The incubation is usually performed at different protein concentrations. The RBP is then captured utilizing its streptavidin binding tag, and the RNA bound to the captured protein is reverse transcribed, barcoded sequencing adapters are added, the RNA is amplified via PCR, and then sequenced. This assay's computational component entails calculating motif enrichment values (R) for each 5, 6, and 7-mer in the selected pool over the input pool. Previous experiments have shown

that the calculated R-value consistently correlates with known protein binding affinities. The influence of secondary structures on protein/RNA interactions can also be detected by combining the RBNS output with the Vienna RNAfold algorithm (Lorenz et al., 2011), which can calculate the intramolecular base-pairing probability for high-affinity motifs. The sequences are then binned based on motifs, protein concentration, and base-pairing, and R is calculated for each bin. The changes of R in relation to these measures can then be tested. A large proportion of RBNS experiments/ resulting data are publicly available on ENCODE (Davis et al., 2018; ENCODE Project Consortium, 2012). A list of available RNA Bind-n-Seq experiments can be found in the Appendix.

1.5 Global, in vivo approaches to studying protein-RNA interactions in vivo

All approaches mentioned above are *in vitro* and therefore fail to accurately consider the cellular environment and its potential effects on RNA-protein interactions and binding specificity (Änkö and Neugebauer, 2012). The highest affinity based on *in vitro* experiments does not necessarily mean that that binding site is the most used one. Further, even weaker binding sites might be biologically relevant or even advantageous. Finally, the contribution of other RBPs is frequently neglected *in vitro* but, as previously discussed, plays an essential role in the RBP regulatory network / trans-regulatory landscape. Therefore *in vivo* experiments play a vital role in elucidating protein/RNA interactions.

Comparative transcriptomics: Knockdown or overexpression of RBPs followed by microarray or next-generation sequencing (NGS) is a popular *in vivo* approach to understanding protein RNA interactions. This approach allows explicitly understanding the functional consequences of these interactions. Typically a knockdown or overex-

pression experiment is accompanied by a wildtype experiment as a control and comparison. Present-day, the preferred method is to prepare polyA-selected RNA sequencing libraries from knockdown or overexpressing cells or tissues and the corresponding wildtype experiment and perform next-generation sequencing on them (Wang et al., 2009). These two sequencing experiments can then be computationally compared for changes in expression between wildtype and knockdown or overexpression (DESeq2 (Love et al., 2014)). Any genes that change significantly between the two experimental conditions are likely functional targets of the RBP of interest. This approach can be used for most RBPs that bind to mRNA (or even small or noncoding RNAs). It is further helpful for investigating splicing factors since RNAseq data not only allows for analysis of differential expression but also of differential splicing patterns. Specifically for splicing analysis, the recent advances in long-read sequencing can be beneficial in these types of experiments. The approach's limitations are the difficulties in determining direct and indirect targets of the protein of interest since knockdown or overexpression experiments tend to lead to significant changes in gene expression in the whole regulatory network. This can typically be amended by combining this assay with an immunoprecipitation experiment (like RIP or CLIP, mentioned below) to determine direct and functional targets. Further, not every RBP can be assayed this way since not every RBP can be knocked down. Many knockdowns lead to lethal phenotypes. Protein overexpression routinely disrupts many gene regulatory and metabolic mechanisms leading to overall dysregulated gene expression (Bolognesi and Lehner, 2018).

RNA immunoprecipitation (RIP) in combination with either microarray analysis (RIP-chip)(Keene et al., 2006) or next-generation sequencing (RIP-seq)(Zhao et al., 2010) was one of the earliest approaches to identifying *in vivo* protein-RNA binding partners. Protein-RNA complexes are immunoprecipitated using antibodies specific to

the protein of interest. The pulled down, and isolated RNA is then analyzed using either microarrays or next-generation sequencing (NGS). The RNA binding specificity of the protein of interest can then be determined *in silico* based on the chip or NGS results. This analysis is typically performed on unaligned sequences. This approach can also be used to determine the composition of mRNPs. Since the protein-protein interactions are not disrupted during the immunoprecipitation, further experiments can be performed to identify the co-immunoprecipitated proteins.

Crosslinking and immunoprecipitation assays (CLIP)(Ule et al. 2003) started as low-throughput approaches. However, they were ultimately expanded for a genome-wide snapshot of interactions between specific RNA binding proteins and their target mRNAs when high-throughput sequencing became more readily available and affordable (HITS-CLIP (Licatalosi et al., 2008))(Figure 1.4A).

The main difference to the RIP-seq protocol is that the direct protein/RNA interactions are captured due to crosslinking. The proteins are cross-linked to their currently bound target mRNA *in vivo* with UV radiation. The crosslinking is only successful when protein and RNA are within 1 angstrom of each other. Following the crosslinking, the proteins and their bound RNA are co-immunoprecipitated using RBP-specific antibodies. The following RNAse digestion leaves only short stretches of 40-60nt of RNA bound by the RBP. Then, non-crosslinked RNA and proteins are removed in the presence of SDS, followed by PAGE and size selection based on the known size of the RBP. After gel extraction, linker sequences are attached, and the RBP is removed from the RNA with proteinase K treatment. Then, the RNA is amplified with RT-PCR and finally sequenced. The computational part of this assay includes mapping the CLIP reads to the genome and confidently identifying enrichment of reads (so-called peaks) in the data, representing the binding sites of the RBP of interest. For that purpose, CLIPper (Lovci

et al., 2013; Yeo et al., 2009), piranha (Uren et al., 2012), and the FAST-iCLIP pipeline (Koch, 2014) are the most widely used tools. Further analyses regarding the binding motifs can be performed using various online and command-line analysis tools, such as Homer (Heinz et al., 2010) or MEME (Bailey et al., 2009).

Over the years, multiple variations to the original HITS-CLIP protocol have emerged. For example, individual nucleotide resolution CLIP (iCLIP) (Huppertz et al., 2014)(Figure 1.4B) is based on the observation that crosslinking induces a sequence change in the RNA so that, in theory, the exact location of the crosslinking site can be determined. The use of intramolecular cDNA circularization solves the previous problem of RT stalling at the crosslinking site, and this protocol subsequently uses the stalling of the RT to detect the crosslinking site. Another remarkable variation of CLIP is enhanced CLIP (eCLIP)(Van Nostrand et al., 2016)(Figure 1.4D). It includes collecting a size-matched input (excision of unbound RNAs at the same location (=size-match) of the gel). Sequencing this size-matched input allows for a more precise peak calling in the computation part of the assay since the size-matched input data serves as background. Finally, photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP)(Danan et al., 2016; Hafner et al., 2010)(Figure 1.4C) incorporates photoreactive ribonucleoside analogs into nascent RNA transcripts. The RNAs labeled this way are excited using UV-A or UV-B, leading to a more efficient crosslinking with their RBP interacting partners. In addition to that, the defining feature of PAR-CLIP is the mutation introduced during reverse transcription at the exact crosslinking site leading to nucleotide-resolution identification of protein/RNA interaction sites.

As a result of the widespread use of CLIP technologies, we have comprehensive binding maps for more than 100 RBPs, mostly eCLIP and iCLIP format, publicly available, many of which are deposited on ENCODE (Davis et al., 2018; ENCODE Project Consortium, 2012). The list of publicly available CLIP datasets can be found in the

Appendix. CLIP technologies were used in Chapter Five of this thesis. Publicly available CLIP datasets were leveraged for a meta-analysis of protein-RNA interactions in Chapter Four of this thesis.

1.6 Global approaches for the quantification of alternative splicing and translation

Alternative splicing describes a co- and post-transcriptional process of assembling mature RNAs out of different combinations of coding and non-coding sequences transcribed from DNA into pre-mRNA. As discussed earlier, alternative splicing can be regulated and changed significantly by various RNA-binding proteins, most prominently splicing factors. Additionally, alternative splicing and the resulting different mRNA transcripts can have far-reaching consequences on post-transcriptional regulation of those mRNA transcripts, such as mRNA stability (Lewis et al., 2003), subcellular localization (Taliaferro et al., 2016), and translation (Sanford et al., 2004).

To quantify alternative splicing and its effects on mRNA translation, the Sanford lab developed a **subcellular fractionation and sequencing (Frac-Seq)** assay (Sterne-Weiler et al., 2013). This assay allows for tracking the enrichment or depletion of mRNA isoforms across different ribosomal fractions. It is an expansion of the ribosome footprinting technique used to track the translation of mRNAs (Ingolia et al., 2009). Isolating more than just one subcellular fraction, specifically, we identified the monosomal, light (p2-4), medium (p5-8), and heavy polysomal fraction (p9+), which allows for high-resolution tracking of the translational state of mRNA isoforms.

Cells or tissues are treated with cycloheximide to arrest the ribosomes in place before cell lysis. The cytosolic extracts are layered onto a sucrose density gradient and separated using velocity/density gradient separation, which separates the mRNAs based

on the number of ribosomes attached to the transcripts. Separate ribosomal fractions can then be isolated and purified using a gradient station. Each isolated fraction and the whole cytosolic extract were subjected to polyA selected RNA sequencing.

Frac-seq experiments result in very rich datasets that can be analyzed for alternative splicing and polysome association patterns. For that purpose, we used JunctionCounts (Figure 1.5), an in-house alternative to the splicing quantifier MISO (Mixture of Isoforms)(Katz et al., 2010). junctionCounts identifies pairwise alternative splicing events using provided transcript annotations as well as de novo assembled transcripts. The quantification of events is based on a simple and straightforward, exon-centric approach, using only reads that map to exon-exon and exon-intron junctions supporting the events. These junction counts are used to calculate the percent spliced in (PSI) for each event, which is the ratio of reads supporting the included isoform over the total number of reads supporting the event. This tool can identify a wider variety of alternative splicing events as its competitors (e.g., SplAdder (Kahles et al., 2016), MISO (Katz et al., 2010), SUPPA2 (Trincado et al., 2018), ASTALAVISTA (Foissac and Sammeth, 2007)). Also, the pipeline utilizes CAT annotations (Fiddes et al., 2018) to make reliable cross-species comparisons of alternative splicing events.

1.7 Conclusion of Introduction

The following chapters of this thesis present the work from multiple projects revolving around protein-RNA interactions that I have contributed to throughout my time as a graduate student. Chapters Two and Three cover investigations of the hypothesis that cis- and trans-regulatory differences in mRNA transcripts brought on by alternative splicing can affect the regulation of mRNA translation. Chapter Four contains a project investigating the interactions between microRNAs and regulatory RBPs for which I conducted a meta-analysis of publicly available protein/miRNA

interaction data that gives way to an in-depth experimental investigation of SRSF1 interacting with mir10b. Chapter Five covers investigations into the RNA-binding specificity of the oncofetal RNA binding protein called IGF2BP3 in leukemogenesis. Chapter Six covers BiocSwirl, an R package and teaching platform that helps bench-level scientists learn the right amount of programming and statistical understanding to do their own data analysis.

Finally, Chapter Seven will summarize and discuss the findings presented in the previous chapters in a broader view and look at the future directions of the projects mentioned above.

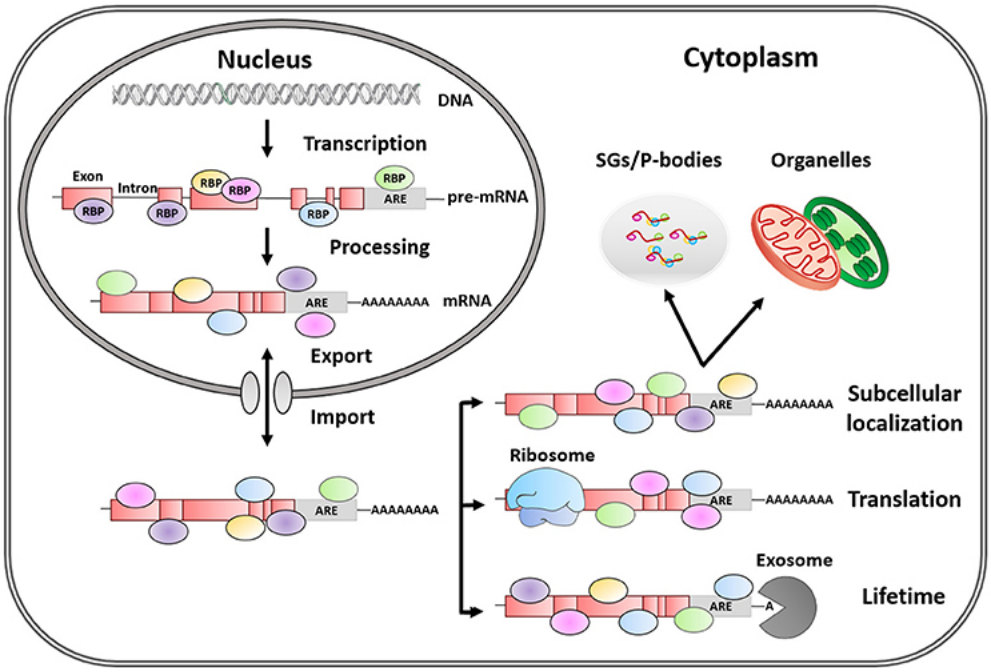

Figure 1.2 Regulation of mRNA fate by RNA-binding proteins throughout its life cycle.
Pre and mature mRNAs are covered in different sets of RNA-binding proteins throughout their whole life cycle. The sum of RBPs on an mRNA is referred to mRNPs. These mRNPs can change in composition and affect the mRNA fate within a cell, from transport, localization to mRNA decay. Adapted from Garcia-Mauriño et al.(García-Mauriño et al., 2017)

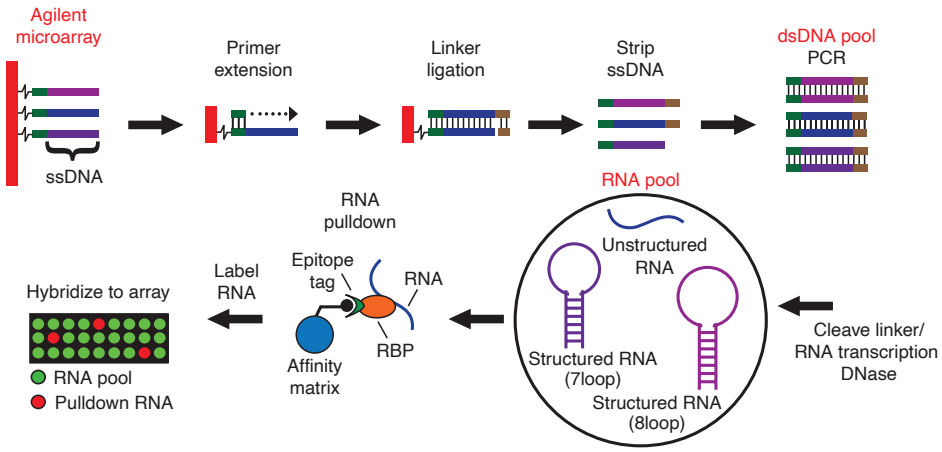

Figure 1.3 Schematic of RNAcompete experimental procedure.
RNAcompete is an in vitro approach to investigate protein-RNA interactions, both sequence and structure-dependent.The initial RNA pool with all possible 10-base sequences and all possible 7 and 8 base stem-loops is incubated with a GST-tagged protein of interest. A GST pulldown is performed, the recovered RNA labeled with Cy5, some of the initial RNA pool labeled with Cy3, and both are co-hybridized to a microarray. Computationally, the enrichment of pulled-down RNA to the original RNA pool can be determined. Schematic adapted from Ray et al., 2009

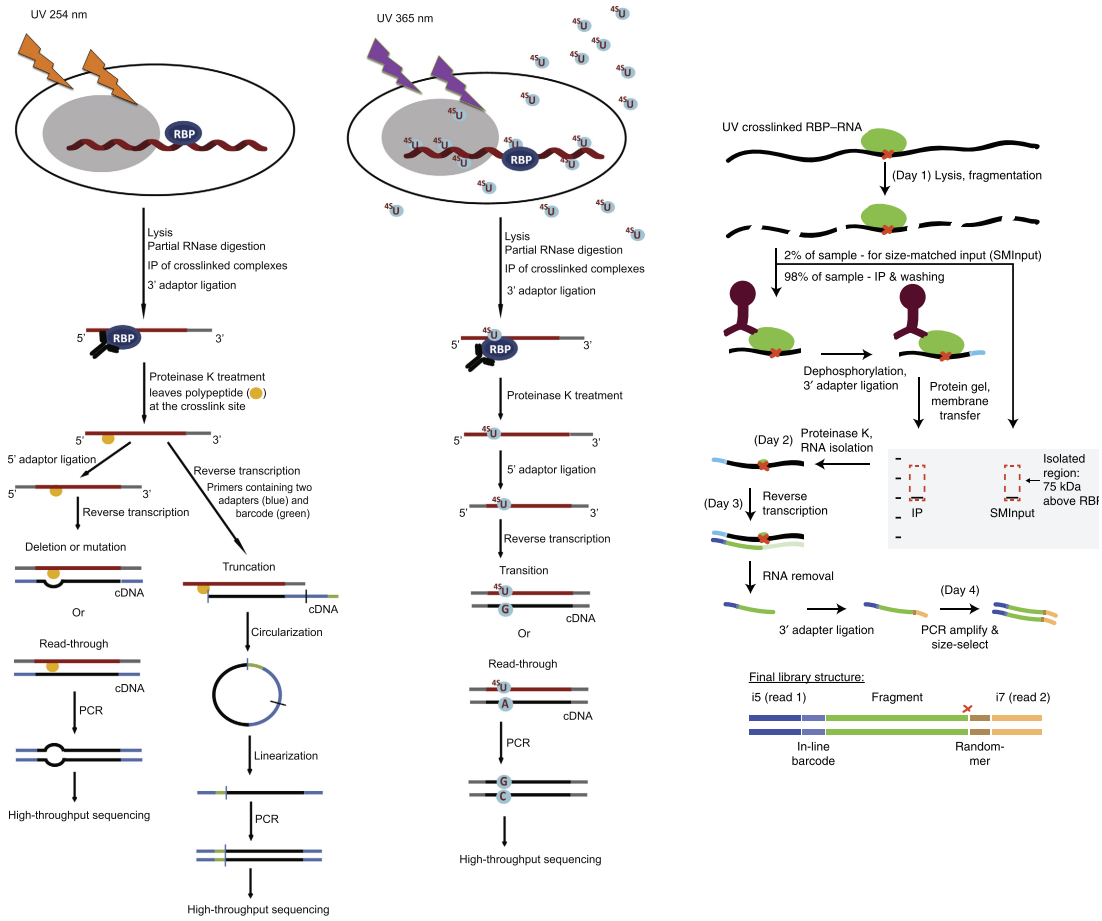

Figure 1.4 Overview over different variations on the CLIP protocol.
Outline of different CLIP procedures. A) HITS-CLIP, B) iCLIP, C) PAR-CLIP, and D) eCLIP. The different CLIP procedures vary mostly in the way the sequencing libraries are prepared. However, PAR-CLIP differs in crosslinking and the use of photoreactive ribonucleoside analogs. The eCLIP protocol includes the collection of size-matched input data for computational normalization. The images were adapted from Li et al., 2014b and Van Nostrand et al., 2016.

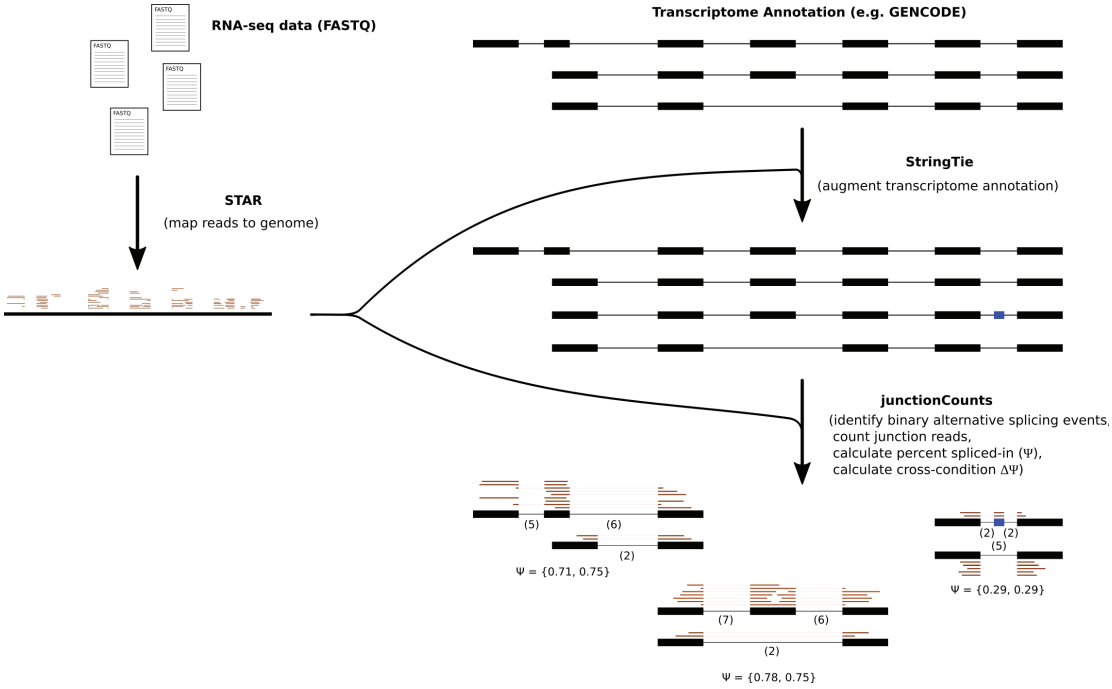

Figure 1.5 Schematic of JunctionCounts.
JunctionCounts uses Stringtie to identify unannotated transcripts and uses CAT transcriptomes, together with the stringtie merge command to generate the final transcriptomes for each species. Pairwise alternative splicing events were identified by pairwise comparison of all transcripts with at least one exon-intron junction in common. An alternative event was defined to be a set of exons unique to one transcript that are surrounded by two exon-intron junctions common to both transcripts or by one junction and the transcript terminus. Alternative splicing events were quantified by counting the number of reads supporting the exon-exon junction and the number of reads supporting the exon-intron junction of each event. PSI (percent spliced in) values were calculated as the ratio of number of reads supporting the included isoform to the number of reads supporting both isoforms. Image created by Andrew Wallace.

2. Chapter 2: Isoform-specific translational control is evolutionarily conserved in primates

Draper[1*], Philipp[1*], Dargyte[1], Wallace[1], Katzman[2] and Sanford[1+]

[1] University of California Santa Cruz, Department of Molecular, Cellular and Developmental Biology, Department of Chemistry and Biochemistry CA, 95064, USA

[2] Center for Biomolecular Science and Engineering, University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95060, USA

* *These authors contributed equally to this work.*

+ *Corresponding Author*

2.1 Abstract

The process of alternative splicing expands protein coding capacity and post-transcriptional regulatory mechanisms. We recently demonstrated that alternative splicing influences mRNA translation. Here we explore alternative splicing coupled translational control (AS-TC) across primate cell lines. We used a Fractionation-Sequencing (Frac-Seq) approach to identify polyribosome associated mRNA isoforms. We discovered orthologous AS-TC events with either conserved or species-specific translation patterns. Exons sequences associated with conserved sedimentation profiles show strong conservation across vertebrates. Orthologous exons with divergent sedimentation profiles drive species-specific expression of luciferase reporters. Together these data show that cis-acting elements regulate AS-TC across primates species.Introduction

2.2 Introduction

Precursor messenger RNA (pre-mRNA) splicing is required for the accurate expression of most protein-coding genes. The removal of Intervening sequences (introns) and ligation of protein coding sequences (exons) is catalyzed by a macromolecular

complex called the spliceosome. The spliceosome assembles de novo on every newly synthesized intron using a handful of essential cis-acting RNA elements including the 5' and 3' splice sites, the branch point sequence and myriad auxillary splicing regulatory elements. Exonic and intronic splicing enhancer or silencer elements (ESE and ESS, respectively) play significant roles in determining whether or not specific exons are included in the mature mRNA transcript. Splicing regulation occurs in part, through the interpration of this complex, context-dependent cis-regulatory landscape by RNA binding proteins (RBPs). Many RBPs, such as the serine- and arginine-rich (SR) proteins, not only influence spliceosome assembly, but also regulate subsequent steps of the gene expression pathway (Howard and Sanford, 2015). These observations raise the intriguing hypothesis that RBPs may couple pre-mRNA splicing to downstream processes such as mRNA export, translation and decay.

Alternative splicing (AS) augments the expression of most human genes by generating transcript isoforms from different combinations of exons sequences. Perhaps, the most obvious impact of alternative splicing is to modify the primary structure of polypeptides (Nilsen and Graveley, 2010). However, AS also influences post-transcriptional regulation of messages in unexpected ways (Figure 2.1a). Alternative splicing events that induce premature termination codons (PTCs) can trigger nonsense-mediated decay (AS-NMD). In many cases, AS-NMD appears to be part of an evolutionarily conserved regulatory mechanism for fine tuning the expression of splicing factors (Lareau and Brenner, 2015; Lewis et al., 2003; Pan et al., 2006; Saltzman et al., 2008). Recently, several groups, including our own, discovered that mRNA isoforms can exhibit differential polyribosome association, suggesting that alternative splicing can be coupled to translational control (AS-TC). Thus, alternative splicing not only expands the protein coding capacity of eukaryotic genes but may also influence the cytoplasmic

fate of the resulting mRNA isoforms.

Comparative genomics and transcriptomics are powerful approaches for studying the evolution of gene regulatory mechanisms. For example, the expression levels of orthologous genes are well conserved between tissues across distantly related vertebrate species (Barbosa-Morais et al., 2012; Calarco et al., 2007; Mazin et al., 2018; Merkin et al., 2012). By contrast, exon usage is poorly conserved across the same species and tissues, suggesting that tissue-specific alternative splicing patterns are rapidly evolving. Indeed the prevalence of alternative splicing varies between species and is highly abundant in the primate nervous system (Barbosa-Morais et al., 2012). Comparative transcriptomic and proteomic analysis of primate cells also demonstrated that although steady state protein levels are similar for orthologous genes, mRNA levels can vary dramatically. This suggests that overall mRNA levels might evolve under less rigorous evolutionary pressure compared to protein expression levels (Khan et al., 2013).

Many alternative exon sequences associated with AS-NMD are ultraconserved, suggesting that regulatory elements and function of these mRNA isoforms are under strong purifying selection. Indeed, ablation or programmed mis-splicing of AS-NMD results in growth defects and loss of tumor suppression in cell culture and mouse embryo models (Jung et al., 2015; McIlwain et al., 2010).

Our previous work (Sterne-Weiler et al., 2013), as well as others' (Floor and Doudna, 2016; Wong et al., 2016), demonstrates that alternative splicing of mRNA isoforms can lead to differential translational control of these isoforms contributing to the discrepancy between mRNA and protein expression levels. By contrast to AS-NMD, the physiological and evolutionary significance of AS-TC is unknown.

In this manuscript, we use comparative genomics (transcriptomics) to test the extent of conservation of AS-TC and to identify functionally important exons that contribute to the coupling. We compared Frac-seq data (subcellular fractionation and high-throughput RNA-sequencing)(Sterne-Weiler et al., 2013) of human, chimpanzee, and orangutan induced pluripotent stem cells (iPSCs). Using these Frac-seq data allows us to identify mRNA isoforms with differential polyribosome association. Our data show that the process of AS-TC is conserved across all three cell lines. We identified alternative splicing events with conserved translational control as well as events with species-specific regulation. We validated the ability of sequence elements associated with isoform-specific polysome association to affect mRNA translation *in vivo* using Luciferase reporters. We further showed that events with conserved translational control show higher sequence conservation specifically in cassette exons (and alternative first exons), indicating the functional relevance of these exons. Taken together our data suggest a conserved mechanism of AS-dependent regulation of mRNA translation

2.3 Results

2.3.1 Frac-Seq analysis of primate iPSCs

We previously presented Frac-seq (Figure 2.1A) as a method to determine the association of alternative mRNA isoforms with the polyribosome. In order to test the conservation of alternative splicing coupled to translational control, we applied the Frac-seq methodology to human, chimpanzee, and orangutan iPSCs. We identified and quantified approximately 3000-600 events spanning a variety of different event types using previously established analysis pipelines (Figure 2.1B). Within each cell line, we identified over 1000 events with PSI values differing between fractions by more than 10%. We further separated these into events that generate non-productive NMD isoforms (AS-NMD) and events maintaining the integrity of their open reading frame. The

latter we consider alternative splicing events that are potentially implicated in translational control (AS-TC events, Figure 2.1C). Between all three cell lines, we identified orthologous alternative splicing events with differing cellular fate (Figure 2.1D) and 362 orthologous alternatively spliced events that are coupled to translational control in all three cell lines (Figure 2.1E).

2.3.2 Identification of orthologous mRNA isoforms with similar or species-specific polyribosome association

To determine if orthologous AS-TC events exhibit similar polyribosome association, we calculated the cumulative interspecies distance of mean PSI between species, across all fractions. Calculating this cumulative distance for all orthologous AS-TC events results in a right-skewed distribution where the far left and right tails represent the AS-TC events with the most and least conserved sedimentation profile, respectively. We visualized these subsets of events in correlation heatmaps. The colors are indicating the Spearman correlation of PSI values in events with low cumulative distance (conserved sedimentation profiles)(Figure 2.2A) and in events with high cumulative distance (species-specific sedimentation profiles)(Figure 2.2B). Interestingly, in events with conserved sedimentation profiles, the same fractions from the different cell lines cluster together neatly, indicating the PSI values in these fractions are more similar to each other than to the other fractions within the same species. Consequently, in events with species-specific sedimentation profiles, the different fractions of each species cluster together, indicating more similarity within the species than the fractions. For example, C12orf29 exhibits a similar sedimentation pattern between species (Figure 2.2C), whereas CNN1 alternative splicing generates isoforms with species-specific sedimentation patterns (Figure 2.2D). In events with species-specific sedimentation patterns, the PSI values across the polysomal fractions differ between species, despite a similar

nuclear output represented by the PSI value in the cytoplasmic fraction.

2.3.3 Evolutionary conservation of orthologous AS-TC events

To test if events with conserved sedimentation patterns also show sequence conservation, we investigated the sequence conservation of AS-TC events compared to non-AS-TC exons. Skipped exons that are not associated with AS-TC are much less conserved compared to their flanking exons. By contrast, exons linked to AS-TC exhibit significantly higher Phastcons scores and are more similar to their flanking exons. Further, AS-TC exons that have conserved sedimentation profiles between species have elevated phastCons scores relative to the other classes (Figure 2.3A). Similar effects can be observed for the PhastCons scores of alternative first AS-TC and non-AS-TC exons (Figure 2.3B). We further observe a higher sequence conservation in the distal alternative first exons compared to the proximal alternative first exons. The high degree of sequence conservation within AS-TC exons suggests the presence of functional elements within these exons that influence polyribosome association.Functional characterization of primate AS-TC events.

2.3.4 Single nucleotide variants in orthologous AS-TC exons influence translation

In order to test the hypothesis that sequence differences associated with isoform-specific sedimentation patterns regulate mRNA translation, we created luciferase reporters for two skipped exon events (Figure 2.4A-G) and three alternative first exon events (Figure 2.4H-Q) from different genes exhibiting AS-TC and performed qPCR experiments . Figure 2.4A shows the schematic for the skipped exon luciferase reporters, where either the human (green) or chimpanzee (orange) skipped exon event were inserted inframe upstream of the firefly luciferase. Frac-seq analysis revealed differential sedimentation profiles for human and chimp isoforms from the GGCX (Figure

2.4B) and SUMF2 (Figure 2.4E) genes. Both pairs of orthologous exon sequences differ at a single position (Figure 2.7/Supplemental Figure 3)

Surprisingly, luciferase reporter constructs containing the chimpanzee-derived GGCX sequence promoted significantly higher luciferase activity compared to the human sequences (Fig 4C, F). This effect was likely due to increased translational efficiency, as the steady state mRNA levels are significantly higher for reporters containing the human SE compared (Fig 4D, G).

We also tested the ability of alternative first exon sequences associated with isoform-specific sedimentation to affect mRNA translation. We created pairs of luciferase reporters from three genes exhibiting isoform-specific sedimentation in human and chimpanzee iPSCs corresponding to the proximal (blue) or distal (red) first exon (Figure 4H). We chose two genes exhibiting species-specific sedimentation profiles (Figure 4I,L) and one gene with conserved sedimentation (Figure 4O). Interestingly, AFE reporters corresponding to the chimpanzee orthologs of CNN1 and UGP2 resulted in stronger luciferase activity compared to their human counterparts. By contrast, orthologous sequences from the MAD2L2 gene resulted in similar expression levels. In all cases, the steady state mRNA levels for each reporter were similar, suggesting that expression differences were likely due to translation.

2.4 Discussion

In this study we use comparative transcriptomic analysis of primate iPSCs to investigate post-transcriptional control of gene expression. We discovered that the process of AS-TC is conserved across all three cell lines. We identified orthologous alternative splicing events between human, chimpanzee, and orangutan iPSCs with isoform-specif-

ic ribosome engagement patterns which suggest coupling of AS to translational control. We classified these events as conserved or species-specific based on their distribution across sucrose gradients. Skipped exons and alternative first exons that are predicted to couple AS with translational control, show a stronger sequence conservation than canonical skipped and alternative first exons. This observation suggests that cis-acting element function in isoform-specific translational control. Sequence elements that were associated with isoform-specific sedimentation profiles based on our analysis of the Frac-seq data influenced the expression of luciferase reporters *in vivo*.

Alternative splicing alters the primary sequence and therefore the landscape of cis and trans-acting translational regulators leading to differential fate of mRNA isoforms. Particularly 5'UTRs, as well as coding sequences, of mRNA transcripts are known to contain numerous cis-acting regulators of translation(Gebauer et al., 2012; Pfeiffer et al., 2012)) such as upstream open reading frames (uORFs), internal ribosome entry sites (IRES), and RBP-binding sites. Their differential inclusion in the final transcripts could lead to drastic variations in the trans-acting landscape and the composition of the mRNPs of the final transcript, leading to isoform-specific translational control. Changes in primary sequence could also contribute to differences in the secondary structure of the mRNA isoforms. RNA secondary structures have previously been shown to regulate mRNA translation as well. For example, more stable secondary structures near the start codon require more energy for unfolding, slowing down translation initiation (Kudla et al. 2009). Alternative splicing within the coding sequence could further change the codon composition of the mRNA transcripts resulting in differential translation rates. Without the ability to calculate the translational efficiency as per Ingolia et al, a potential explanation for differential ribosome association of isoforms could also be the difference in length of the coding region of transcripts.

Comparative genomics studies reveal functional regulatory elements and the evolution of alternative splicing (coupled with translational control). The high prevalence and complexity of alternative splicing in primates makes them an excellent system for comparative genomics studies. Previous comparative genomics studies (Barbosa-Morais et al., 2012; Merkin et al., 2012) have shown that most species-specific splicing patterns are caused by cis-regulatory elements. Comparing human, chimpanzee, and orangutan iPSCs allowed us to identify events of AS-TC that are highly conserved and therefore likely to be of functional importance as well as highly species-specific events that can give insight into the evolution of AS and it's coupling to translational control.

Alternative splicing has previously been shown to be implicated in developmental regulation (Baralle and Giudice, 2017; Su et al., 2018). Species-specific, and therefore evolutionary, differences in the coupling of alternative splicing and translational control could indicate a potential role for AS-TC in developmental regulation of gene expression. It would, therefore, be interesting to conduct a time course experiment collecting Frac-seq datasets from different stages of development, e.g. different stages of neuronal differentiation, and identify changes in AS-TC over developmental time. We would expect to find similar changes in AS-TC throughout the different stages of development compared to the evolutionary changes.

Our data shows that the coupling of alternative splicing to translational control is conserved across multiple primate cell lines, representing approximately 13 million years of primate evolution. We were able to demonstrate that alternative exons implicated in AS-TC are more conserved than canonical exons, indicating a functional role in regulating AS-TC. Through luciferase reporters, we showed that sequence elements

associated with AS-TC, are able to influence expression as predicted *in vivo*. Taken together, our data supports the coupled fate hypothesis and points to a conserved mechanism coupling alternative splicing and translational control.

2.5 Materials & Methods

2.5.1 iPSC generation and culture

Integration-free human, chimpanzee, and orangutan induced pluripotent stem cells (iPSC) were generated from primary fibroblasts by Field et al. (Field et al., 2017) as previously published.

2.5.2 Fractionation, polyribosome profiling, RNAseq

Frac-seq experiments were performed as previously published (Sterne-Weiler et al., 2013) using human, chimpanzee, and orangutan IPSCs. Cytosolic extracts from cell lines/tissues are fractionated by sucrose gradient centrifugation. We collected the total cytosolic lysate, the monoribosomal fraction (80s), as well as light (P2-4), medium (P5-8), and heavy (P9+) polyribosomal fractions. We then sequenced the polyA+ selected RNA from these fractions. RNA sequencing was performed on the previously mentioned fractions using paired end 125x125 sequencing, resulting in approx 75-150M reads per sample (Figure 2.5/ Supplementary Figure 1) with approximately 40-50% junction reads per sample.

2.5.3 Mapping of Illumina short read RNAsequencing

The reads were mapped to the human genome assembly hg38, the chimpanzee (*Pan troglodytes*) genome assembly panTro6, and the sumatran orangutan (*Pongo abelii*) genome assembly ponAbe3 using STAR v2.7 (Dobin et al., 2013). Repeat sequences were masked by mapping to repeatMasker sequences (Smit et al. RepeatMasker Open-4.0 at http://repeatmasker.org) using Bowtie2 (Langmead et al., 2009) . PCR Duplicate

removal was performed by collapsing fragments with common start and end positions and CIGAR strings using in house scripts. All data collection and parsing was done with bash and python2.7. Statistical analyses and data visualization were performed using R programming language version 3.5.1.

2.5.4 Identification and quantification of orthologous alternative splicing events

The reads were mapped to the human genome assembly hg38, the chimpanzee (*Pan troglodytes*) genome assembly panTro6, and the sumatran orangutan (*Pongo abelii*) genome assembly ponAbe3 using STAR v2.7 (Dobin et al., 2013). Repeat sequences were masked by mapping to repeatMasker sequences (Smit et al. RepeatMasker Open-4.0 at http://repeatmasker.org) using Bowtie2 (Langmead et al., 2009) . PCR Duplicate removal was performed by collapsing fragments with common start and end positions and CIGAR strings using in house scripts. All data collection and parsing was done with bash and python2.7. Statistical analyses and data visualization were performed using R programming language version 3.5.1.

2.5.5 Cross-fraction comparison/ Cross-species comparison / Identification of conserved and species-specific orthologous events

All events identified were filtered to be supported by at least 15 junction reads (per comparison). Alternative splicing events undergoing translational control (AS-TC) events were defined as events with a change in PSI value (delta PSI) between any two adjacent fractions of at least 0.1. Consequently, alternative splicing (AS) events not undergoing translational control were defined as events with a minimum PSI > 0 and delta PSI < 0.1. Alternative splicing events leading to nonsense-mediated decay (AS-NMD) were identified using *in silico* translation of raw transcripts and subsequent identification of premature termination codons (PTCs) (technically CDSinsertion)

.

For estimating the difference/conservation of the polysome association pattern we calculated the Manhattan distances for each event between each two species. The Manhattan Distance is the sum of differences in mean psi between two species across all fractions. Min/max normalization of the Manhattan distance allowed us to identify events with overall different sedimentation profiles as opposed to events with similar sedimentation profiles at a different y-axis intercept. We ranked all AS-TC and AS-NMD events based on their min/max normalized Manhattan distance and used the top and bottom 10% (= 350 events) for further analysis, considering them the least and most conserved set of events respectively.

2.5.6 Determination of sequence conservation

To determine the sequence conservation of ASTC events with conserved or species-specific sedimentation profiles as well as AS events, phastCons (Siepel, 2005; Siepel and Haussler) scores were obtained from the UCSC genome browser (Kent et al., 2002). For skipped exons events, phastCons scores were obtained for 100nt windows around both splice sites of the cassette exon as well as around the upstream 5'ss and the downstream 3'ss. For alternative first exon events, the scores were obtained for the 300 nucleotides downstream of both transcription start sites as well in 100nt windows around the 5'ss of the two alternative first exons. The scores were visualized with local nonlinear smoothing using a generalized additive model.

2.5.7 RNA purification and RT-qPCR

Total RNA was isolated using the Direct-zol RNA MiniPrep Kit (Zymo Research). 800ng of the RNA were treated with RQ DNase (per protocol). The DNase (Promega) treated RNA was reverse transcribed using the High-Capacity cDNA reverse transcriptase kit (Applied Biosystems). 1:200 dilutions of the cDNA were made. For the qPCR,

we used Luna 2x SYBR premix (total volume 20µl per reaction), 0.25 nM primers and 5µl diluted cDNA. qPCR was performed on QuantStudio 3 Real-Time PCR System (Applied Biosystems, Thermo Fisher) according to MIQE guidelines (Bustin et al., 2009a).

2.5.8 Luciferase Reporters

Luciferase activity was assayed 24 hours post transfection using Dual-Glo Luciferase Assay System (Promega). For a 6 well plate, transfections were performed with lipofectamine with either 2µg pLCS plasmid (previously published Sanford et al.) plus 125ng control plasmid (rluc) (for skipped exon events) or 1µg p5UTR (pLightSwith_5UTR, from Switch Gear) 1µg plus 250ng control plasmid (pmir) (for alternative first exon events) per well.

Dissertation author contribution

JP: Design and execution of experiments, writing of manuscript

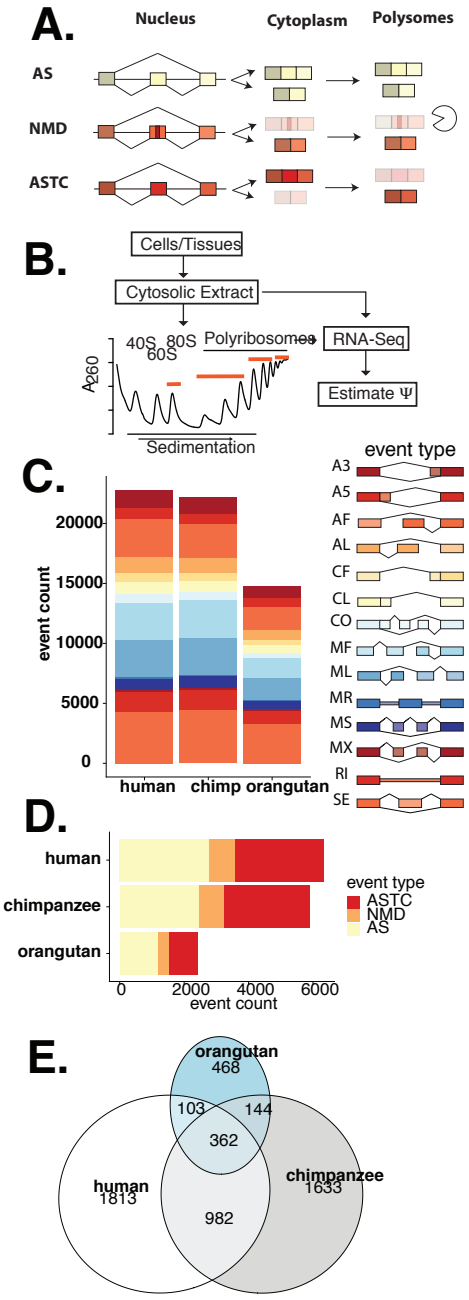

Figure 2.1 Frac-seq reveals polyribosome associated mRNA isoforms.
A) Alternative splicing can influence multiple post-transcriptional regulation pathways. A) Frac-seq (subcellular fractionation and subsequent sequencing of polyA+ selected RNA from fractions) was performed on human, chimpanzee, and orangutan iPSCs. RNA from the total cytosolic lysate, the mono-ribosome (80s), the light (P2-P4), medium (P5-P8), and heavy polyribosome (P9+) was sequenced. B) Identification and quantification of alternatively spliced events was performed using junctionCounts. This pipeline allowed the identification of 14 different event types. C) Alternative splicing events were further classified into AS, ASTC, and AS-NMD events The proportions of these three event groups are comparable between the three cell lines. D)) Out of the events categorized as AS-TC, over 300 events were identified in all three cell lines.

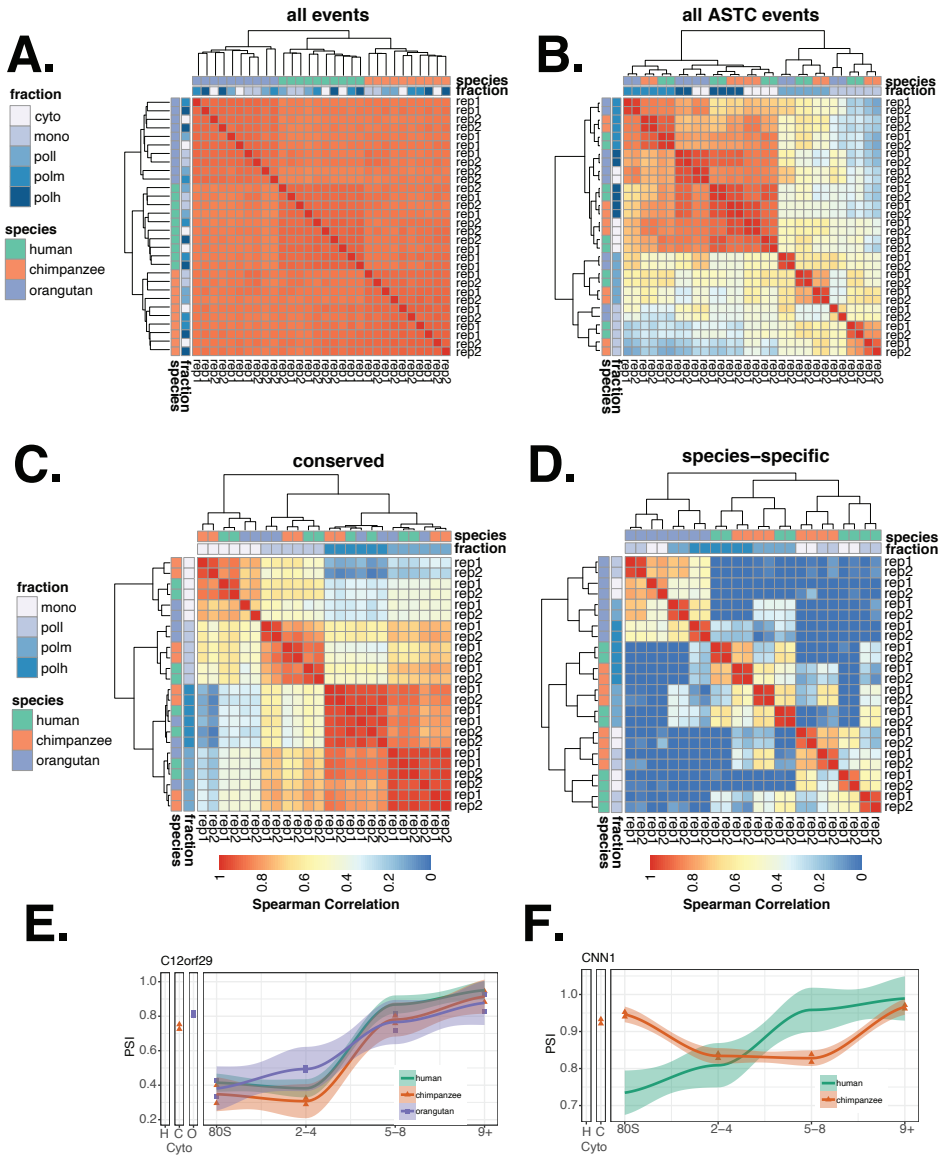

Figure 2.2. Orthologous AS-TC events exhibit either conserved or species-specific sedimentation profiles.

A) Heatmap of Spearman correlation of PSI values of all identified orthologous events. The columns and rows represent the total cytosolic lysate and the 4 subcellular fractions in each cell line. The colors represent the Spearman correlation of PSI values between pairs of fractions (red = high correlation, blue = low correlation). B) Heatmap of Spearman correlation of PSI values of all orthologous ASTC events. C+D) Heatmaps of Spearman correlation of PSI values of orthologous ASTC events. C) The events in this exhibit sedimentation profiles consistent across all three cell lines as shown in the example in panel E. D) The events in this heatmap exhibit species-specific sedimentation profiles as shown in the example in panel F. E) The skipped exon event within C12orf29 is an example of an alternative splicing event with conserved sedimentation profiles across all three cell lines. F) The alternative first exon event of CNN1 is an example of an alternative splicing event with species-specific sedimentation profiles.

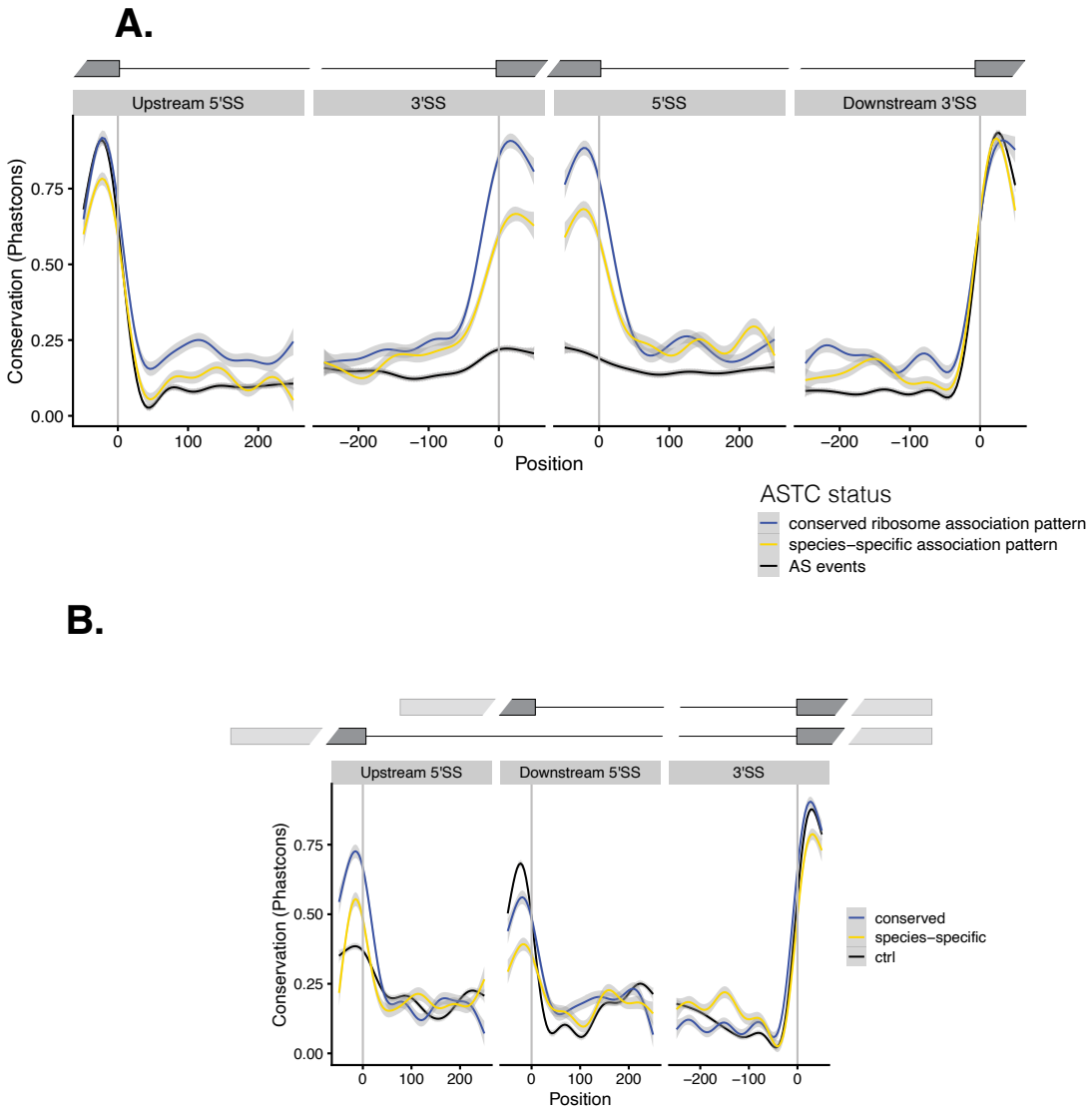

Figure 2.3 Alternative splicing events with sedimentation profiles consistent across species show higher sequence conservation.

A) Sequence conservation of exon/intron boundaries of ASTC and AS skipped exon events represented by phastCons (phastcons scores. A lower score indicates less conservation. ASTC events with conserved sedimentation profiles in blue, ASTC events with species-specific sedimentation profiles in yellow. B) Sequence conservation of exon/intron boundaries of ASTC and AS alternative first exon events represented by phastCons scores.

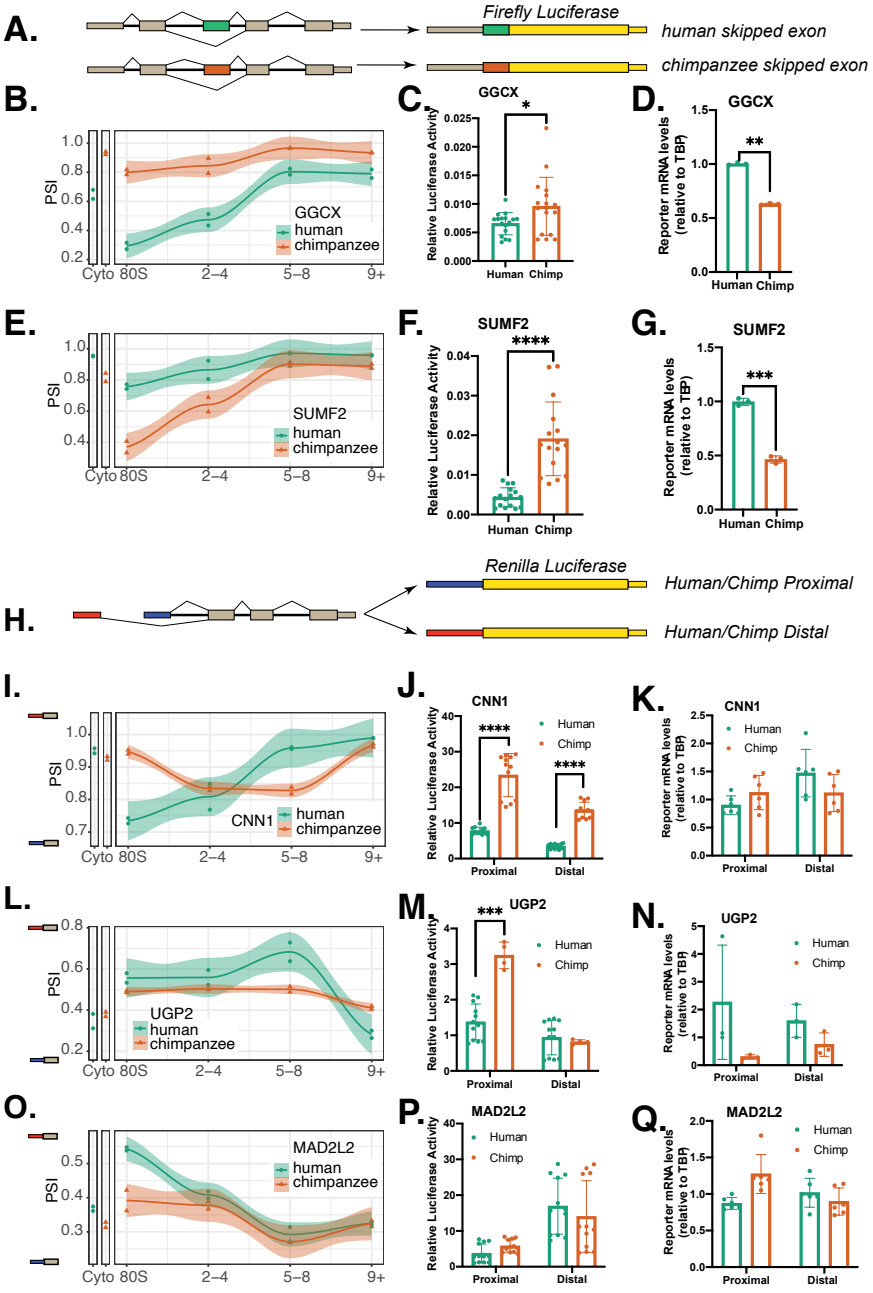

Figure 2.4 AS-TC cassette exons drive isoform-specific expression.

A) Schematic diagram of the pairs of luciferase reporter constructs containing either the human (green) or chimpanzee (orange) cassette exon from different genes exhibiting AS-TC. B,E) Polyribosome sedimentation profiles for isoforms from the GGCX and SUMF2 genes (respectively) in human and chimpanzee iPSCs. C,F) Dual luciferase assays in HEK cells for the two skipped exon events. D,G) qPCR of SE reporter mRNAs. H) Schematic diagram of the pairs of luciferase reporter constructs containing either the proximal (dark blue) or distal (red) alternative first exons from different genes exhibiting AS-TC. I,L,O) Polyribosome sedimentation profiles for isoforms from the CNN1, UGP2, and MAD2L2 genes (respectively) in human and chimpanzee iPSCs. J,M,P) Dual luciferase assays in HEK cells for the three alternative first exon events. K,N,Q) qPCR of AF reporter mRNAs.

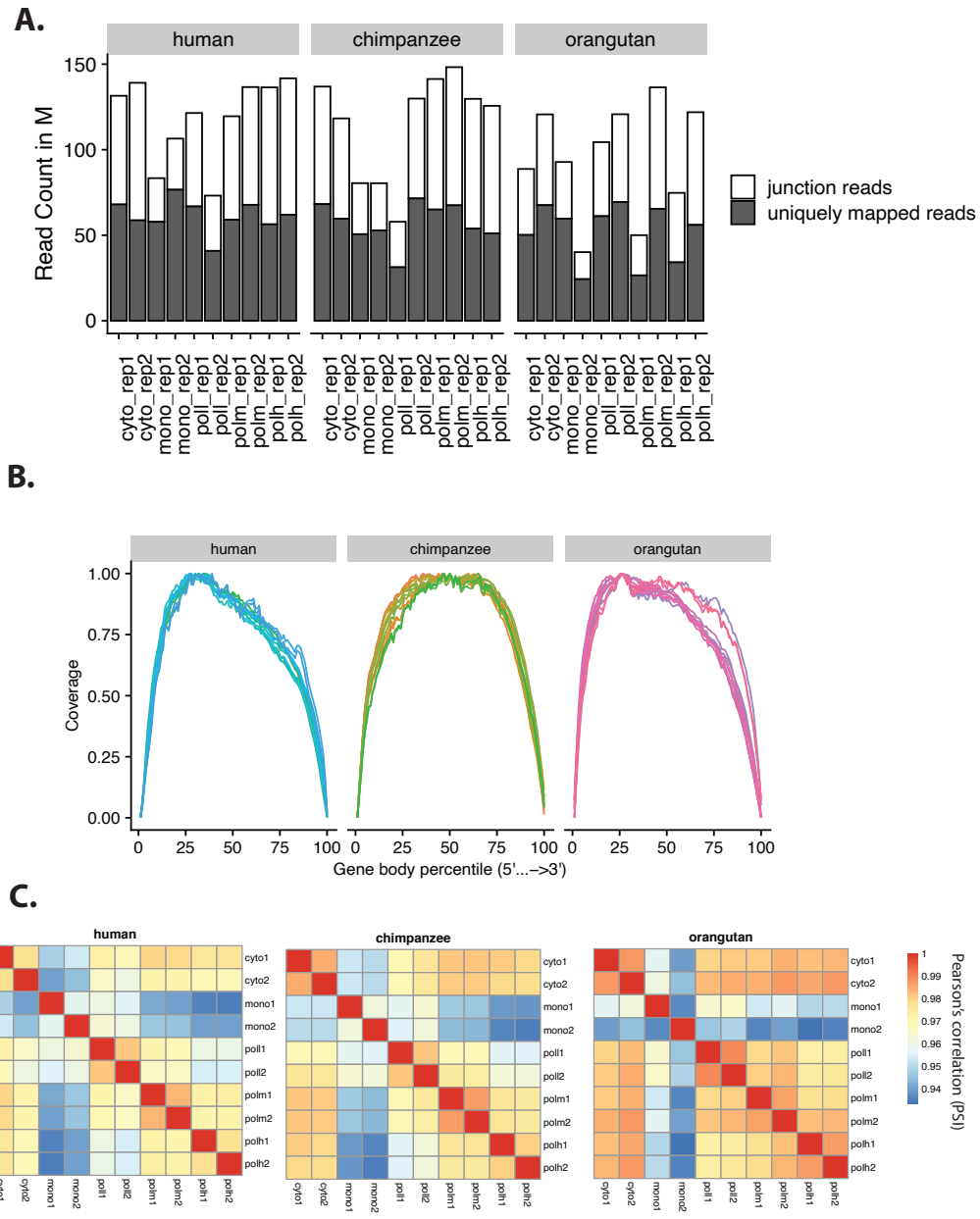

Figure 2.5 [Supplementary Figure 1] Quality control of Frac-seq data
A) Mapping total read counts and junction counts. B) Gene body coverage for each replicate in each of the species. C) Correlation of PSI values between replicates for each species.

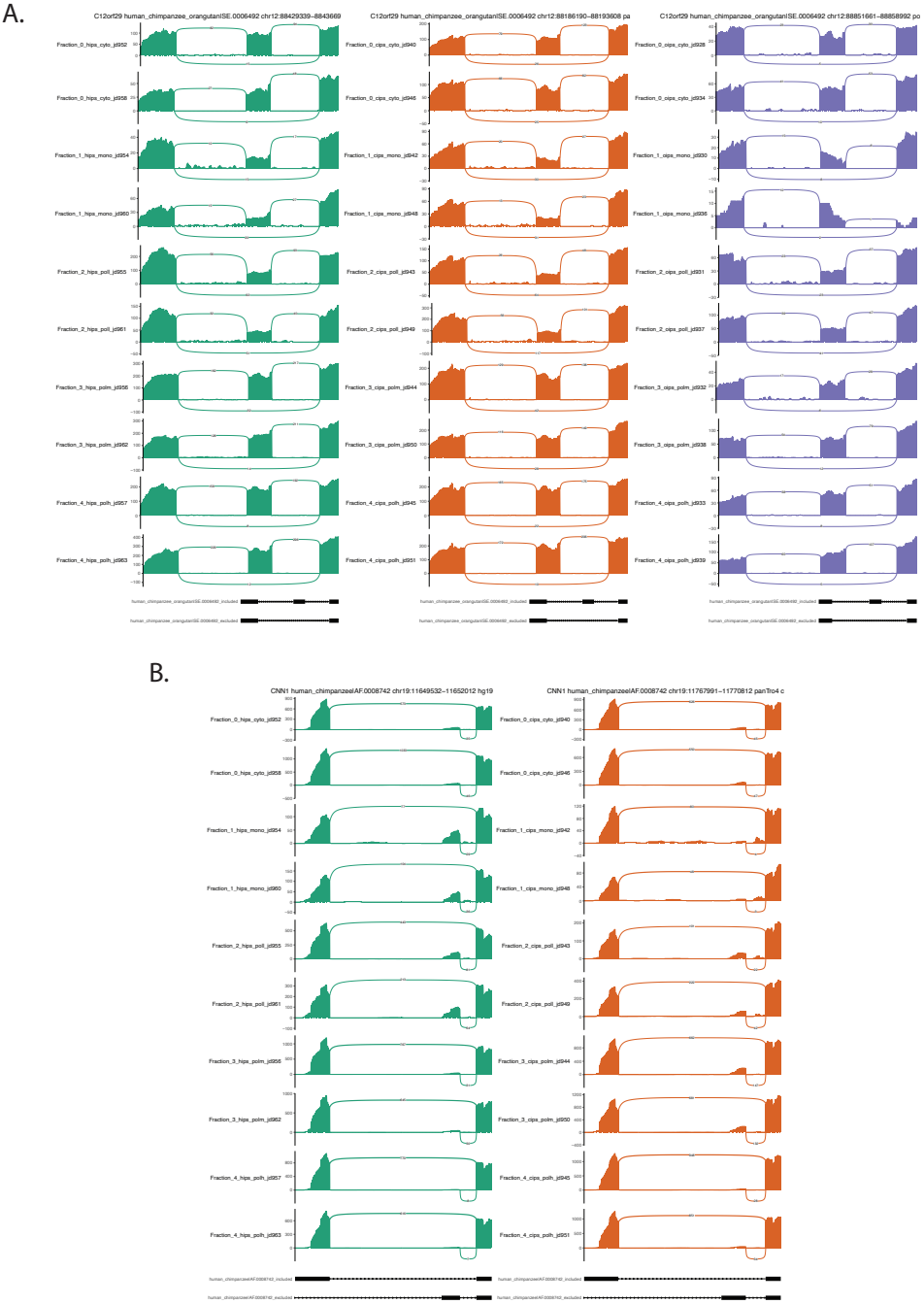

Figure 2.6 [Supplementary Figure 2] Sashimi plots of representative conserved and species specific ASTC events.

A) Sashimi plots for all fractions and all three species visualizing the sequencing reads mapping to the splice junctions of the conserved skipped exon ASTC event C12orf29 (Figure 2.2E) in human (green), chimpanzee (orange), and orangutan (purple). B) sashimi plots for all fractions and two species visualizing the sequencing reads mapping to the splice junctions of the species-specific alternative first exon ASTC event CNN1 (Figure 2.2F) in human (green), chimpanzee (orange), and orangutan (purple).

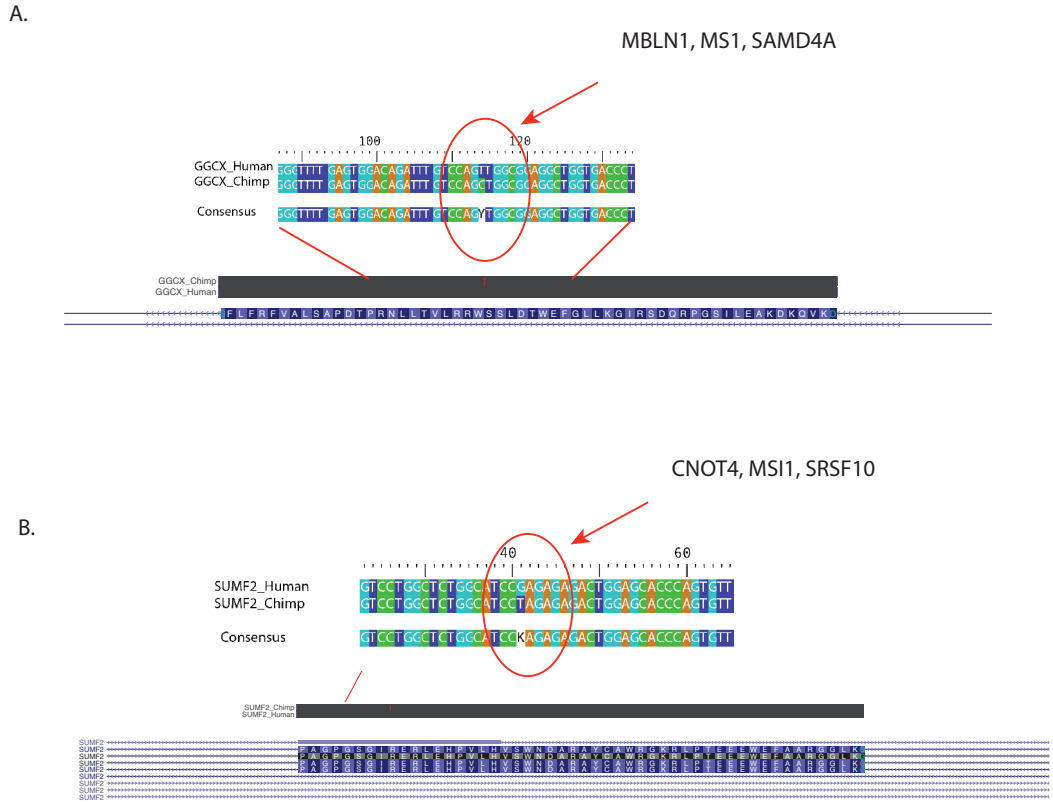

Figure 2.7 [Supplementary Figure 3] Pairwise Alignments of exons tested in luciferase reporters showing subtle differences that might regulate AS-TC.
Pairwise Alignments of skipped exons tested in luciferase reporters showing subtle differences that might regulate AS-TC. A) Pairwise alignment of human and chimpanzee GGCX skipped exon sequences with mismatches highlighted and Genome Browser Blat results for GGCX human and chimpanzee skipped exon sequences. B) Pairwise alignment of human and chimpanzee SUMF2 skipped exon with mismatches highlighted and Genome Browser Blat results for SUMF2 human and chimpanzee skipped exon sequences.

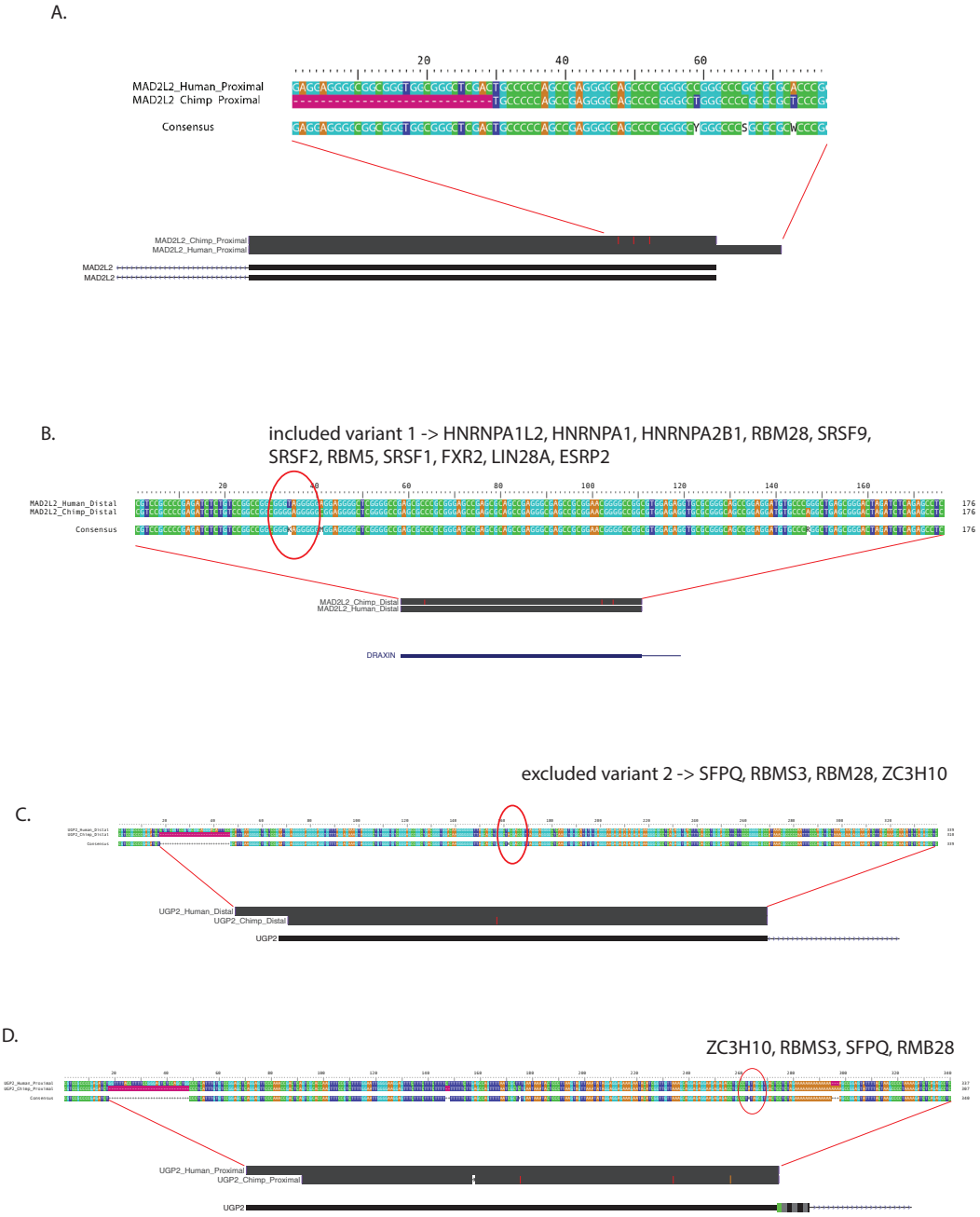

Figure 2.8 [Supplementary Figure 4]Pairwise Alignments of alternative first exons tested in luciferase reporters showing subtle differences that might regulate AS-TC.

A) Pairwise alignment of human and chimpanzee MAD2L2 proximal first exon sequences with mismatches highlighted and Genome Browser Blat results for the same sequences. B) Pairwise alignment of human and chimpanzee MAD2L2 distal first exon sequences with mismatches highlighted and Genome Browser Blat results for the same sequences. C) Pairwise alignment of human and chimpanzee UGP2 proximal first exon sequences with mismatches highlighted and Genome Browser Blat results for the same sequences. D) Pairwise alignment of human and chimpanzee UGP2 distal first exon sequences with mismatches highlighted and Genome Browser Blat results for the same sequences.

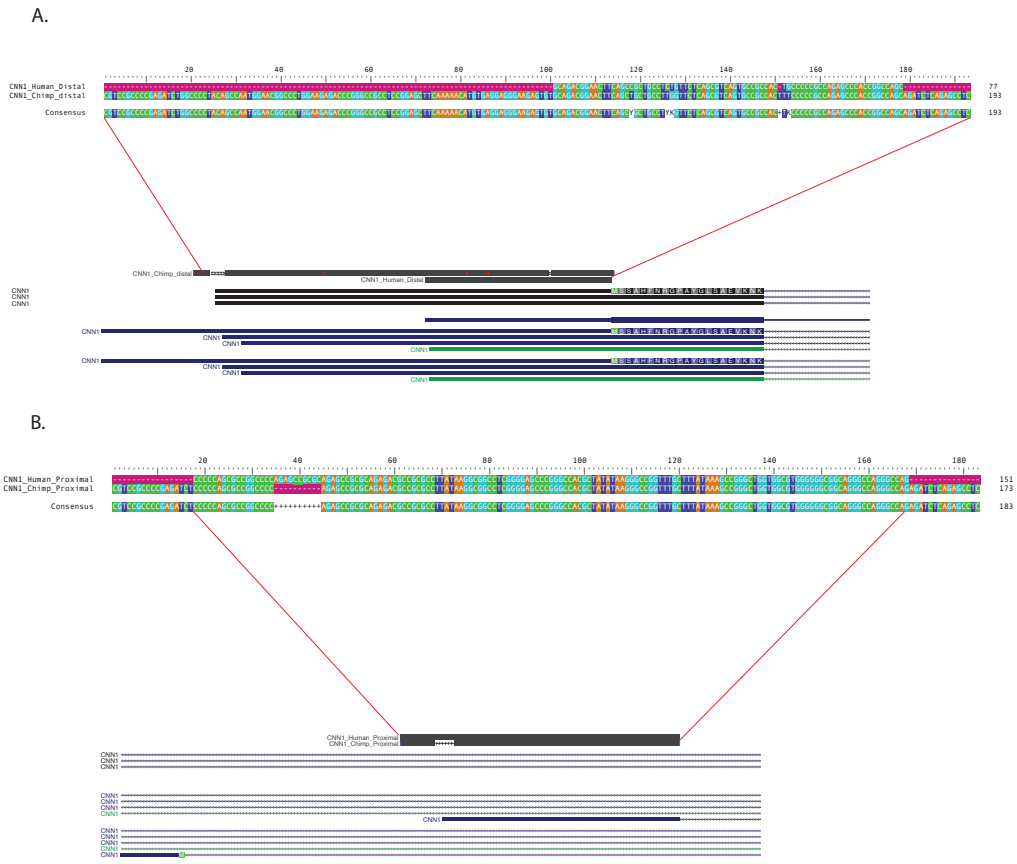

Figure 2.9 [Supplementary Figure 5]Pairwise Alignments of alternative first exons tested in luciferase reporters showing subtle differences that might regulate AS-TC, continued.
A) Pairwise alignment of human and chimpanzee CNN1 distal first exon sequences with mismatches highlighted and Genome Browser Blat results for the same sequences. B) Pairwise alignment of human and chimpanzee CNN1 proximal first exon sequences with mismatches highlighted and Genome Browser Blat results for the same sequences.

3. Chapter 3: The cis-regulatory landscape controlling isoform-specific translation in primates

3.1 Abstract

Steady-state mRNA levels and protein expression levels correlate poorly, indicating intricate post-transcriptional regulation of gene expression (de Sousa Abreu et al., 2009). This can be explained by the variation in mRNA stability, the regulation of translation initiation and elongation, and protein degradation. However, up to 30% of this discrepancy remain unaccounted for (Vogel et al., 2010). To close this gap, we conducted a comparative transcriptomics analysis to identify cis-regulatory elements implicated in the regulation of mRNA translation. We used previously collected Fractionation-Sequencing (Frac-seq) data from human, chimpanzee, and orangutan iPSCs, in which we had identified alternative splicing events that undergo translational regulation (AS-TC) in both a conserved or species-specific manner. We identified single nucleotide variants (SNVs) between two species at a time and evaluated the potential of these SNVs to change the trans-acting regulatory landscape of mRNA transcripts by utilizing publicly available protein-RNA interaction data. We identified multiple SNVs that have the potential to affect interactions between mRNA and RNA-binding proteins (RBPs). We demonstrated a weak correlation between codon optimality and polysome association in skipped AS-TC exons. We showed that the predicted overall secondary structure of event isoforms does not correlate with polysome association in alternative first AS-TC exons. Taken together, our prediction data suggest that single nucleotide variants are worth testing for their contribution to the species-specific translation of mRNA transcripts as they seem to perturb the binding of regulatory RBPs.

3.2 Introduction

Most metazoan/eukaryotic pre-mRNAs (the unprocessed transcripts of protein-coding genes) consist of multiple stretches of coding sequence (exons) interrupted by non-coding sequences (introns). Assembly of mature mRNAs involves post-transcriptional excision of introns and ligation of exon sequences by the spliceosome. The mature mRNAs can then be exported from the nucleus and translated by the ribosomes in the cytoplasm. The majority of genes with more than one exon can be spliced together in multiple, often cell type-specific, combinations (isoforms) in a process called alternative splicing (AS). This process is widespread: about 80% of genes undergo alternative splicing (Floor and Doudna, 2016), resulting in an average of 5 different isoforms (ENCODE Project Consortium, 2012; Tung et al., 2020). Alternative splicing, together with alternative transcription initiation and polyadenylation, can lead to substantial variation in the length, primary sequence, and secondary structure of both untranslated regions (UTRs) and coding sequences (CDS) of mRNA isoforms.

Alternative Splicing (AS) is an essential step in the regulation of eukaryotic gene expression. mRNA isoforms resulting from AS expand the cells' proteome with functionally distinct protein isoforms (Maniatis & Tasic 2002). Alternative splicing is further coupled to other post-transcriptional regulation processes: AS can designate transcripts to undergo nonsense-mediated decay by introducing premature termination codons (Brogna and Wen, 2009), thus regulating transcript decay and abundance. Alternative splicing, specifically of 3' UTRs, can also affect the subcellular localization of mRNAs (Taliaferro et al., 2016), affecting local regulation and protein synthesis. Since alternative splicing can result in differences in the primary sequence of mRNA isoforms, we expect the alternative use of UTR and coding exons to change the cis-regulatory landscape of mRNAs.

RNA-binding proteins (RBPs) are essential components in regulating the translation of mRNAs and their stability, subcellular localization, and many other processes. mRNAs are coated in a dynamic set of RBPs (ribonucleoproteins = mRNPs) throughout their whole lifecycle, which determines their cellular fate. While some of these RBPs change with the stage and location of the mRNA, others remain bound to the mRNA, coordinating and coupling multiple steps of post-transcriptional regulation (García-Mauriño et al., 2017). For example, exonic splicing enhancer and shuttling factor SRSF1 binds to nuclear mRNAs to aid in their nuclear export and stimulates the translation of these mRNAs in the cytoplasm (Sanford et al., 2004). Since many RBPs bind to mRNA in a sequence-dependent manner, alternative splicing and the resulting differences in the primary sequence of isoforms could give way to transcript-specific mRNP compositions and, therefore, transcript-specific cellular fates. Similarly, the stability of secondary structures depends on the primary sequence of an mRNA transcript and can affect the transcript-specific translation/ cellular fate. A prominent example for this mechanism is the alternative splicing of the Oskar pre-mRNA in *Drosophila melanogaster*: The exclusion of the first intron of this message allows the formation of secondary structure, which creates binding sites for trans-acting factors required for the correct cellular localization of the final mRNA (Ghosh et al., 2012; Hachet and Ephrussi, 2004). Finally, changes in the primary sequence of coding regions mRNAs could also affect codon optimality.

Our previous work (Sterne-Weiler et al., 2013), as well as others' (Floor and Doudna, 2016; Wong et al., 2016), demonstrates that alternative splicing of mRNA isoforms can lead to differential translational control of these isoforms contributing to the discrepancy between mRNA and protein expression levels. Our work further demonstrated that the coupling of alternative splicing with translational control (AS-TC) is a con-

served process across human, chimpanzee, and orangutan cell lines. We also identified AS-TC events with species-specific ribosome association with minimal differences in primary transcript sequence. These sequence differences were sufficient to change the activity of luciferase reporters *in vivo*, suggesting the intriguing hypothesis that these sequences affect the regulatory landscape of transcripts leading to differential translation.

In this chapter, we further analyzed the previously collected Frac-seq data (see Chapter Two). We conducted pairwise alignments of alternative splicing events coupled with translational control, identified SNVs, and tested the effect of these SNVs on potential protein-RNA interactions using publicly available RNAcompete data. This analysis led to the identification of promising SNVs that we propose to test for their ability to affect translation *in vivo* using luciferase assays. This analysis provides rich datasets for future hypothesis generation regarding which cis- and transacting elements in the regulatory landscape of mRNAs are sufficient to alter the translation of mRNA transcripts. Also, we examined the correlation of codon optimality and translation of isoforms and the changes in codon optimality as a consequence of alternative splicing and observed small to moderate correlations. The codon optimality analyses might point to an involvement of these factors in regulating isoform-specific translation but can not be singled out as determining factors. Finally, we tested the correlation between predicted mRNA secondary structures and translation of isoforms and observed no apparent correlations. This analysis does not suffice to exclude mRNA secondary structure from regulating isoform-specific translation but merely indicates the need for more refined analyses. Taken together, these data show that the regulation of mRNA translation is a complex process with many regulatory factors involved while also providing intriguing hypotheses for future research.

3.3 Results

3.3.1 Frac-seq allows the identification of orthologous mRNA isoforms with similar or species-specific polyribosome association)

We previously identified orthologous alternative splicing events with both consistent and species-specific ribosome association patterns (see Chapter Two). We visualized these subsets of events in correlation heatmaps (Figure 2.2C/D). Interestingly, in events with conserved sedimentation profiles, the same fractions from the different cell lines cluster together neatly, indicating the PSI values in these fractions are more similar to each other than to the other fractions within the same species. Consequently, in events with species-specific sedimentation profiles, the different fractions of each species cluster together, indicating more similarity within the species than the fractions. We further demonstrated more substantial sequence conservation in skipped (Figure 2.3A) and alternative first exons (Figure 2.3B) of events associated with isoform-specific translation compared to canonical AS exons, giving rise to the hypothesis that these highly conserved sequences are functionally relevant.

3.3.2 Identifying sequence differences between orthologous mRNA isoforms from different species

To test the hypothesis that AS-TC exons contain functional elements that influence polyribosome association, we identified single nucleotide variants between human and chimpanzee and human and orangutan ASTC exons and tested their potential to disrupt RBP binding (Figure 3.1A). Using pairwise alignments, we compared human and chimpanzee alternative first (AF) AS-TC exons and skipped (SE) AS-TC exons. A parallel analysis of human and orangutan AS-TC events was performed as well (Figure 3.8/Supplementary Figure 1). Normalized by sequence length, we observe the highest SNV frequency in ASTC exons with species-specific polysome association patterns followed by canonical AS exons, followed by ASTC exons with conserved sedimenta-

tion (Figure 3.1B). We observe the same pattern in skipped exon events (Figure 3.1D). Separating the SNVs into the different base for base substitutions, we see that the most common substitutions are A to G, C to T, G to A, and T to C in both AF (Figure 3.1C) and SE events (Figure 3.1E). These are all transitions, which are substitutions of purines for purines or pyrimidines for pyrimidines. These are the more commonly occurring substitutions (Collins and Jukes, 1994). Next we explored, where the SNVs are located in relation to the splice sites of the tested exons and found an enrichment of variants around 50bp away from the splice sites in the exon sequence in both AF and SE events (Figure 3.1.F-I). This enrichment agrees with previously observed positional biases of SNPs around splice sites (Majewski and Ott, 2002). The sharp drop off in SNV density around 100-150bp away from splice sites is likely due to the length distribution of the tested exons.

To test the hypothesis of these SNVs disrupting the binding of RNA binding proteins, we used a sliding window approach (Eickhardt et al., 2016; Soemedi et al., 2017) and publicly available RNAcompete data (Ray et al., 2009, 2013, 2017) for human to record the possibility of binding of 80 RNA-binding proteins (102 different binding motifs) in a 6 to 7nt window around the SNVs in both species of the comparison (Figure 3.1A). We recorded the predicted binding affinity in both species based on PWM matching and the difference in binding affinity.

3.3.3 Single nucleotide variants between orthologous mRNA isoforms affect predicted RBP binding affinity

To identify single nucleotide variants and RBPs that might affect translational control of alternative splicing events, we compared the predicted changes in RBP binding between AS and ASTC events, as well as in events considered ASTC in both species. Figure 3.2A shows SNVs between human and chimpanzee AF events classified AS in one

species and ASTC in the other that lead to significant predicted binding in (one or) both species. SNVs leading to similar binding in both species are located on and around the line of equality (x=y), while SNVs that lead to differential predicted binding are located closer to the x- or y-axis. We observe SNVs that lead to weaker binding in AS events and some that lead to weaker binding in ASTC, while some do not cause a change in predicted binding between human and chimpanzee iPSCs. Testing SNVs in events that are considered ASTC in both species results in very similar patterns. As a background, we tested SNVs in events that are considered AS in both species. The patterns are very similar here as well. We performed parallel analyses for between human and orangutan alternative first exons (Figure 3.2D/E/F), human and chimpanzee skipped exons (Figure 3.2G/H/I), and human and orangutan skipped exons (Figure 3.2J/K/L). Testing SNVs in different groups of events, especially with AS events as a background set, allows us to test the frequencies of which RBP binding sites are affected by SNVs and compare. We tested the significance of the frequency binding site changes between different groups with the chi square-test for two proportions. We identified multiple RBPs with a higher frequency of binding site changes in ASTC/ASTC or ASTC/AS sets compared to the background (Figures 3.9-3.12/ Supplementary Figures 2-5). Significantly over- or underrepresented RBPs in the different comparisons are visualized in Figures 3.3 and 3.4. These RBPs could be ASTC regulatory candidates and should be pursued in further experiments. It is interesting that similar proteins appear to be overrepresented in multiple event groups, multiple of which are known translational regulators such as multiple members of the SR family.

3.3.4 Global effects of cis-regulatory differences and transacting factors

We then investigated how likely an SNV is to lead to a difference in binding affinity ("change") between two species in the form of odds ratios (OR) (Figure 3.5A-C).

We compared the SNVs occurring in events with species-specific polysome association patterns to AS-TC events with consistent polysome association patterns across the compared species (conserved). We found no significant differences when pooling the SNV information for both alternative first exons (Figure 3.5A). However, we found that SNVs in the distal first exon alone are more likely to cause a change in the binding affinity in the species-specific AS-TC group compared to the conserved group (Figure 3.5B). Further, the SNVs identified in the proximal first exons are more likely to not cause a change ("neutral") in the binding affinity in the species-specific AS-TC group compared to the conserved group (Figure 3.5C). This, together with the differential sequence conservation we observed between the two AF exons (Figure 2.3B), could indicate a more prominent role for the distal first exon in the translational regulation of AS-TC exons.

Next, we tested whether the SNVs that were identified in events with species-specific sedimentation are more likely to increase ("gain") or decrease ("loss") the binding affinity compared to AS-TC events with conserved sedimentation (Figure 3.5D-F). We performed this analysis from a human-centric point of view. Increase in binding affinity, therefore, means increase in binding affinity in human compared to chimpanzee. According to the OR, the SNVs in distal first exons are more likely to cause an increase in binding affinity (Figure 3.5E). We found no significant differences in proximal first exons (Figure 3.5F). The pooled analysis agrees with the distal first exon (Figure 3.5D). We observed similar behavior in the human to orangutan comparison (Figure 3.13/ Supplementary Figure 6). Testing SNVs in skipped exons in both human/chimpanzee and human/orangutan comparisons did not show any significant results (Figure 3.14/ Supplementary Figure 7).

To determine if specific RBP binding sites are more susceptible to change through single nucleotide variants, we investigated the log-fold enrichment of binding sites for

each of the 102 binding motifs (PWMs) AS-TC exons relative to a background of SNVs in canonical AS events. The heatmaps visualize the log-fold enrichment of binding sites for neutral or change ASTC SNPs over non-ASTC SNPs (background) per tested PWM (Figure 3.5G-I). The column on the left represents the significance of the over- or underrepresentation according to a binomial distribution (see methods). Very few RBPs, e.g. SRSF10 and YBX2, stand out to be significantly overrepresented in either of the exons of the combined analysis (Figure 3.5H,I). This lack of significant overrepresentation indicates that it might be challenging to narrow down AS-TC regulation as an effect of few and specific RBPs. This is not too surprising given the combinatorial nature of RBP-dependent regulation.

3.3.5 Does codon usage correlate with isoform-specific polysome association?

To test if alternative splicing, specifically within the coding region of a transcript (CDS), can alter the codon optimality sufficiently to affect the mRNA translation, we calculated the codon content of AS and AS-TC events, specifically skipped exon events, since those are the most common alternative splicing events to occur in the coding region of an mRNA. The codon content was calculated as overall GC content, GC content in the third position of codon triplets, and GC content in the third position of four-fold degenerate codons (GC4) as previously described (Mordstein et al., 2020). We correlated those measures with the measure of polysome association (PA, formula see Methods) aiming to represent the polyribosome association pattern of an mRNA isoform, where a higher PA indicates a stronger association with the heavy polyribosome over the light polyribosome or the monosome (Figure 3.6). Alternatively, we tried a measure of ribosome association (RA) and a variation of polysome association (PA2), both of which did not result in any significantly stronger correlations with codon content (Figures 3.15 and 3.16 / Supplementary Figures 4 and 5, respectively). We observed a

medium correlation between GC content and PA in both AS and ASTC (Figure 3.6A), testing the difference between the correlation with the Fisher's r-to-z test did not result in a significant difference in correlation. We further observed rather small correlations of GC3 and GC4 with PA. While the R value tends to be higher in AS-TC events, there were no significant differences between the AS and AS-TC exons (Figure 3.6B,C). To try and investigate the changes in codon content between pairs of isoforms, we also tested the correlation in GC, GC3, and GC4 between included and excluded isoforms for both AS and AS-TC events (Figure 3.17 / Supplementary Figure 10). We observed high overall correlations in GC, GC3 and GC4 content, which is to be expected. While not significantly different, the R value tended to be higher in ASTC exons (Figure 3.17B,C,D).

3.3.6 Does predicted mRNA secondary structure correlate with polysome association?

Finally, to test the hypothesis that complex mRNA secondary structures could affect mRNA translation, we tested the correlation of predicted secondary structure against the PA measure in AS and ASTC (Figure 3.7). For that purpose, we used the Vienna RNA fold tool to predict the free Gibbs energy for each isoform of both skipped exons and alternative first exons of events associated with ASTC or AS. For translation, we used all three measures described above: RA, PA, and PA2 for both types of alternative splicing events in human. We did not observe any correlation between free Gibbs energy and any translation measures in any of our comparisons (Figure 3.7). These observations hardly mean that mRNA secondary structures are not involved in regulating isoform-specific mRNA translation. Most evidence for mRNA secondary structures regulating translation is based on secondary structures in the 5'UTR of the transcripts and around the translation initiation site. Highly structured sequences can make sites less accessible for translation initiation factors and therefore regulating the rate of ini-

tiation (Mustoe et al., 2018). Therefore, it would be beneficial for this type of analysis to limit the structure prediction to either the 5'UTRs of the transcripts or a specific window around the translation start site (start codon), e.g., Mordstein et al. (Mordstein et al., 2020) conducted structure prediction on a window of 42nt around the translation initiation site.

3.4 Discussion

Here we present a comparative transcriptomics analysis of previously published Frac-Seq data to identify cis-regulatory elements involved in the translational regulation of mRNAs, modified by alternative splicing. We previously identified alternative splicing events that are predicted to be implicated in translational control based on their dynamic polyribosome association patterns. We categorized these events into events with conserved polyribosome association patterns and events with species-specific patterns. Based on these findings, we were able to identify single nucleotide variants between pairs of species and evaluate their potential to disrupt RBP binding sites.

Our primary hypothesis was that the change in protein-RNA interactions due to sequence differences driving isoform-specific mRNA translation. Using RNAcompete is an exciting first pass at predicting protein-RNA interactions that yielded mostly isolated SNVs that will be intriguing to test.

To test if the computationally identified single nucleotide variants between two species can affect mRNA translation as predicted based on our analysis in conjunction with the publicly available RNAcompete data, we propose a variation of pairs of luciferase reporters as previously described (Chapter Two). Based on our previous reporter results, where we were able to show differential luciferase activity from orthologous

exons, we would be creating pairs of luciferase reporters with identical sequences, e.g., based on the human exon, with a single nucleotide difference, i.e., the chimpanzee SNV. If the single nucleotide can disrupt the binding of a regulatory RNA binding protein, we would expect a significant change in luciferase activity between the two reporters. The most attractive SNV to test in this case are the ones that occur pretty isolated, e.g., not many other SNVs within the same event. This makes testing the hypothesis that this particular SNV affects mRNA translation very straightforward. If the SNV does not affect translation, it is easier to conclude that other factors, not RBP binding, are involved in translation regulation. Based on previous experiments and our predictions, we have identified the skipped exons of GGCX and SUMF2 and the alternative first exons of CNN1 and UGP2 *in vivo* testing with luciferase reporters. All of these sequences demonstrated significant differences in luciferase activity between species (Figure 2.4). Both GGCX and SUMF2 appear to only have one SNV between the human and chimpanzee sequences, making these the obvious targets for testing (Figures 2.7). According to our prediction using the RNAcompete data, the SNV in GGCX would disrupt MBLN1, MSI1, and SAMD4A binding sites, the SNV in SUMF2 would disrupt CNOT4, MSI1, and SRSF10 binding sites (Figure 2.7). The sequence differences in the alternative first AS-TC events include both insertions as well as SNV that disrupt many RBP binding sites in MAD2L2, UGP2, and CNN1 (Figures 2.8 and 2.9), that will be more complex to untangle and will have to be properly prioritized for testing.

The computational prediction approach as well as the luciferase reporter assays could be followed up by analysis of both RNA bind-and-seq data, which is also predictive but will help cover a broader range of RBPs, and publicly available iCLIP datasets, which are more *in vivo* data. However, it is crucial to keep some of the limitations of these approaches in mind. The primary Frac-seq experiment was performed in human

and primate iPSCs, which are cell lines and species for which we have little if no binding data available that applies precisely to these experimental conditions. The luciferase reporter assays we have performed in the past (Chapter Two) and are proposing in this chapter are performed in HEK cells and are therefore only an approximation. The RBPome, and the RBP binding affinities, could look quite different in the primate species. An experimental approach that could be very interesting would be to perform RNA affinity chromatography on specific mRNA transcripts that have already been tested/validated for isoform-specific translation (e.g., CNN1, UGP2, GGCX, and SUMF2; see Chapter Two). This assay would allow us to identify specific RBPs that bind to these RNA sequences and identify differences in the mRNPs between species or alternative splicing events. Further, redoing this Frac-seq experiment using long-read sequencing could be very informative. It would allow us to identify full-length mRNA transcripts and, consequently, to identify the whole predicted RBPome and allow us to learn more about the combinatorial effects of RBPs on translational control of alternative splicing events.

Alternative to the RBP hypothesis, we suggested that changes in codon optimality of alternative splicing events within the coding sequence could also affect isoform-specific translation. To test this, we tested correlations of codon content and translation measures and found relatively small correlations. However, it is essential to consider what kind of correlations we would expect from such an experiment. It is already clear that many factors affect mRNA translation. Therefore examining just one potential factor will not lead to perfect correlation coefficients. We observe slightly higher correlations between codon content and translation measures for AS-TC events than AS, indicating the potential relevance of codon content for translation. These analyses could also be followed up with more in-depth computational analyses as well as experimental

approaches. It would be interesting and potentially more straightforward than the correlation analysis to identify which codons are commonly changed through SNVs between the species and if the SNVs lead to more or less optimal codons and synonymous or nonsynonymous codons. Ultimately, this hypothesis could also be tested by comparing luciferase reporters containing, e.g., the original human exon, a codon-optimized, and a less codon-optimized version of the same sequence. It is further important to note that codon and GC content does not actually test or represent codon optimality. Even though GC3 and GC4 content have been used in correlation analyses of mRNA fate (Mordstein et al., 2020), metrics like RSCU (Relative Synonymous Codon Usage) or tAI (tRNA adaptation index) are likely more appropriate for the analysis of mRNA translation or translatability.

Our second alternative hypothesis regarding the regulation of isoform-specific translation involves differential secondary structures that could change the trans-regulatory landscape by changing the accessibility of mRNA transcripts. As mentioned before, this analysis is very much preliminary. We did not observe any correlation between free Gibbs energy of ASTC isoforms and their translation, which is not enough to refute our hypothesis. For further analysis, it would be helpful to narrow down the area of the mRNA transcript that is tested for mRNA secondary structures, e.g., the 5'UTR or a window around the translation initiation site. After that, experimental exploration of differential secondary structures using selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) (Wilkinson et al., 2006) on specific alternative isoforms could be a promising way of validating the structure predictions.

In this chapter, we explored different hypotheses for the isoform-specific regulation of mRNA translation. The mechanisms we hypothesized are that sequence differ-

ences caused by alternative splicing affect the cis and trans-regulatory landscape (e.g., RBP binding sites, secondary structure, codon use), ultimately leading to differential translation. Comparing different species where orthologous and mostly identical sequences lead to differential translation patterns allows us to narrow down which sequences might be relevant to translation regulation. Our analysis of SNVs and their potential to disrupt or increase RBP binding provides the data for many new and testable hypotheses, which could ultimately lead to the discovery of new or multifunctional regulatory sequences. Our codon usage and secondary structure prediction analyses, albeit preliminary, might be the starting point of more in-depth and, more importantly, experimental approaches. Taken together, our data do not refute any of our three hypotheses but provide new, testable hypotheses for protein/RNA interactions in the context of translation and indicate that regulation of translation is a complex and multilevel process that is difficult to narrow down to one factor or another.

3.5 Materials & Methods

3.5.1 Alignment and identification of SNVs and indels

The sequences of alternative regions of AS-TC events from human and chimpanzee were globally (Needleman-Wunsch alignment) aligned using the R Biostrings (Pagès et al., 2020) function pairwiseAlignment with default settings. Mismatches, insertions, and deletions were identified in the pairwise alignments using R Biostrings. The distribution of single nucleotide variants in relation to alternative exon boundaries (splice sites) was visualized using the density function from ggplot2 (Wickham and Others, 2009) which computes and plots the kernel density estimate.

3.5.2 RBP binding analysis

RNAcompete (Ray et al., 2009, 2013) data sets available on ENCODE (ENCODE Project Consortium, 2012) were used to predict RBP binding sites affected by single

nucleotide sequence differences between human and chimpanzee ASTC events. The matchPWM (Wasserman and Sandelin, 2004) function from the R Biostrings package (Pagès et al., 2020) was used to score PWMs based on the RNAcompete data in a sliding window across the identified sequence differences. Matches achieving at least 80% of the maximum score were recorded for both human and chimpanzee datasets. The matches were compared between the species in form of a deltaPWMscore (e.g. PWMscore(human) - PWMscore(chimp)). Z-scores were calculated for all deltaPWMscores. Mismatch positions that resulted in a z-score > 2 or z-score < -2 were considered significant. However, for unless explicitly stated, any differences in binding affinities were included in the analyses, not just significant ones according to the z-score cutoff.

3.5.3 Change, gain and loss of binding

RBP binding to a kmer is predicted based on the score of the PWMmatch function (see above). Change of binding is defined as any differences in the PWMscore for an RBP between two species. Neutral means there is no difference in PWMscores. Gain and loss of binding are defined in a human centric way, where gain is PWMscore(human) > PWM(chimpanzee/orangutan), and loss where PWMscore(human) < PWMscore(chimpanzee/orangutan).

3.5.4 Odds ratios and binomial estimation

We counted the numbers of SNPs causing change or now change in binding (change/neutral) or causing gain or loss in binding (gain/loss). An odds ratio (OR), was calculated given

$$\mathrm{OR} = \frac{P(event|species\ sp.)/[1 - P(event|species\ sp.)]}{P(event|conserved)/[1\ - P(event|conserved)]}$$

where the event can be loss or gain of an RBP binding site (gain or loss of predicted binding affinity) and P(loss|data set) = 1- P(gain|data set). Odds ratios are plotted as bars, with 95% confidence intervals (two-tailed). The error bars are calculated using standard methods (Pagano and Gauvreau 2000).

To assess the enrichment of SNVs causing change in RBP sites over no change or gain/loss in binding over no change, we visualized the log2(change/neutral) or log2(gain/neutral) and log2(loss/neutral) in heatmaps. To determine the significance of enrichment of SNVs within certain RBP binding sites, we used the binomial distribution as described by Sterne-Weiler et al., 2011. The p-values were corrected for multiple testing using Benjamini-Hochberg false discovery rate (FDR) of 5% (Benjamini and Hochberg 1995).

3.5.5 Measures of (poly)ribosome association based on Frac-seq data

The calculation of measures of (poly)ribsome association is based on the calculations presented by Mordstein et al., 2020. However, adjustments were made based on the unavailability of the free RNP fraction from our Frac-seq data.

Ribosome association (RA) for each event isoform was calculated as the sum of junction reads in the monosomes, light polysomes, medium polysomes, heavy polysomal fractions, divided by the junction reads found in the cytoplasmic fraction:

$$\frac{mono + poll + polm + polh}{cyto} = \text{ribosome association (RA)}$$

Polysome association (PA) for each event isoform was calculated as the sum of junction reads in the light, medium, and heavy polysomal fractions, divided by the junction reads in the cytoplasmic fraction.

$$\frac{poll + polm + polh}{cyto} = \text{polysome association (PA)}$$

A variation of the polysome association (PA2) for each event isoform was calculated as the sum of junction reads in the light, medium, and heavy polysomal fractions, divided by the junction reads in the monosomal fraction:

$$\frac{poll + polm + polh}{mono} = \text{polysome association (PA2)}$$

3.5.6 Codon content

GC content and codon content (GC1, GC2, GC3) was calculated using functions from the R package seqinr (Charif and Lobry, 2007). GC(1/2/3) indicates the GC content at the first, second, and third positions of the codon. GC4, the GC content at fourfold degenerate bases, was calculated with custom R scripts.

3.5.7 Structure prediction

The mRNA secondary structure prediction was performed using UNAfold (Zuker 2003; Markham and Zuker 2008). The minimum free energy (dG) of predicted mRNA secondary structure was calculated using the hybrid-ss-min program version 3.8 (default settings: NA = RNA, t = 37, [Na+] = 1, [Mg++] = 0, maxloop = 30, prefilter = 2/2).

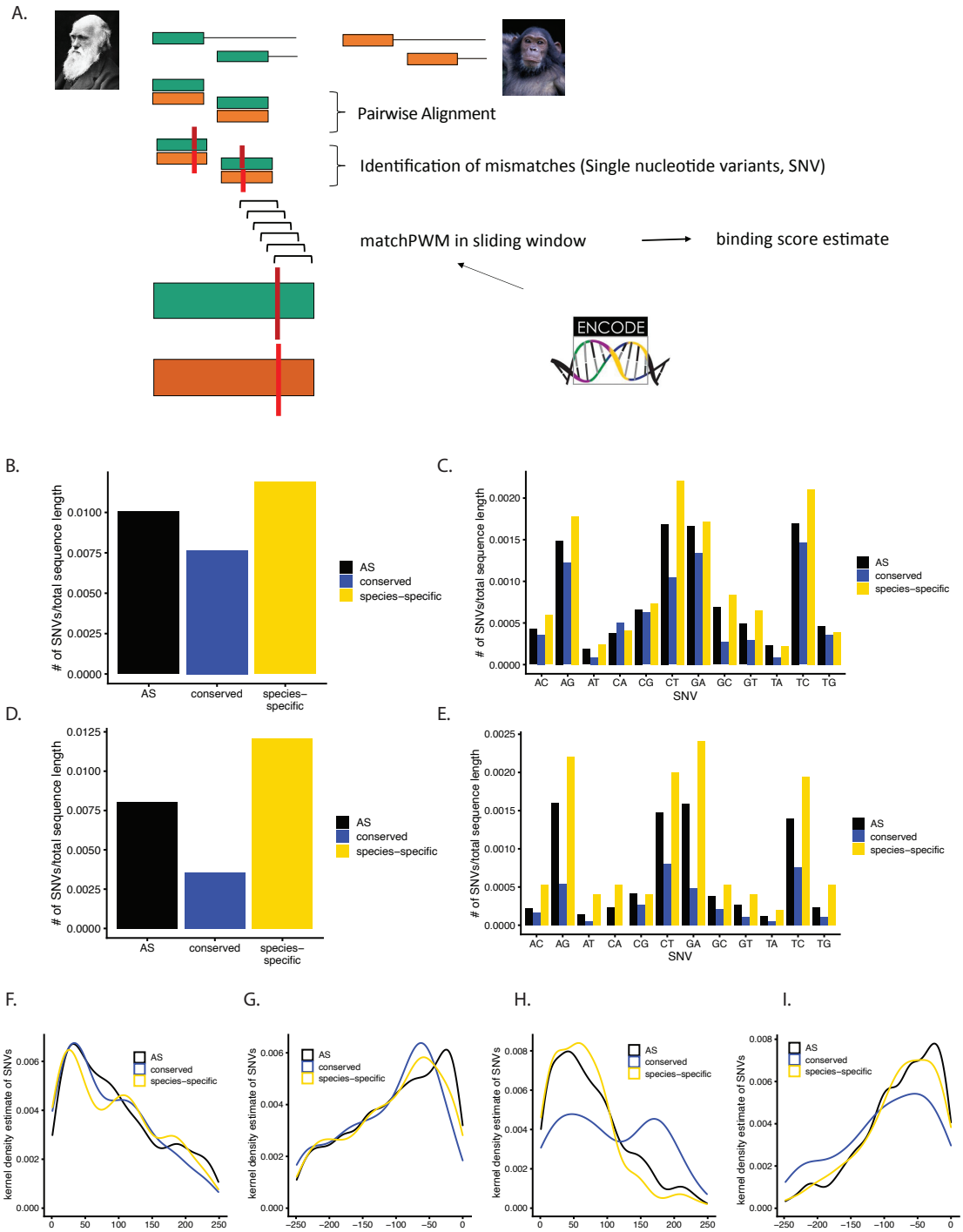

Figure 3.1 Distribution of SNVs and sliding window approach for RBP binding prediction.
A) Approach to identify SNVs between e.g., human and chimpanzee AF exons and then predict their potential to change RBP binding. B) Frequency of SNVs in AF ASTC exons with conserved or species-specific sedimentation compared to canonical AF exons C) Frequency of different SNVs in AF exons. D) Frequency of SNVs in ASTC skipped exons with conserved or species-specific sedimentation compared to canonical skipped exons. E) Frequency of different SNVs in skipped exons. F-I) Spatial distribution of SNVs around exon boundaries in AF and SE events.

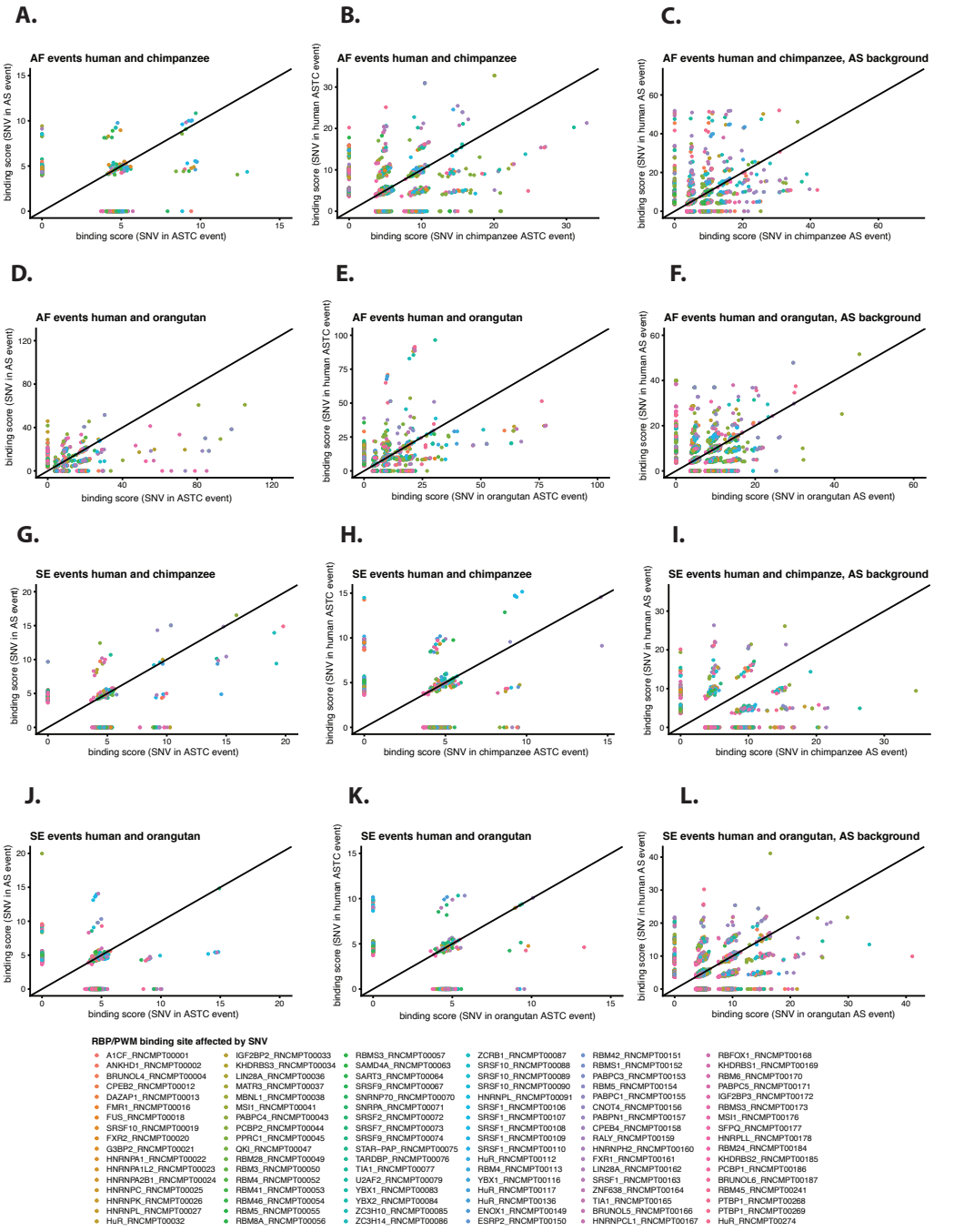

Figure 3.2 SNVs found in pairwise alignments of alternative first AS-TC events have the ability to disrupt or generate new RBP binding sites

SNVs identified in pairwise alignments in different groups of events that are predicted to be within kmers bound by one of the 102 PWMs tested and their predicted binding scores. A,D,G,J) Events considered AS in one, but ASTC in the other species. B,E,H,K) Events considered ASTC in both species. C,F,I,L) Events considered AS in both species. This group serves as control group.

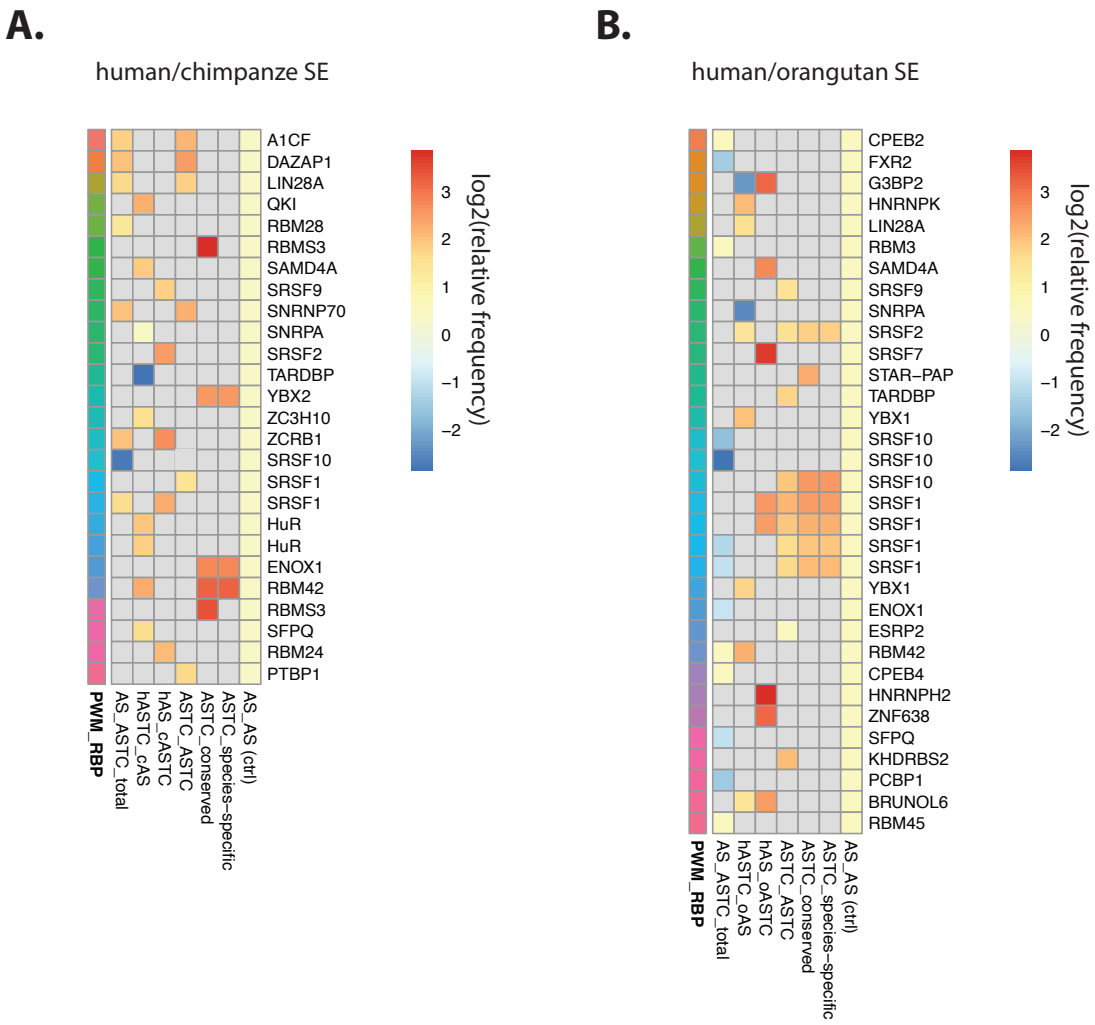

Figure 3.3 Identification of ASTC regulatory candidates in skipped exon events.
Heatmaps summarizing which RBP binding sites are over- or underrepresented in the number of binding sites affected by single nucleotide differences in skipped exons compared to the AS control set in A) human and chimpanzee sequences and B) human and orangutan sequences. Each row represents an RBP, each column represents a test set of AS or ASTC exons or the control set. The color of each cell shows the log2 of the relative frequency of RBP between test and control set. Red indicates an overrepresentation compared to the control group and blue indicates an underrepresentation compared to the control group.

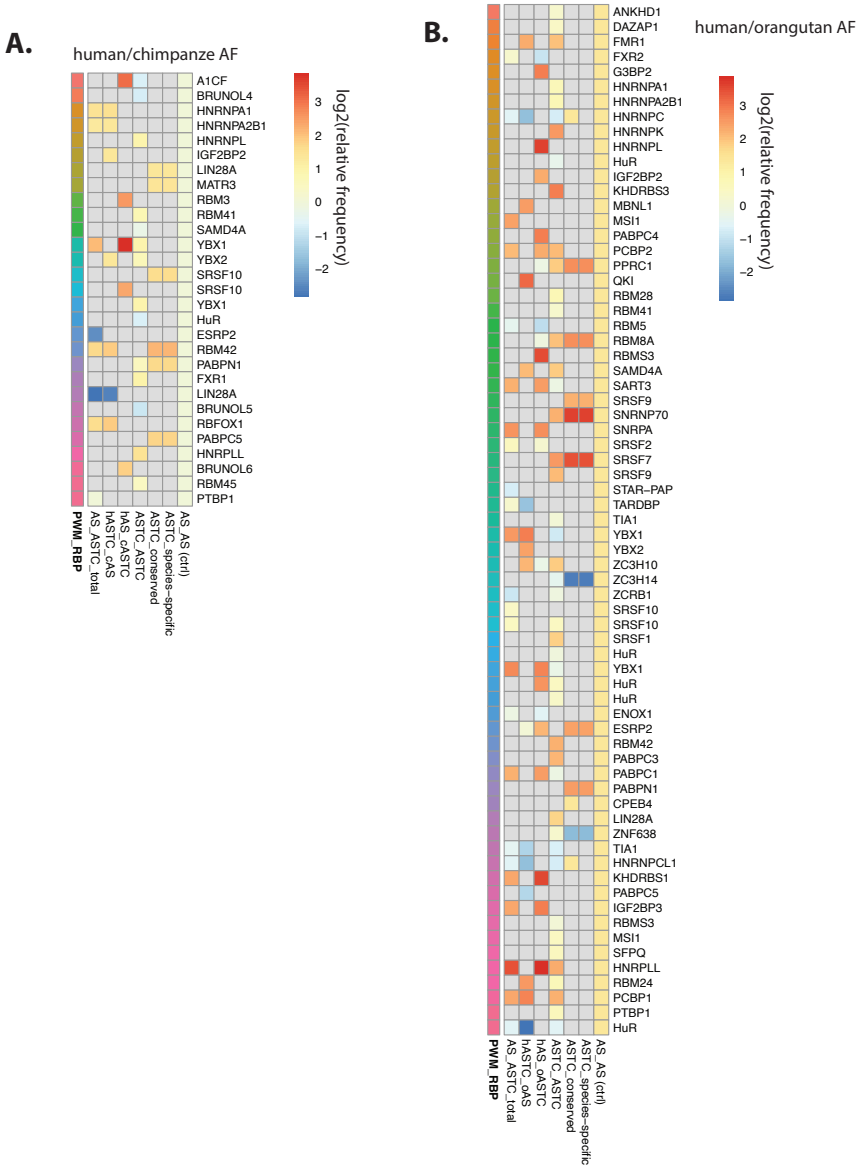

Figure 3.4 Identification of ASTC regulatory candidates in alternative first exon events.
Heatmaps summarizing which RBP binding sites are over- or underrepresented in the number of binding sites affected by single nucleotide differences in skipped exons compared to the AS control set in A) human and chimpanzee sequences and B) human and orangutan sequences. Each row represents an RBP, each column represents a test set of AS or ASTC exons or the control set. The color of each cell shows the log2 of the relative frequency of RBP between test and control set. Red indicates an overrepresentation compared to the control group and blue indicates an underrepresentation compared to the control group.

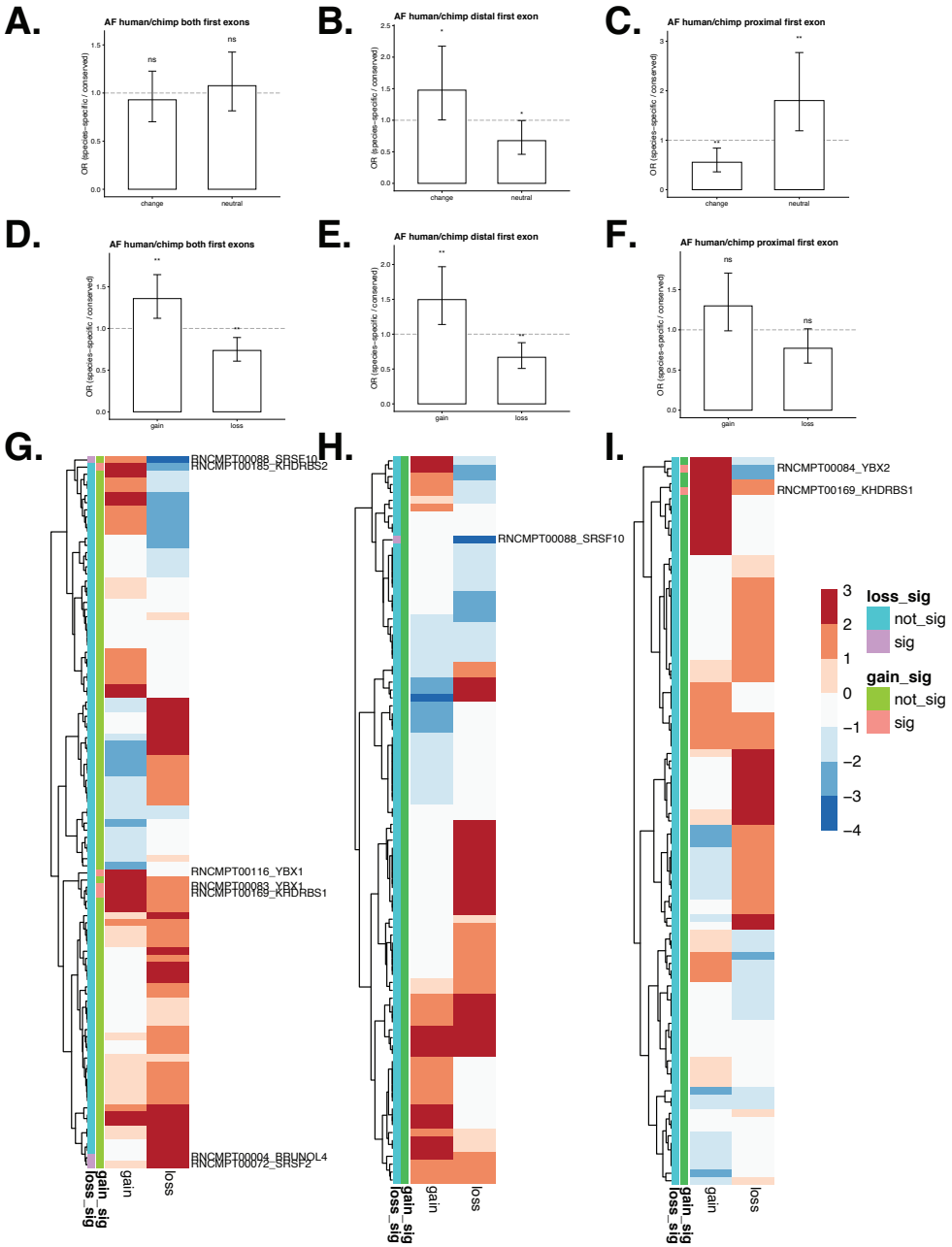

Figure 3.5 Global effects of SNVs found alternative first AS-TC events.
A-C) Odds ratios (OR) of SNVs identified between human and chimpanzee AS-TC events that lead to a change or no change ("neutral") in RBP binding as predicted based on RNAcompete data. For both first exons combined, for the distal first exon, and for the proximal first exon. D-F) Odds ratios of SNVs identified between human and chimpanzee alternative first AS-TC events that have been identified to change RBP binding affinity to lead to a gain or loss of a binding site. For both first exons combined, for distal first exon only, for proximal first exon. Bar height indicates OR. Error bars represent the two-tailed 95% confidence interval for the bar height. G-I) Heatmaps showing the enrichment of PWMs in events with an increase in RBP affinity (gain) or with a decrease in RBP affinity (loss) around the identified SNVs. The red/blue color scale indicates the log-fold enrichment over AS event background. Green and teal columns to the side indicate the significance of this enrichment based on a binomial distribution

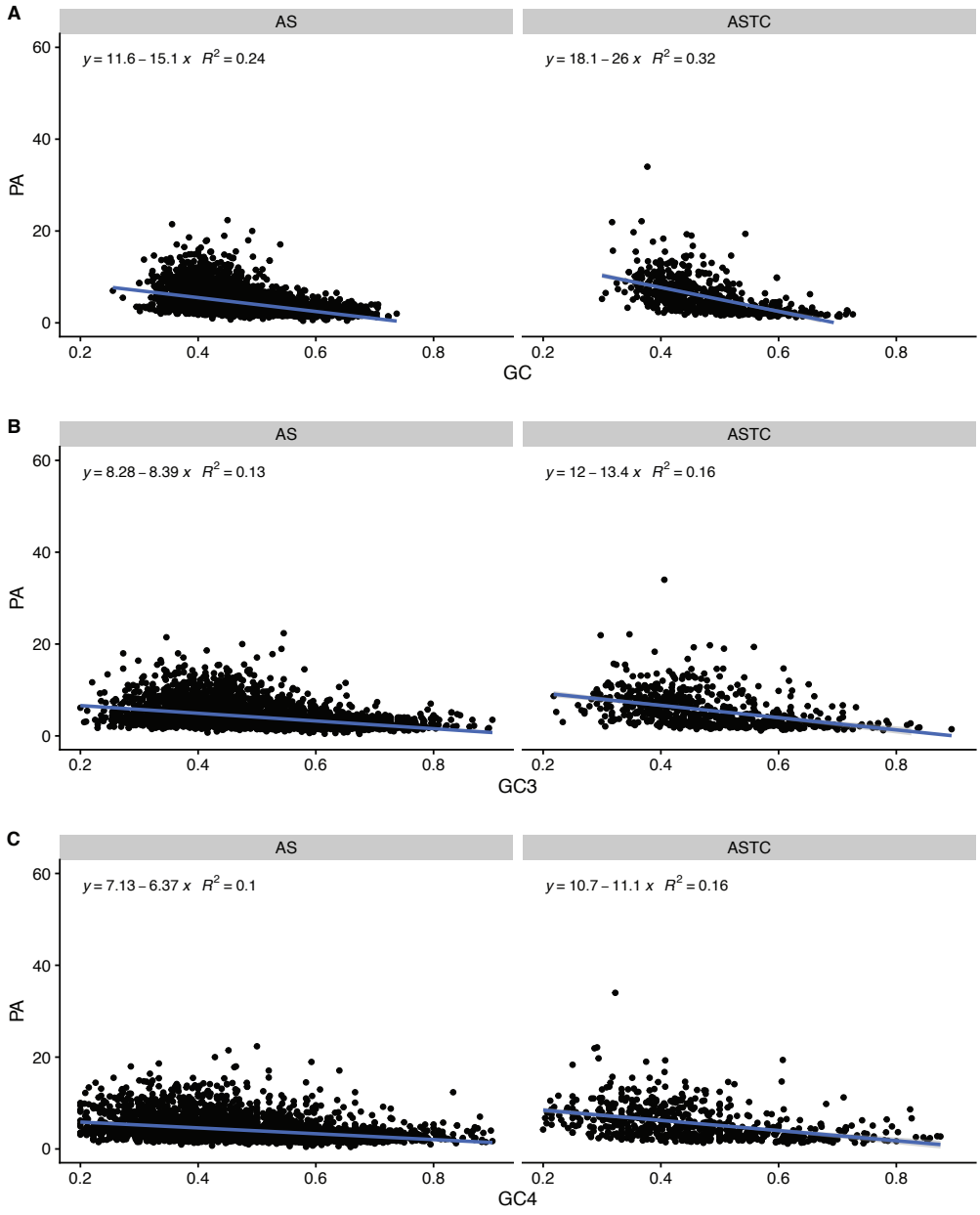

Figure 3.6 Codon content weakly correlates with polyribosome association for AS and AS-TC skipped exons.

A) Correlation between GC content and polysome association (PA) in AS (left) and AS-TC (right) skipped exon events. B) Correlation of GC3, GC content at the third position of all codons, and polysome association in AS and AS-TC skipped exons. C) Correlation of GC4, GC content at the third position of four-fold degenerate codons, and polysome association in AS and AS-TC exons.

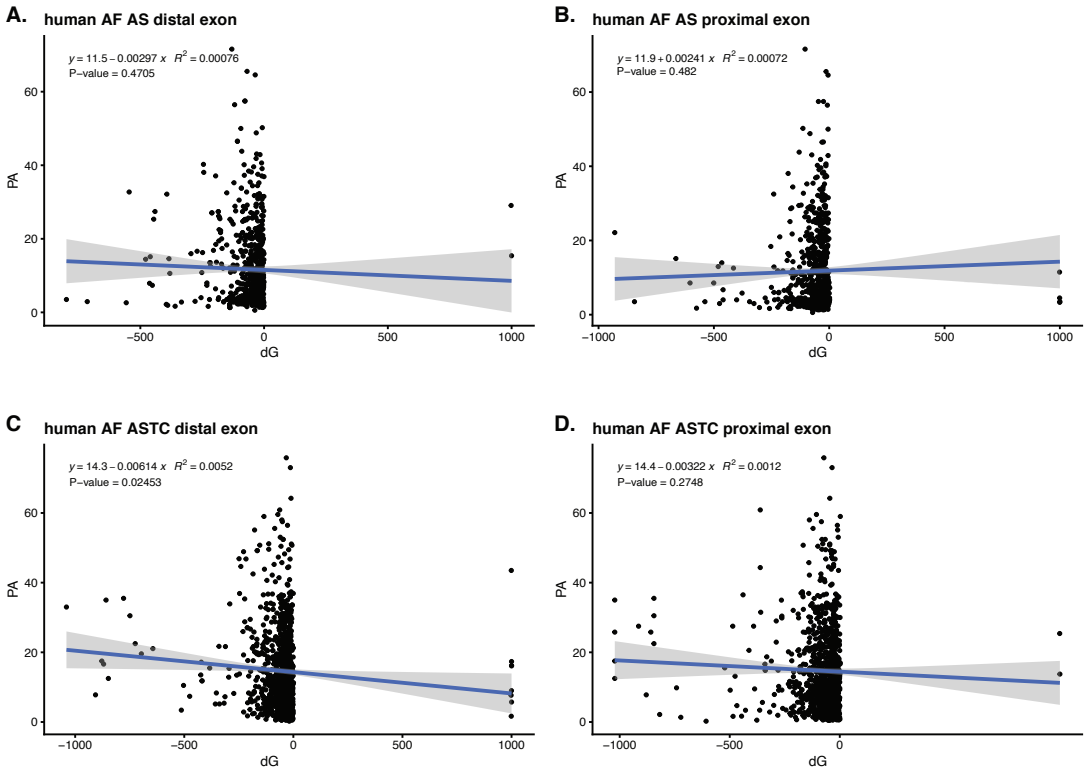

Figure 3.7 Predicted mRNA secondary structure (gibbs free energy) does not correlate with isoform translation measures.

A) Correlation of Gibbs free energy (dG) and polysome association (PA) in AS (left) and AS-TC (right) skipped exon events. B) Correlation of Gibbs free energy (dG) and polysome association in AS and AS-TC skipped exons. C) Correlation of Gibbs free energy (dG) and polysome association in AS and AS-TC exons.

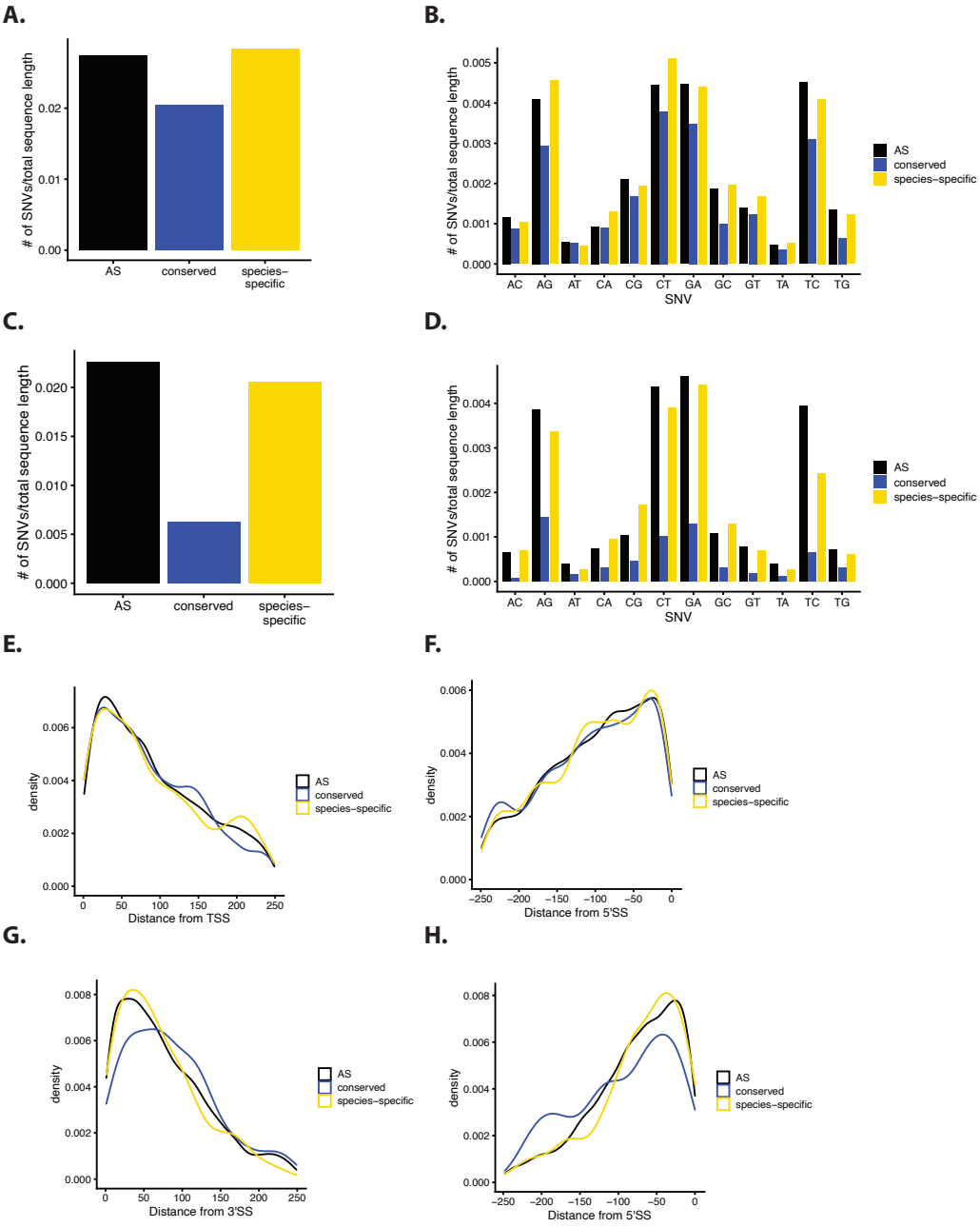

Figure 3.8 [Supplementary Figure 1] Distribution of single nucleotide variants between human and orangutan AS-TC exons.

A) Frequency of SNVs in AF ASTC exons with conserved or species-specific sedimentation compared to canonical AF AS exons in between human and orangutan sequences. B) Frequency of the different SNVs in AF exons. C) Frequency of SNVs in ASTC skipped exons with conserved or species-specific sedimentation compared to canonical skipped exons between human and orangutan sequences. D) Frequency of different SNVs in skipped exons. E,F) Spatial distribution of SNVs around exon boundaries in AF events. G,H) Spatial distribution of SNVs around exon boundaries in SE events.

Figure 3.9 [Supplementary Figure 2] Over- and underrepresented RBPs within the total binding sites affected by SNVs between human and chimpanzee in skipped exon events.
The bars indicate the frequency of each RBP's binding sites being affected by SNVs in the different test groups and the control group. Significant differences in frequencies were determined with the chi square test of two proportions. * indicates p≤ 0.05; ** indicates p≤ 0.01; *** indicates p≤ 0.001.

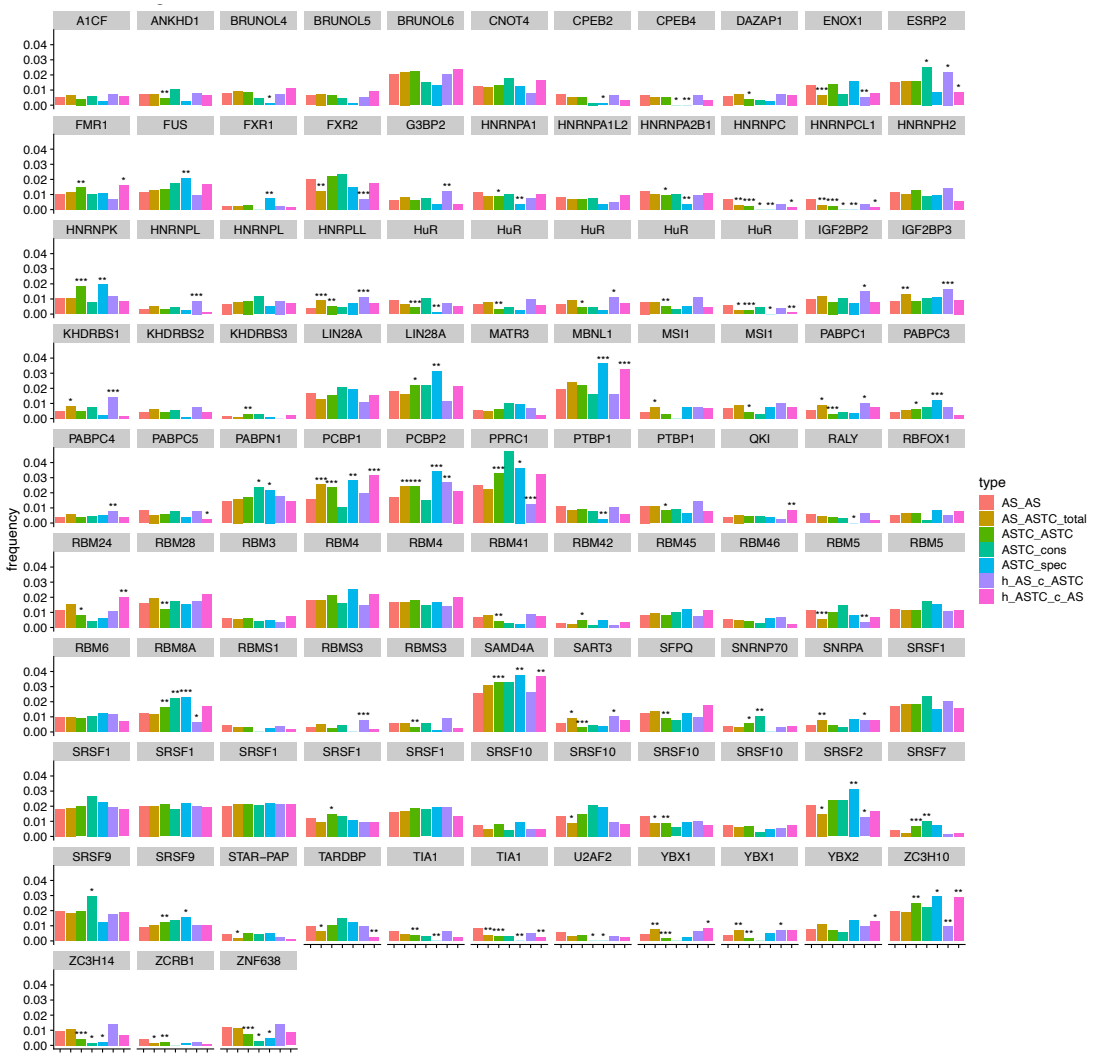

Figure 3.10 [Supplementary Figure 3] Over- and underrepresented RBPs within the total binding sites affected by SNVs between human and orangutan in skipped exon events.
The bars indicate the frequency of each RBP's binding sites being affected by SNVs in the different test groups and the control group. Significant differences in frequencies were determined with the chi square test of two proportions. * indicates p≤ 0.05; ** indicates p≤ 0.01; *** indicates p≤ 0.001.

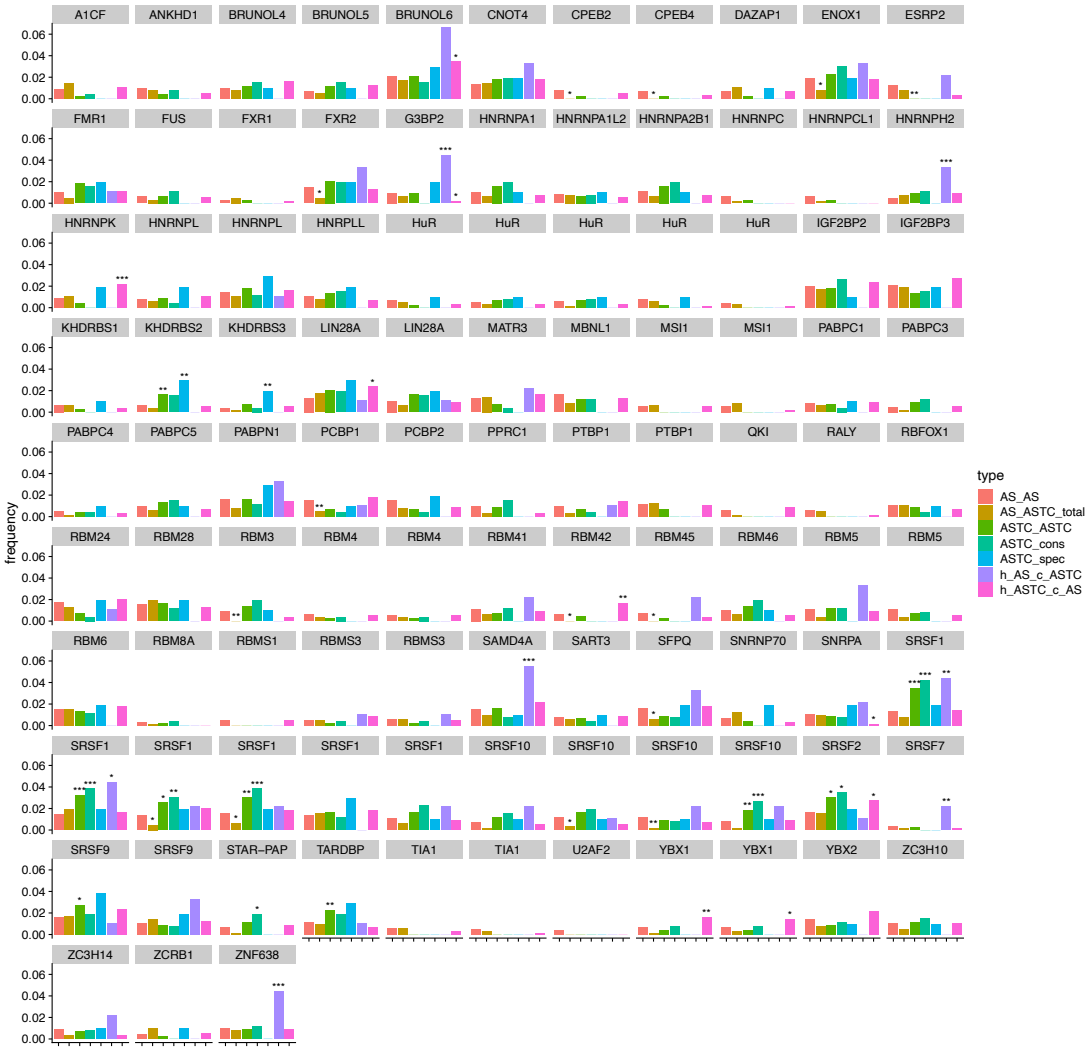

Figure 3.11 [Supplementary Figure 4] Over- and underrepresented RBPs within the total binding sites affected by SNVs between human and chimpanzee in alternative first exon events.
The bars indicate the frequency of each RBP's binding sites being affected by SNVs in the different test groups and the control group. Significant differences in frequencies were determined with the chi square test of two proportions. * indicates p≤ 0.05; ** indicates p≤ 0.01; *** indicates p≤ 0.001

Figure 3.12 [Supplementary Figure 5] Over- and underrepresented RBPs within the total binding sites affected by SNVs between human and orangutan in alternative first exon events.
The bars indicate the frequency of each RBP's binding sites being affected by SNVs in the different test groups and the control group. Significant differences in frequencies were determined with the chi square test of two proportions. * indicates p≤ 0.05; ** indicates p≤ 0.01; *** indicates p≤ 0.001

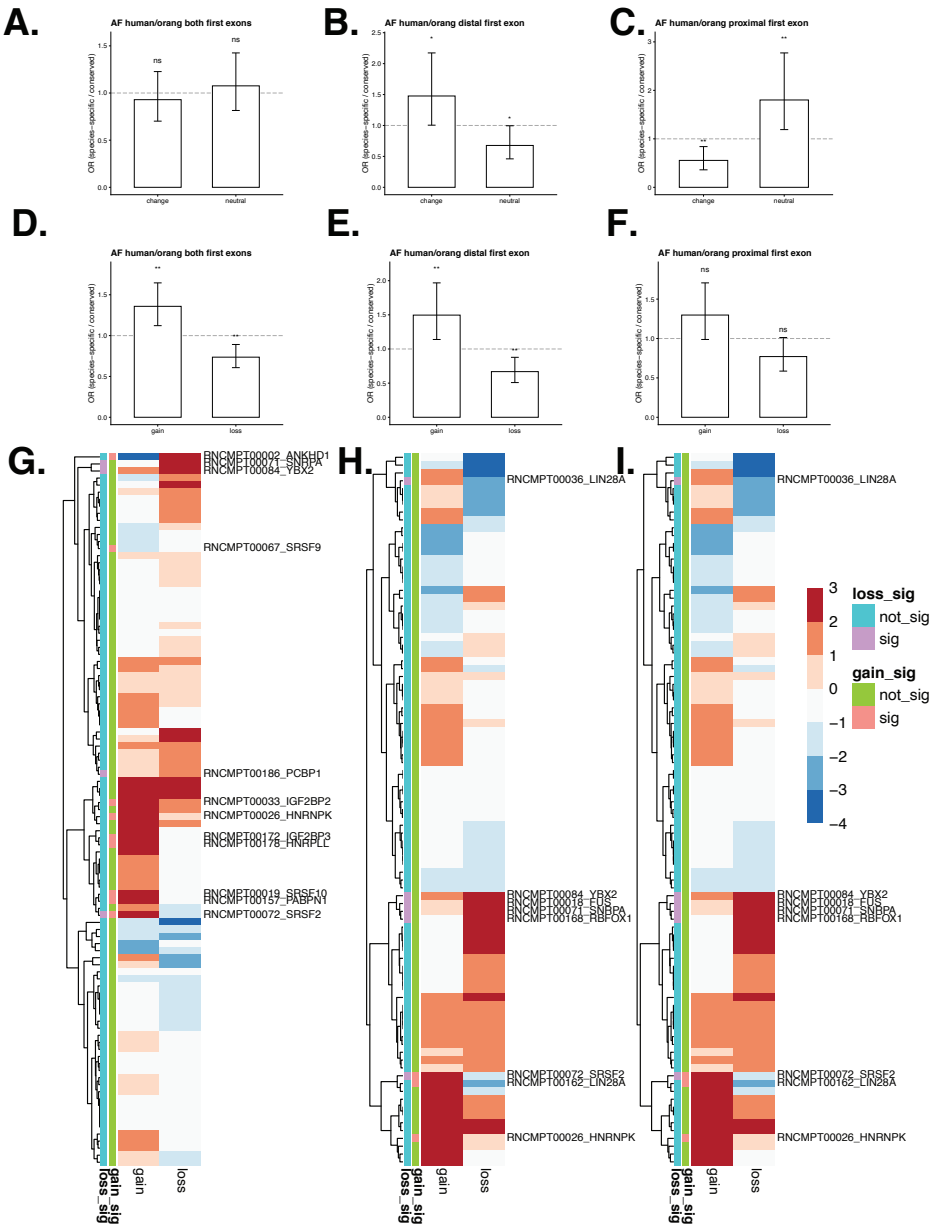

Figure 3.13 [Supplementary Figure 6] Global effects of SNVs found in human and orangutan alternative first AS-TC events.

A-C) Odds ratios (OR) of SNVs identified between human and orangutan AS-TC events that lead to a change or no change ("neutral") in RBP binding as predicted based on RNAcompete data. For both first exons combined, for the distal first exon, and for the proximal first exon. D-F) OR of SNVs identified between human and orangutan alternative first AS-TC events that have been identified to change RBP binding affinity to lead to a gain or loss of a binding site. For both first exons combined, for distal first exon only, for proximal first exon. Bar height indicates OR. Error bars represent the two-tailed 95% confidence interval for the bar height. G-I) Heatmaps showing the enrichment of PWMs in events with an increase in RBP affinity (gain) or with a decrease in RBP affinity (loss) around the identified SNVs. For both first exons combined, for distal first exon only, for proximal first exon. The red/blue color scale indicates the log-fold enrichment over AS event background. The green and teal columns to the side indicate the significance of this enrichment based on a binomial distribution (see methods).

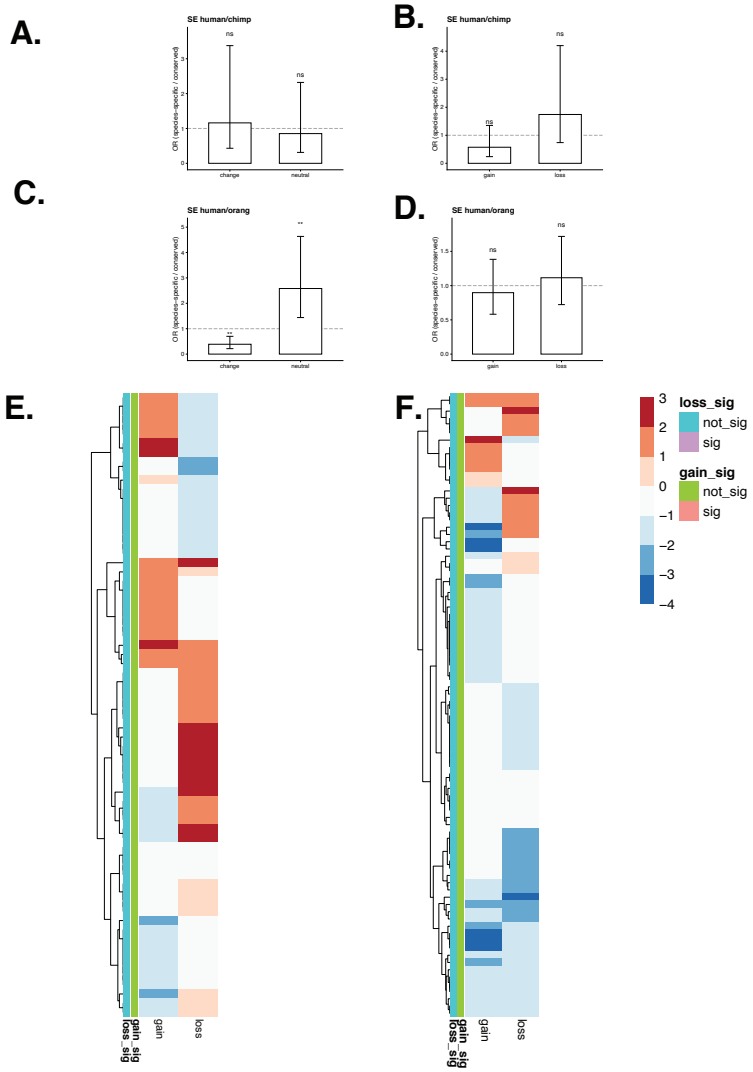

Figure 3.14 [Supplementary Figure 7] Global effects of SNVs found in human and orangutan alternative first AS-TC events.

A) Odds ratios (OR) of SNVs identified between human and chimpanzee skipped AS-TC events that lead to a change or no change ("neutral") in RBP binding as predicted based on RNAcompete data. B) OR for SNVs identified between human and orangutan skipped AS-TC events that lead to a change or no change ("neutral") in RBP binding as predicted based on RNAcompete data. C) OR of SNVs identified between human and chimpanzee skipped AS-TC events that have been identified to change RBP binding affinity to lead to a gain or loss of a binding site. D) OR of SNVs identified between human and orangutan skipped AS-TC events that have been identified to change RBP binding affinity to lead to a gain or loss of a binding site. Error bars represent the two-tailed 95% confidence interval for the bar height. E) Heatmaps showing the enrichment of PWMs in events with an increase in RBP affinity (gain) or with a decrease in RBP affinity (loss) around the identified SNVs from human and chimpanzee sequences. F) . Heatmaps showing the enrichment of PWMs in events with an increase in RBP affinity (gain) or with a decrease in RBP affinity (loss) around the identified SNVs from human and orangutan sequences. The red/blue color scale indicates the log-fold enrichment over AS event background. The green and teal columns to the side indicate the significance of this enrichment based on a binomial distribution (see methods).

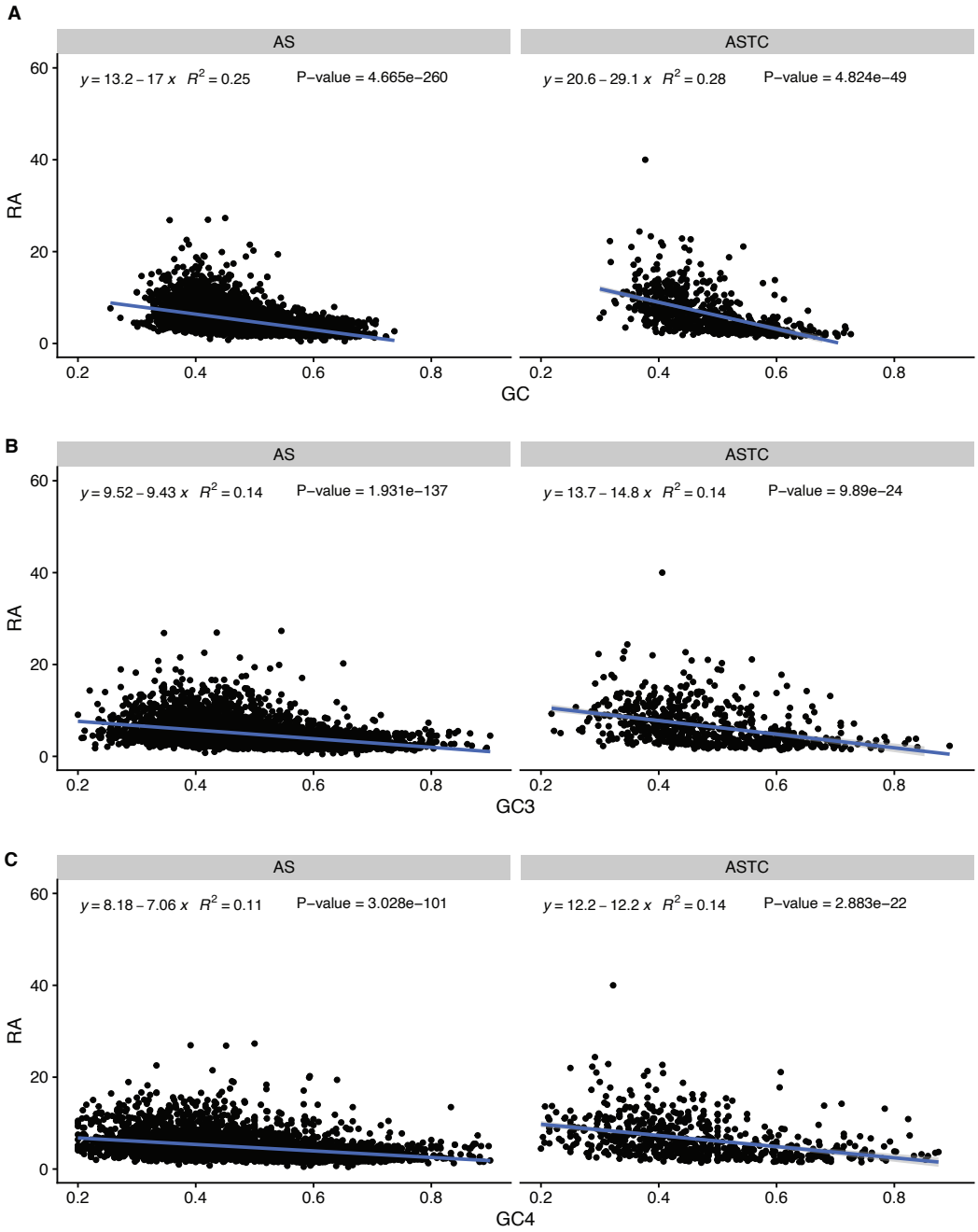

Figure 3.15 [Supplementary Figure 8] Codon content weakly correlates with ribosome association (RA) for AS and AS-TC skipped exons.

A) Correlation between GC content and ribosome association (RA) in AS (left) and AS-TC (right) skipped exon events. B) Correlation of GC3, GC content at the third position of all codons, and ribosome association in AS and AS-TC skipped exons. C) Correlation of GC4, GC content at the third position of four-fold degenerate codons, and ribosome association in AS and AS-TC exons.

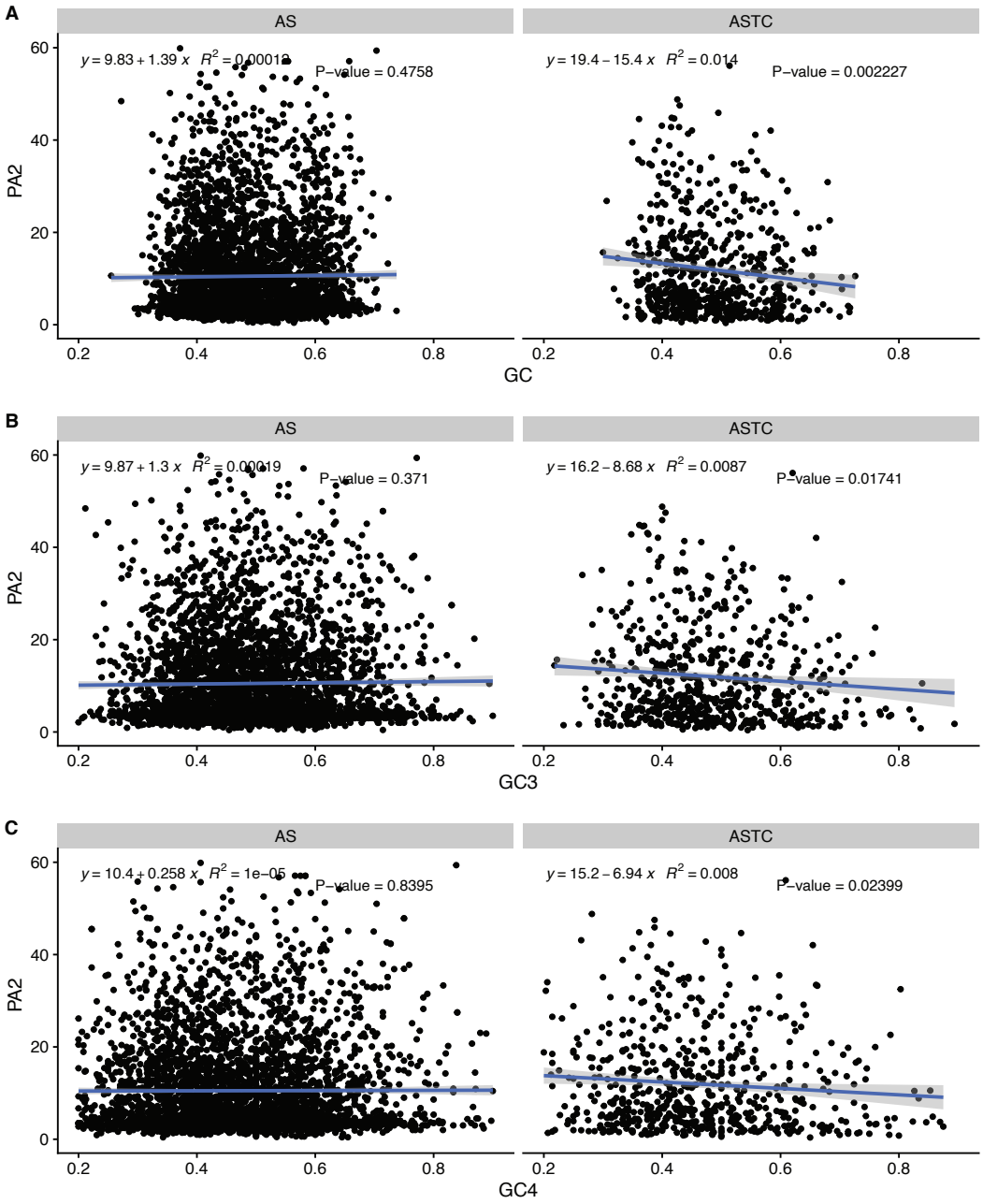

Figure 3.16 [Supplementary Figure 9] Codon content does not correlate with a modified polysome association measure (PA2) for AS and AS-TC skipped exons.

A) Correlation between GC content and a modified polysome association (PA2, see methods) in AS (left) and AS-TC (right) skipped exon events. B) Correlation of GC3, GC content at the third position of all codons, and PA2 in AS and AS-TC skipped exons. C) Correlation of GC4, GC content at the third position of four-fold degenerate codons, and PA2 in AS and AS-TC exons.

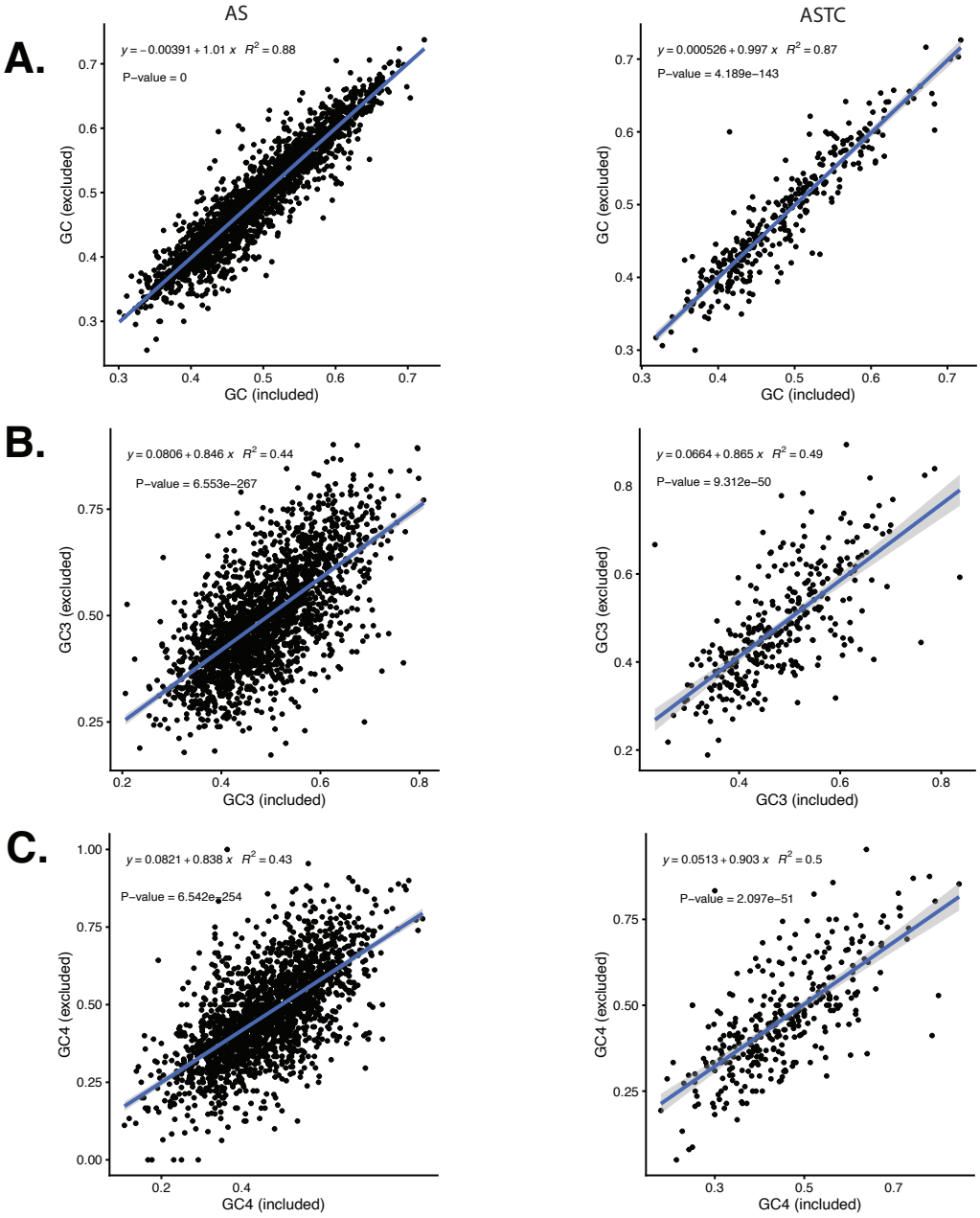

Figure 3.17 [Supplementary Figure 10] Correlation of GC content between included and excluded isoform pairs per ASTC event.

Correlation of GC and codon content between included and excluded isoform pairs per AS or ASTC event. A) Correlation of GC content between included and excluded isoform in AS (left) and ASTC (right) events in human. B) Correlation of GC3 content between included and excluded isoform in AS (left) and ASTC (right) events in human. C) Correlation of GC4 content between included and excluded isoform in AS (left) and ASTC (right) events in human.

4. Chapter 4: Splicing Factor SRSF1 expands the regulatory logic of microRNA expression

Marija Dargyte[1], Julia Philipp[1], Christina D. Palka[2], Michael D. Stone[2], and Jeremy R. Sanford[1*]

[1] University of California Santa Cruz, Department of Molecular, Cellular and Developmental Biology, CA, 95064, USA

[2] University of California Santa Cruz, Department of Chemistry and Biochemistry, CA,95064, USA

**Corresponding Author*

4.1 Abstract

The serine and arginine-rich splicing factor SRSF1 is an evolutionarily conserved, essential pre-mRNA splicing factor. Through a global protein-RNA interaction survey we discovered SRSF1 binding sites 25-50nt upstream from hundreds of pre-miRNAs. Using primary miRNA-10b as a model we demonstrate that SRSF1 directly regulates microRNA biogenesis both *in vitro* and *in vivo*. Selective 2' hydroxyl acylation analyzed by primer extension (SHAPE) defined a structured RNA element located upstream of the precursor miRNA-10b stem loop. Our data support a model where SRSF1 promotes initial steps of microRNA biogenesis by relieving the repressive effects of cis-regulatory elements within the leader sequence.

4.2 Introduction

MicroRNAs (miRNAs) are important regulators of post-transcriptional gene expression. Nearly 60% of human protein coding genes contain conserved miRNA target sites (Friedman et al. 2009). Given the importance of miRNAs in gene regulation, it is not surprising that spatial and temporal expression patterns of miRNAs are tightly

regulated. Canonical miRNA biogenesis begins with transcription of a primary miRNA (pri-miRNA) by RNA polymerase II. In the nucleus, the pri-miRNA folds into a hairpin structure which is excised by the Microprocessor complex consisting of Drosha and DGCR8, yielding a precursor miRNA (pre-miRNA). Upon transport to the cytoplasm the hairpin is cleaved, by Dicer, into a 22nt miRNA duplex. The less thermodynamically stable strand is preferentially loaded into RISC by catalytic Argonaute protein, Ago2 (Noland & Doudna 2013). Although the major catalytic steps of miRNA biogenesis and downstream RISC targeting are well understood, the regulatory checkpoints are only emerging.

RNA binding proteins are broadly implicated in miRNA biogenesis. The terminal loop region of the hairpin is a central target for many RBPs (Nussbacher & Yeo 2018; Treiber et al. 2017). For example, Lin28 binds to the terminal loop of let-7 family members recruiting TUT4 for uridylation (Heo et al. 2009). Competition between KSRP and hnRNP A1 binding to the terminal loop of pri-miR-18a influences processing by Drosha/DGCR8 (Guil & Cáceres 2007; Michlewski & Cáceres 2010). The functional importance of the terminal loop in regulation of miRNA biogenesis is underscored by strong phylogenetic conservation of this sequence element across vertebrates. In addition to the terminal loop, other sequence elements within the pri-miRNA are implicated in regulation of biogenesis (Michlewski & Caceres 2018).

The serine and arginine-rich (SR) protein family are evolutionarily conserved RNA binding proteins. Named for their Arg-/Ser-rich carboxyl terminal domain (RS domain), these proteins have diverse functions in post-transcriptional gene regulation including, pre-mRNA splicing, mRNA export, mRNA decay, nonsense mediated decay and mRNA translation (Howard & Sanford 2015). SR proteins are essential splicing fac-

tors and required for pre-mRNA splicing *in vitro* and *in vivo* (Zahler et al. 1993; Krainer et al. 1991; Li & Manley 2005). During spliceosome assembly, SR proteins, through the RS domain, promote splice site recognition via splicing factor recruitment (Zhu & Krainer 2000). Alternatively, the RS domain may function to promote RNA-RNA interactions by neutralizing electrostatic interactions between U snRNAs at the 5'ss and branch point sequence (Shen & Green 2006).

Previous work from our lab and others demonstrated that SR proteins interact with non-coding mRNA transcripts (Sanford et al. 2009; Royce-Tolland et al. 2010; Tripathi et al. 2010). By contrast to their roles in pre-mRNA splicing, the functional roles of SR proteins in small RNA expression remain poorly described. Two members of the SR protein family, SRSF1 and SRSF3, have been implicated in miRNA biogenesis. SRSF3 recognizes a sequence determinant located downstream of the basal junction in hundreds of pri-miRNAs (Kim et al. 2018; Auyeung et al. 2013). Whereas, SRSF1 promotes processing of pri-miR-7 by binding to the lower stem, although its mechanism remains unclear (Wu et al. 2010).

Here we report the discovery of a new sequence determinant of miRNA biogenesis. Using ENCODE eCLIP data, we discovered that a wide array of RBPs interact with pri-miRNAs. Remarkably, we found the region 25-50nt upstream of miRNA hairpins was a frequent ligand for RBPs, including the pre-mRNA splicing factor SRSF1. We validated these data using iCLIP, which identified hundreds of pri-miRNAs in HEK293T cells with strong cross linking signals 35-50nt upstream of the 5' of the hairpin, which we name the 5' leader sequence. We demonstrate that SRSF1 expression levels correlate with decreased levels of pri-miRNAs and a concomitant increase in functional miRNA activity. Using pri-miR-10b as a model, we determine that SRSF1 binding sites are nec-

essary for SRSF1-dependent stimulation of miRNA biogenesis. Taken together our data demonstrate, for the first time, an upstream determinant required for SRSF1 directed regulation of miRNA biogenesis.

4.3 Results and Discussion

4.3.1 Global analysis of primary miRNA-protein interactions

To identify RBPs that preferentially interact with sequences outside of the hairpin, we used the ENCODE consortium enhanced crosslinking immunoprecipitation and high throughput sequencing (eCLIP-seq) data (Van Nostrand et al. 2016). We compiled more than 120 protein-RNA interactions in HepG2 and K562 cells (Supplemental Table 1). We set a range to genomic regions 100nt upstream and 200nt downstream of the 5' end of pre-miRNAs, as annotated by Gencode. Using aggregated eCLIP peaks for all RBPs in the ENCODE database, we observed a wide array of interactions across pri-miRNAs, including a prominent region near the terminal loop region (Figure 4.1B). We also noted pronounced, but broadly distributed binding sites upstream of the 5' end of pre-miRNA. (Figure 4.1B) To determine how specific RBPs interact with pri-miRNAs we plotted the binding site density for individual RBPs, with binding sites in at least 7 unique miRNAs. Unsupervised hierarchical clustering revealed that different RBPs preferentially associate with specific regions of pri-miRNAs (Figure 4.1A). For example, Lin28B interacts specifically with a region encompassing the terminal loop, a finding that is well-aligned with previous studies (Choudhury & Michlewski 2012). By contrast, we noted several splicing factors, including SRSF1 and U2AF1, with preferential binding sequences upstream of the pre-miRNA (Figure 4.1A and Supplemental Table 2).

Using published CLIP-seq and iCLIP experiments from our lab we validated the interaction of SRSF1 and the 5' end of pre-miRNAs (Howard et al. 2018; Sanford et al.

2009). As expected, most SRSF1 binding sites identified by CLIPper in protein coding genes were associated with exonic sequences (Supplemental Figure S1C). We also observed a purine-rich motif enriched in sequences corresponding to SRSF1 binding sites (Supplemental Figure S1D). At a single nucleotide resolution, crosslinking density was significantly higher in exon than intron sequences, consistent with previous studies (Supplemental Fig. S1E; Sanford et al. 2009; Sanford et al. 2008; Änkö et al. 2012). We used the 5' end of SRSF1 iCLIP reads to approximate the crosslinking position of SRSF1 on hundreds of pri-miRNAs (Supplemental Table 2). In agreement with eCLIP data, we observed a non-uniform distribution of SRSF1 crosslinking density relative to the 5' end of pre-miRNAs, with a strong bias to positions 50nt upstream of the 5' end of the pre-miRNA (Figure 4.1C). SRSF1 was previously linked to regulation of miRNA processing, although the mechanism was not described (Wu et al. 2010). A curious finding from the prior study was that SRSF1 recognized a consensus binding motif located in the basal region of the pre-miR-7 stem loop. By contrast, eCLIP and iCLIP show SRSF1 interacts with sequences upstream of pre-miRNAs.

To discriminate between these two hypotheses, we asked if SRSF1 overexpression influenced mature miRNA activity. We generated luciferase reporters containing target sites for specific miRNAs within their 3'UTR. Individual miRNA reporter constructs or a control reporter lacking the heterologous miRNA target site were co-transfected with T7-SRSF1 or a control plasmid into HEK293T cells. If SRSF1 stimulates either mature miRNA activity or expression we expect to see a decrease in reporter activity or an increase in repression. In all cases, we observed significant reduction in reporter activity relative to controls upon T7-SRSF1 overexpression (Figure 4.2B). These data suggest that SRSF1 promotes maturation of miRNAs rather than simply reducing pri-miRNA levels. To determine if these changes in reporter activity are specific to SRSF1 we also co-transfected HEK293T cells with the same reporter constructs as well as hnRNPA1,

another RBP linked to the biogenesis of specific miRNAs. As expected, over-expression of hnRNPA1 enhanced miR-17 activity (Kooshapur et al. 2018). By contrast, hnRNPA1 had no effect on let-7-a1 or miR-10b reporter activity (Supplemental Figure S3B).

SRSF1 shuttles continuously from the nucleus to the cytoplasm and is intimately involved in mRNA processing, stability and translation (Das & Krainer 2014). To determine if SRSF1 influences a nuclear or cytoplasmic step in miRNA biogenesis we co-transfected luciferase reporters with wild type SRSF1 or a non-shuttling mutant that is retained in the nucleus (Cazalla et al. 2002). If SRSF1 promotes pre-miRNA export from the nucleus or Dicer activity in the cytoplasm, then we predict that the non-shuttling mutant would be unable to stimulate miRNA activity. By contrast, we observed that relative to wild type, the non-shuttling mutant (SRSF1-NRS) exhibits enhanced repression of the miR-10b reporter (Supplemental Figure S3C). These data suggest that SRSF1 promotes a nuclear step in the miRNA biogenesis pathway, as previously suggested by the processing of miR-7 (Wu et al. 2010).

iCLIP revealed SRSF1 interactions with pri-miRNA 5' leader sequences at single nucleotide resolution. To determine if these points of interaction are functionally relevant for miRNA processing, we generated a series of pri-miR-10b expression constructs containing point mutations at SRSF1 crosslinking sites. If SRSF1 directly promotes miRNA biogenesis, then we predict that mutation of SRSF1 interaction sites could attenuate the effect of SRSF1 on miRNA activity and expression. As expected, driving pri-miR-10b expression up in HEK293T cells strongly reduced luciferase activity relative to the negative control expression construct (Figure 4.3A). Overexpression of SRSF1 further reduced miR-10b luciferase reporter activity. By contrast, pri-miR-10b expression constructs containing crosslinking site mutant 2 attenuated the effect of SRSF1 on miR-10b luciferase reporters. Similarly we observe a loss of detectable mature miR-10b with

mutant 2 overexpression compared to wild type pri-miR-10b (Figure 4.3B). Likewise, we do not observe significant luciferase repression changes between wild type or mutant pri-miR-10b for control experiments lacking SRSF1 overexpression (Figure 4.3A). Taken together this experiment reveals at least one cis-acting RNA element recognized by SRSF1 functions in regulation of miR-10b expression.

To determine if crosslinking site mutations interfere with SRSF1 pri-miR-10b interactions, we performed filter binding assays using purified recombinant SRSF1 (rSRSF1) and RNA binding-deficient mutants (Supplemental Figure S4A). Recombinant SRSF1 binds pri-miR-10b with an apparent Kd of 31.64nM (Figure 4.5A and C). As expected, rSRSF1 harboring point mutations of two solvent exposed phenylalanines in RRM1 are mutated to aspartates (FF->DD) reduce affinity for RNA binding (Supplemental Figure S5A). Likewise, deletion of the RS domain reduces RNA binding. As expected, the point mutation in pri-miR-10b which attenuates SRSF1-dependent regulation of miR-10b activity and expression weakens the affinity of SRSF1 for pri-miR-10b *in vitro* (Supplemental Figure S5). Although affinity for the pri-miR-10b mutants is reduced *in vitro* we cannot discount any *in vivo* interactions that are not accounted for by filter binding. These data indicate that the SRSF1 RS domain is required for binding to pri-miR-10b and that mutations within the leader do reduce SRSF1 affinity for pri-miR-10b.

4.3.2 Identification of a repressive element in the 5' leader of pri-miR-10b

The experiments described above indicate that sequences beyond the hairpin regulate pri-miRNA processing. To test this hypothesis, we created a series of deletion mutants from the 5' or 3' of pri-miR-10b (Figure 4.4). If either the 5' or 3' flanking sequences are required for mature miRNA activity, we expect an increase in miR-10b luciferase reporter activity. If the mutations remove repressive elements, we expect a de-

crease in luciferase activity. To distinguish between these possibilities we co-transfected expression constructs for wild type pri-miR-10b or 5' or 3' deletion mutants, along with the miR-10b luciferase reporter. We observed significant decrease in luciferase activity for the more extreme 5'd2 and 5'd3 mutants, but not the more conservative 5'd1 mutant (Figure 4.4B). These data suggest that there are sequence or structural repressive elements within the SRSF1 binding sites 5' of the hairpin.

To determine if the 5' leader of pri-miR-10b contains structured RNA elements we performed chemical probing using 1-methyl-7-nitroisatoic anhydride (1M7) SHAPE reagent. 1M7 modifies the 2' hydroxyl of unpaired residues. Modified ribose residues are revealed as termination sites by primer extension. Using reactive positions to constrain secondary structure predictions reveals the presence of a canonical miR-10b hairpin containing the embedded mature miRNA (Figure 4.4C). Surprisingly, a well defined stem loop structure emerges just upstream of the hairpin (Figure 4.4C). By contrast, the point mutations that reduce the effect of SRSF1 on miR-10b activity significantly reduce reactivity within the loop region of this novel structural element, suggesting a change in secondary structure of the pri-miR-10b leader sequence (Supplemental Figure S6).

To determine if a structured 5' leader was a general feature of pri-miRNAs bound by SRSF1, we compared the thermodynamic stability of pri-miRNAs predicted to be bound by SRSF1 to those lacking iCLIP signal signal. Using the DINAmelt web server application, Quikfold, we were able to generate -dG values for predicted secondary structures of pri-miRNAs (Markham & Zuker 2005). We observed a slight, yet significant difference in the distribution of -dG between those primary miRNAs bound by SRSF1 and those that are not (Supplemental Figure S7A). These data suggest that perhaps there is a structured element within the 5' leader sequence of SRSF1 bound pri-miRNAs.

4.3.3 SRSF1 directly influences miRNA biogenesis

Taken together, our results suggest that SRSF1 promotes a nuclear step of miRNA processing, and likely before initial cleavage by Drosha. Therefore we reasoned that SRSF1 may enhance Microprocessor complex activity. To test if SRSF1 directly influences the Microprocessor step of miRNA biogenesis we performed *in vitro* miRNA processing assays with immunopurified Drosha/DGCR8 in the presence or absence of rSRSF1 (Figure 4.5A). In control reactions without rSRSF1 we observed a gradual increase in product formation over the course of the reaction (Figure 4.5A, lanes 1-6). However, when pri-miR-10b was incubated in the presence of rSRSF1 we observed a significant increase in the rate of product formation (Figure 4.5A, lanes 7-10). Quantification of replicate experiments revealed that SRSF1 enhances rates of product formation compared to control reactions (Figure 4.5B). Because SRSF1 promotes pri-miR-10b processing by Drosha/DGCR8 and that Drosha contains an RS domain, it is possible that SRSF1 recruits Drosha to the pri-miRNA transcript. To test if SRSF1 directly interacts with the Microprocessor complex we probed proteins coprecipitated with Drosha by western blot. We were unable to observe any RNA-dependent or -independent interactions between exogenously expressed SRSF1 and the Microprocessor complex (Supplemental Figure S7C). Overall, our data suggests that SRSF1 promotes pri-miRNA biogenesis by altering the conformation of the 5' leader sequence prior to Drosha cleavage.

In this study we showed that the SR protein SRSF1 promotes the first steps in miRNA processing. Global analysis of protein-RNA interactions by iCLIP and eCLIP revealed that SRSF1, as well as other splicing factors, engage binding sites upstream of pre-miRNAs (Figure 4.1). Reporter assays demonstrated that SRSF1 enhances miRNA function *in vivo* and that cis-acting SRSF1 binding sites within pri-miR-10b are required. Our data suggests that this 5' leader sequence is inhibitory, and needs to be

relieved for efficient processing. Alleviating a repressive domain for miRNA biogenesis has been previously described and well supported by our data as well (Du et al. 2015). This observation is strongly supported by *in vitro* processing assays, which show that rSRSF1 accelerated the cleavage rate of pri-miR-10b. Coimmunoprecipitation experiments failed to detect an interaction between SRSF1 and Drosha, arguing against a recruitment model. Instead, we suggest that SRSF1 may influence the conformation of the pri-miRNA. Using SHAPE we noted the presence of a strong stem loop structure within the 5' leader region of primary miR-10b. Deletion analysis suggests the 5' leader region interferes with miR-10b expression. Taken together our data suggest that SRSF1 binding to pri-miR-10b alters the conformation of an inhibitory stem loop structure.

Despite decades of research, the mechanisms through which SR proteins regulate post-transcriptional gene expression remain unclear. Competing models include RS domain recruitment of splicing factors and RNA-RNA interaction chaperones (Graveley & Maniatis 1998; Shen & Green 2006). Previously, ATP-independent RNA annealing activity was copurified with SRSF1 (Krainer et al. 1990), suggesting that SRSF1 disrupted intramolecular RNA structure formation to promote intermolecular annealing at temperatures well below the Tm. One prediction is that SRSF1 relieves inhibitory secondary structures in the 5' leader sequence. We believe such a mechanism is consistent with our observations using pri-miR-10b as a model. This structural change could serve as a checkpoint in hairpin selection by the Microprocessor. A similar licensing step was described for processing of the pri-miR-17 92 cluster (Du et al. 2015).

The results presented here, demonstrate that SRSF1 promotes miRNA processing without directly recruiting the Microprocessor. Given the recent discovery that SRSF3 influences miRNA processing through interactions with the basal junction (Kim et al. 2018). We hypothesize that SRSF1 and SRSF3 may function collaboratively, by 5' and 3' interactions respectively, to define the hairpin for miRNA processing. This process

likely involves remodeling inhibitory secondary structure adjacent to the stem loop and consistent with an RNA chaperone function for SRSF1 in miRNA biogenesis.

4.4 Materials and Methods

4.4.1 Analysis of eCLIP and iCLIP datasets

eCLIP data was downloaded from the ENCODE consortium through their dashboard. Peak definitions from HEPG2 cells were aligned relative to the 5' end of miRNA precursors. Data were visualized following unsupervised hierarchical clustering. Only RBPs with at least 7 annotated binding sites near miRNAs were considered in this analysis (Supplemental Table 1 and 2). iCLIP data for SRSF1 was downloaded from (GSE #GSE83923). Reproducible crosslinking sites were defined as previously described (Howard et al. 2018). Crosslinking density was calculated for all SRSF1 crosslinking data relative to the 5'end of miRNA precursors.

4.4.2 Cell culture and transfections

Hek293T cells were grown in 6 well plates with DMEM supplemented with 10% FBS. At 70% confluence cells were transfected with plasmids using polyethylenimine (PEI) and 0.35M NaCl. Each transfection was performed a minimum of three times with two technical replicates per experiment.

4.4.3 RNA purification and RT-qPCR

Total RNA for RT-qPCR was isolated using Direct-zol RNA MiniPrep Kit (Zymo Research) for all other experiments RNA was isolated using standard Tri-reagent (Sigma) protocol. cDNA was reverse transcribed from 1ug of total RNA using High-Capacity cDNA reverse transcriptase kit (Applied Biosystems). qPCR was performed using Titanium Taq (Clontech) and SYBR Green on a Roche Lightcycler 480 (Roche Diagnostics) according to MIQE guidelines (Bustin et al. 2009).

4.4.4 Luciferase reporter assays

Seed sites for let-7a-1, miR-15b, miRNA17, miR 19a, miR 93, and miR 128a were inserted into the 3'UTR of pMIR luciferase reporter (Life Scientific). miR-10b reporters described previously (Ma et al. 2007) were obtained from AddGene. Reporters were co-transfected with Renilla luciferase (Promega) reporter as a transfection efficiency control. Luciferase activity was assayed 24 hours post transfection using Dual-Glo Luciferase Assay System (Promega). For a 24-well plate, each well was transfected with 100ng of TK-rLUC (Promega), 800ng or 1ug of T7-SRSF1 or control plasmid (Cáceres et al. 1997), 400ng of pMIR Luciferase reporter (Life Scientific). Experiments in which exogenous pri-miR-10b was used, cells were transfected with 200ng of pGK (control) or pGK 10b (Ma et al. 2007; Cáceres et al. 1997).

4.4.5 In vitro transcription

20ug of linearized plasmid with BamHI (New England Biolabs) of which 2ug was transcribed with MEGAscript T3 polymerase (ThermoFisher). Transcripts were labeled with alpha-32P UTP for *in vitro* processing and filter binding. Following transcription, RNA was phenol/chloroform extracted and ethanol precipitated. RNA was resolved on a 6% denaturing polyacrylamide gel and extracted with a clean razor. Gel containing RNA was incubated overnight at 42°C in elution buffer (0.3M NaOAc pH 5.5, 2% SDS). RNA was ethanol precipitated and stored at -20°C until use.

4.4.6 In vitro miRNA processing

FLAG-immunoprecipitate *in vitro* processing assays were performed based on (Lee et al. 2003). Briefly, HEK293T cells were transfected with equal concentrations of FLAG-Drosha and HA-DGCR8 which were FLAG immunoprecipitated according to protocol. Processing reactions were incubated as a time course up to 90 minutes at

37°C. RNA was phenol/chloroform extracted and resolved on a 10% denaturing polyacrylamide gel. After drying, gel was exposed on a phosphor screen, visualized with a Typhoon image scanner (GE Healthcare), and was analyzed using ImageJ. Data was standardized between gels by normalizing the ratio of pre-to pri-miRNA bands from the 90 minute time point.

4.4.7 Northern blot

10-25ug of total RNA was resolved on a 12.5% denaturing polyacrylamide gel and transferred to a positively charged nylon membrane (GE Healthcare) using 1x TBE. After UV crosslinking, membrane was prehybridized with ULTRAhyb hybridization buffer (Invitrogen) for 30 minutes at 42°C. The membrane was hybridized to a 32P end labeled oligo probe overnight at 42°C. Membrane was washed with 2X SSC, 0.05% SDS, and 0.1X SSC, 0.1% SDS respectively for 30 minutes at 42°C. Blot was visualized using Typhoon image scanner (GE Healthcare) and analyzed using ImageJ.

Dissertation author contribution:

JP: Meta-analysis of publicly available eCLIP data and differential expression anaysis of small RNA-seq

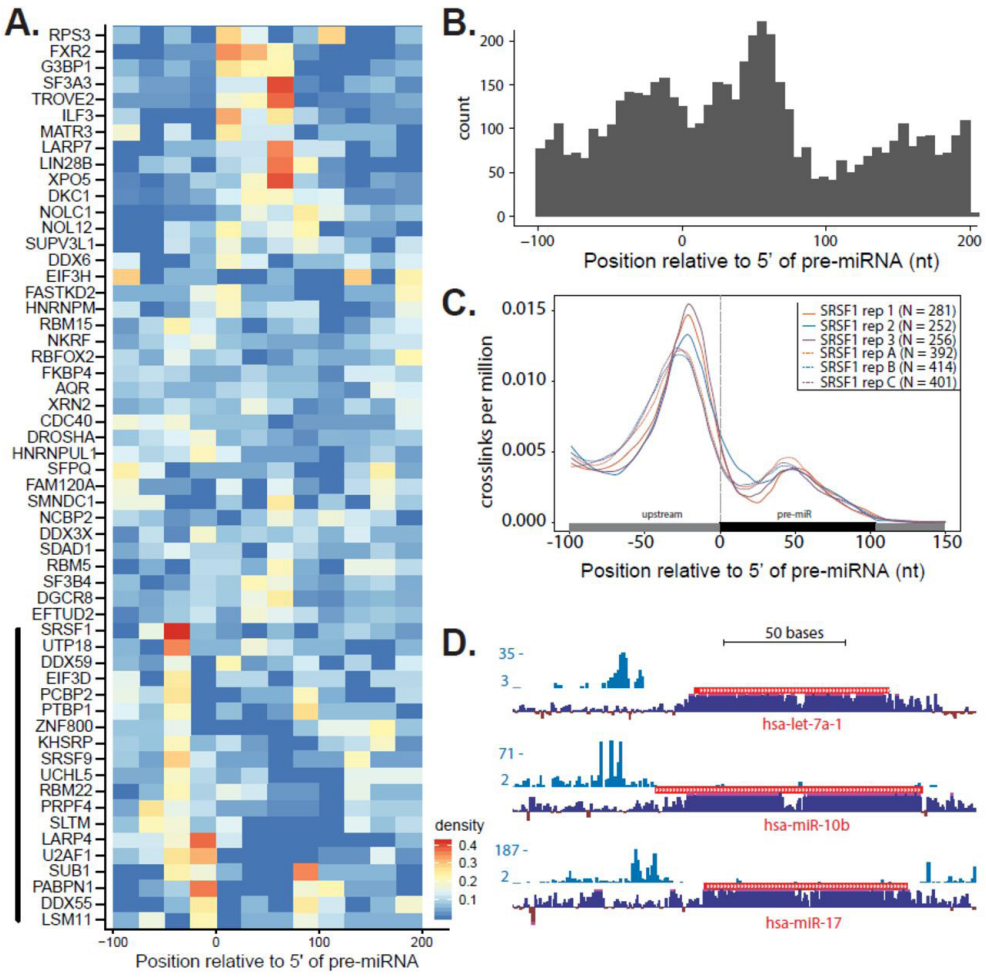

Figure 4.1 Meta analysis of eCLIP data and iCLIP data characterizes a relationship between RBP binding and pri-miRNAs.

(A) Heatmap depicting specific RBP interactions along a subset of pri-miRNA transcripts using HepG2 eCLIP data. Horizontal axis denotes distance from the 5' end of miRNAs by bins in 25nt. (B) Histogram of all RBP localizations in HepG2 cells relative to pri-miRNAs. 0 denotes 5' end of pre-miRNAs annotated in the UCSC genome browser. (C) SRSF1 iCLIP crosslinks density relative to pri-miRNA for six replicates under two conditions. (D) UCSC genome browser screenshots of three exemplar pre-miRNAs with SRSF1 binding. Blue histogram is SRSF1 crosslinking density. Red track is pre-miRNA genes. Purple histogram is 100 vertebrate conservation.

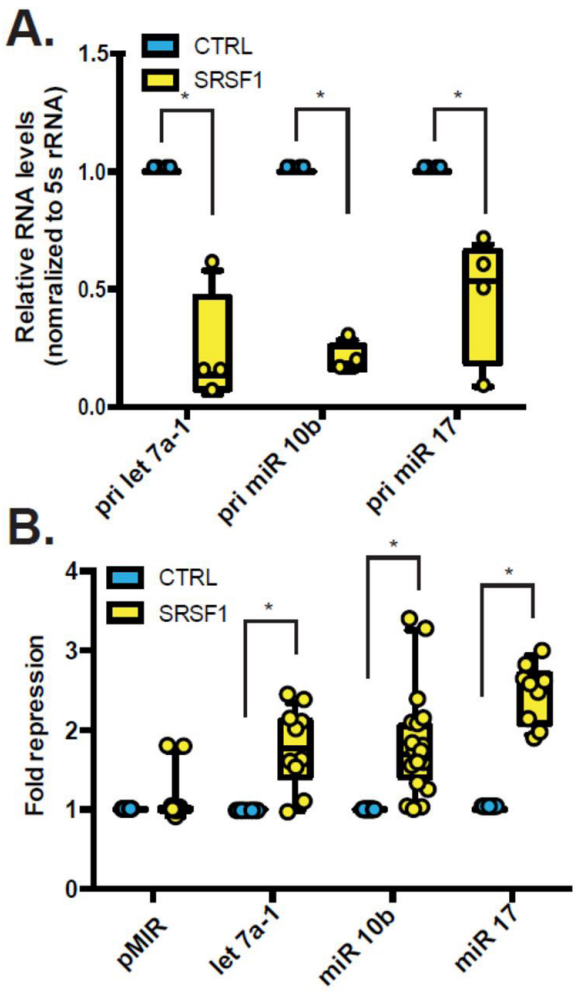

Figure 4.2 SRSF1 dependent miRNA expression and activity.
(A) Relative quantification of RT-qPCR of pri-miRNAs let 7a-1, miR-10b, and miR-17. (B) Normalized luciferase reporter expression (fold repression) for let 7a-1, miR-10b, and miR-17 reporters for control (cyan) or SRSF1 overexpression (yellow). (*) P <0.05 using unpaired t-test.

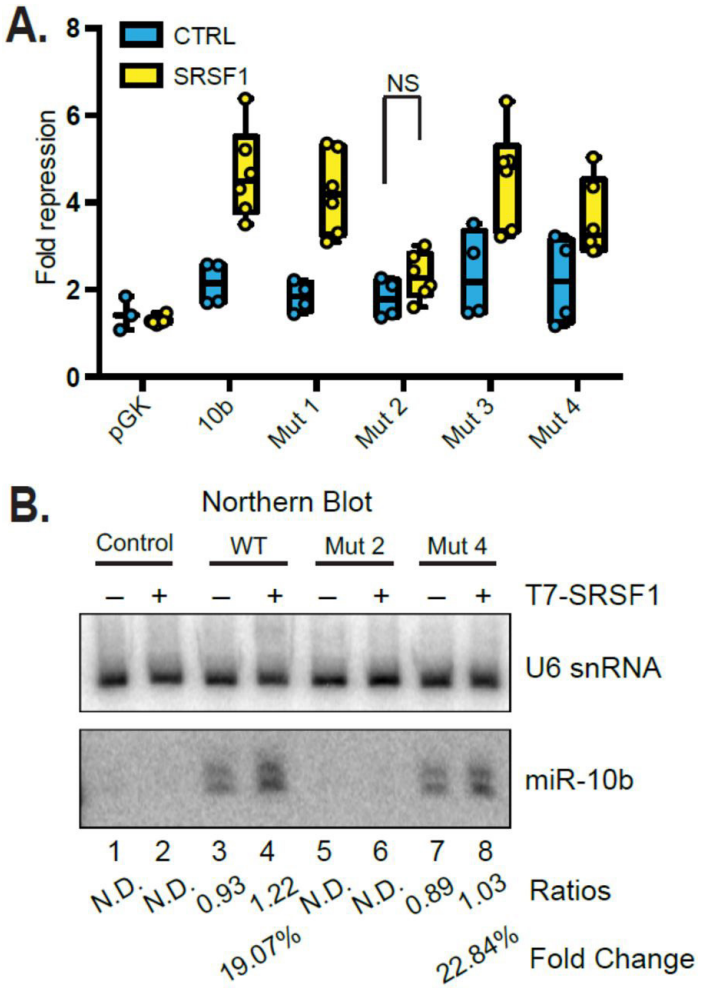

Figure 4.3 Mutations within SRSF1 binding site alter miR-10b expression and activity.
(A) Normalized luciferase reporter expression (fold repression) for miR-10b reporter for control plasmid (cyan) or SRSF1 overexpression cells(yellow). (B) Northern blot of U6 snRNA (control) and mature miR-10b with overexpression of exogenous pri-miR-10b constructs along with SRSF1.

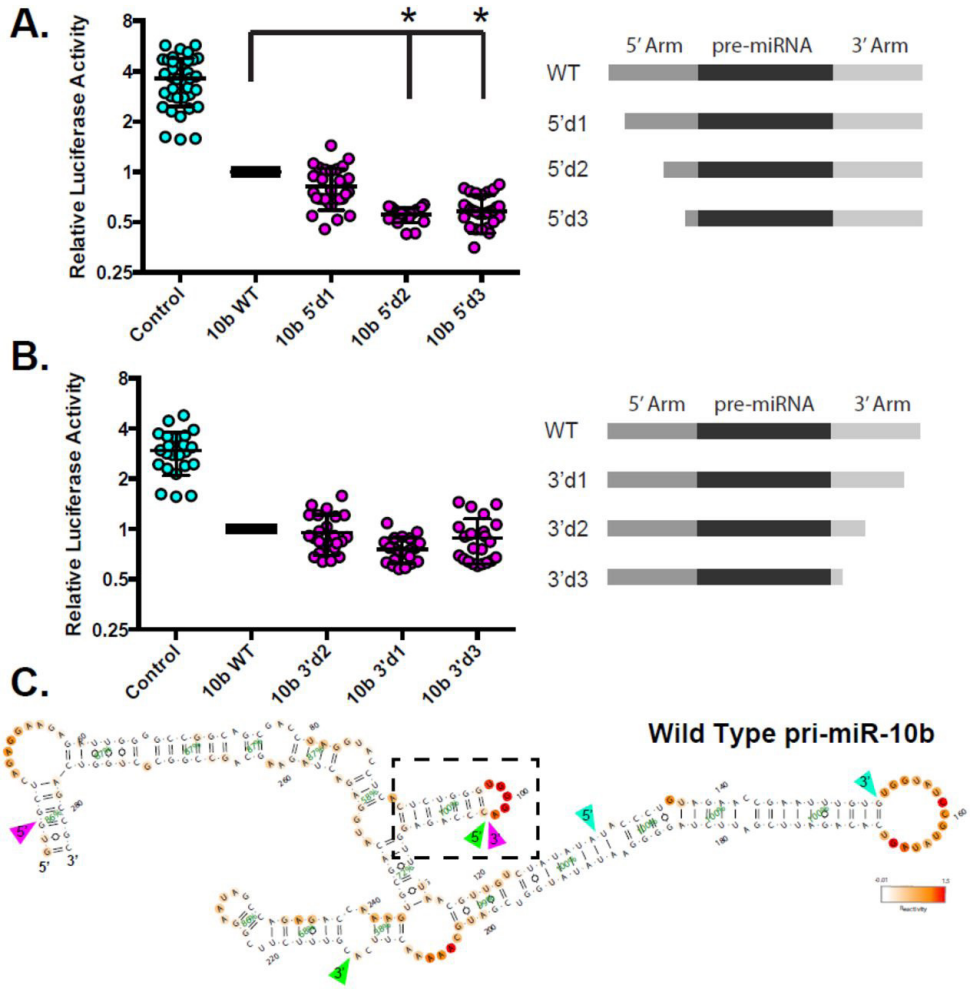

Figure 4.4 An upstream structure of pri-miR-10b influences mature miR-10b activity.
(A,B) Relative luciferase activity when HEK293T cells are transfected with exogenous pri-miR-10b truncation mutations. Schematic depicts regions of truncations relative to pre-miR-10b. (*) P <0.05 using unpaired t-test. (C) Chemical probing of pri-miR-10b by 1M7. Cyan arrows denote the parameter of embedded mature miR-10b, green arrows denote the parameter of embedded pre-miR-10b, and pink arrows denote the parameter of SRSF1 crosslinking region from iCLIP. Nucleotide accessibility to 1M7 are marked by warmer colors. Note boxed, the presence of a small and stable hairpin upstream of the 5' apical stem of miR-10b.

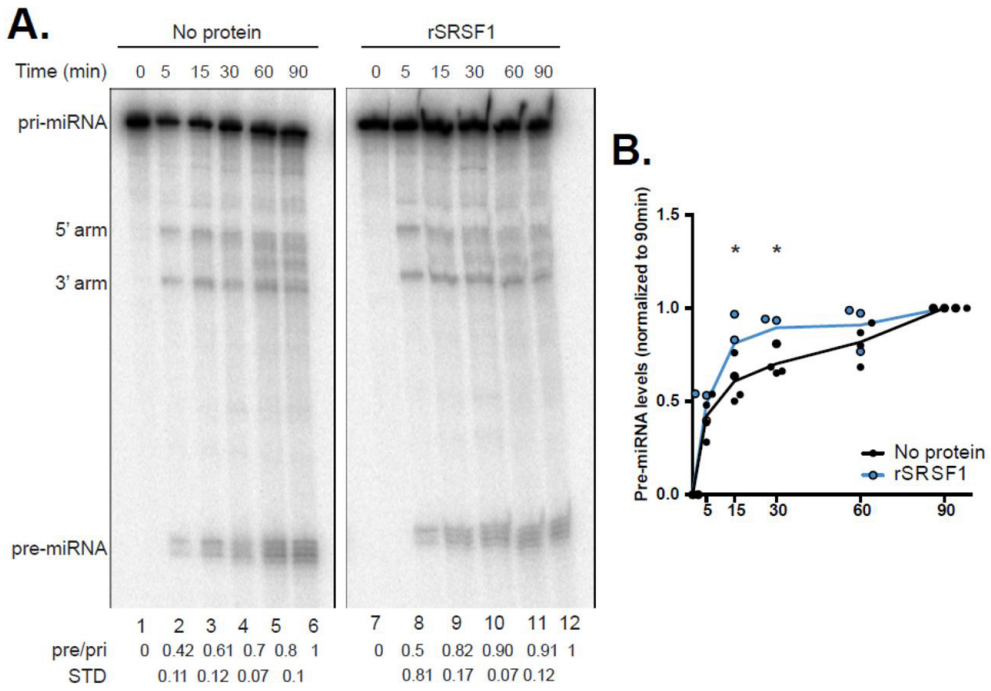

Figure 4.5 SRSF1 directly alters the rate of pri-miRNA processing.
(A) In vitro processing time course of pri-miR-10b constructs by FLAG pulldown Microprocessor complex in presence or absence of rSRSF1. 5' or 3' arms cleaved during processing are labeled. Pri-to pre-miR-10b ratios are calculated for three replicate experiments. (B) Quantification of pre-miR-10b accumulation over 90 minutes. In presence of SRSF1 (blue) pre-miR-10b accumulated significantly faster starting at 15 minutes.

5. Chapter 5: Post-transcriptional gene regulation by the RNA binding protein IGF2BP3 is critical for MLL-AF4 mediated leukemogenesis

Tiffany M Tran[1,2], Julia Philipp[3], Jaspal Bassi[1], Neha Nibber[1], Jolene Draper[3], Tasha Lin[4,5], Jayanth Kumar Palanichamy[1,6,] Amit Kumar Jaiswal[1], Oscar Silva[7], May Paing[1], Jennifer King[8], Sol Katzman[3], Jeremy R. Sanford[3], Dinesh S. Rao[1,2,9,10,11*]

[1] Department of Pathology and Laboratory Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California, USA.

[2] Molecular, Cellular and Integrative Physiology Graduate Program, UCLA, Los Angeles, California, 90095, USA.

[3]Department of Molecular, Cellular and Developmental Biology, UCSC, Santa Cruz, California, 95064, USA.

[4]Division of Hematology/Oncology, Department of Medicine at UCLA, Los Angeles, California, 90095, USA.

[5] Molecular Biology Graduate Program, UCLA, Los Angeles, California, 90095, USA.

[6]Department of Biochemistry, All India Institute of Medical Sciences, New Delhi, 110029, India

[7]Department of Pathology, Stanford University School of Medicine, Stanford, California, 94305, USA.

[8]Division of Rheumatology, Department of Medicine at UCLA, Los Angeles, California, 90095, USA.

[9]Jonsson Comprehensive Cancer Center (JCCC) and

[10]Broad Stem Cell Research Center, UCLA, Los Angeles, California, 90095, USA.

[11]Lead Contact

5.1 Abstract

Despite recent advances in therapeutic approaches, patients with MLL-rearranged leukemia still have poor outcomes and a high risk of relapse. Here, we found that MLL-AF4, the most common MLL fusion protein in patients, transcriptionally induces IGF2BP3 and that IGF2BP3 strongly amplifies MLL-Af4 mediated leukemo-

genesis. Deletion of Igf2bp3 significantly increases the survival of mice with MLL-Af4 driven leukemia and greatly attenuates disease, with a minimal impact on baseline hematopoiesis. At the cellular level, MLL-Af4 leukemia-initiating cells require Igf2bp3 for their function in leukemogenesis. eCLIP and transcriptome analysis of MLL-Af4 transformed stem and progenitor cells and MLL-Af4 bulk leukemia cells reveals a complex IGF2BP3-regulated post-transcriptional operon governing leukemia cell survival and proliferation. Regulated mRNA targets include important leukemogenic genes such as those in the Hoxa locus and numerous genes within the Ras signaling pathway. Together, our findings show that IGF2BP3 is an essential positive regulator of MLL-AF4 mediated leukemogenesis and represents an attractive therapeutic target in this disease.

5.2 Introduction

Chromosomal rearrangements of the mixed-lineage leukemia (MLL, also known as KMT2A) gene are recurrently found in a subset of acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), and in acute leukemia of ambiguous lineage (Krivtsov and Armstrong, 2007). Despite recent advances in therapeutic approaches, patients with MLL-rearranged (MLLr) leukemia have very poor outcomes, a high risk of relapse, and show resistance to novel targeted therapies (Moorman et al., 2010; Pui et al., 2011) (Haddox et al., 2017; Rayes et al., 2016; Wei et al., 2008). MLL encodes a H3K4 methyltransferase that has been shown to be required for hematopoietic stem cell (HSC) development during both embryonic and adult hematopoiesis (Ernst et al., 2004a; Jude et al., 2007). Many of the translocation partners for MLL, including AF4 (AFF1), encode proteins that regulate transcriptional elongation (Lin et al., 2010; Mohan et al., 2010; Smith et al., 2011). Of more than 90 translocation fusion partner genes, MLL-AF4 (KMT2A-AFF1) is the most common MLL fusion protein in patients (Meyer et al., 2018). Biologically, MLL-AF4-driven leukemia is a distinct entity, with a unique

gene expression profile showing significant overlap with stem cell programs (Krivtsov et al., 2008; Somervaille et al., 2009).

At the post-transcriptional level, emerging evidence suggests a role for microRNAs, RNA-binding proteins, and other novel mechanisms in regulating gene expression during leukemogenesis (Ennajdaoui et al., 2016; Jønson et al., 2014; Nguyen et al., 2014; Palanichamy et al., 2016; Park et al., 2015). We recently identified the oncofetal RNA binding protein (RBP) Insulin like growth factor 2 mRNA binding protein 3 (IGF2BP3) as an important regulator of gene expression in MLL-rearranged B-ALL (Palanichamy et al., 2016). IGF2BP3 is expressed during embryogenesis, lowly expressed in healthy adult tissues, and strongly re-expressed in cancer cells (Mueller et al., 2003; Mueller-Pillasch et al., 1999). Several studies have shown that elevated levels of IGF2BP3 expression are correlated with diminished patient survival in many cancer types and may be a marker of disease aggressiveness in B-ALL (Kobel et al., 2009; Lochhead et al., 2012; Schaeffer et al., 2010; Stoskus et al., 2011). Previously, we determined that overexpression of IGF2BP3 in the bone marrow (BM) of mice leads to a pathologic expansion of hematopoietic stem and progenitor cells (HSPC), in a manner dependent on RNA binding. IGF2BP3 interacts primarily with the 3'UTR of its target transcripts, as with MYC and CDK6, resulting in an upregulation of transcript and protein (Palanichamy et al., 2016). In the case of MYC and CDK6, IGF2BP3 binding resulted in upregulation of target transcript and protein, with attendant effects on pathologic hematopoietic stem and progenitor cell expansion. Together, these studies illuminated a novel role for post-transcriptional gene regulation in the pathologic proliferation of HSPCs.

Experimentally, MLL-AF4 driven leukemogenesis has been studied using a range of *in vitro* and *in vivo* models leading to significant progress in our understanding of

MLL-rearranged leukemia (Bursen et al., 2010; Chen et al., 2006; Krivtsov et al., 2008; Metzler et al., 2006; Montes et al., 2011; Tamai et al., 2011). Here, we explicitly tested the requirement for Igf2bp3 in a bona-fide *in vivo* model of MLL-Af4 driven leukemogenesis (Lin et al., 2016). Deletion of Igf2bp3 significantly increased the survival of MLL-Af4 transplanted mice and decreased the numbers and self-renewal capacity of MLL-Af4 leukemia-initiating cells (LICs). Mechanistically, we found that IGF2BP3 targets and modulates the expression of transcripts within the Hoxa locus and components of the Ras signaling pathway, both key regulators of leukemogenesis, through multiple post-transcriptional mechanisms (Downward, 2003; Milne et al., 2010). Together, our findings show that IGF2BP3 is a critical regulator of MLL-AF4 mediated leukemogenesis and a potential therapeutic target in this disease (Downward, 2003; Milne et al., 2010).

5.3 Results

5.3.1 The MLL-AF4 fusion protein transcriptionally induces IGF2BP3

To determine the functional impact of IGF2BP3 expression on MLL-AF4-mediated gene expression, we compared IGF2BP3-regulated targets with a published dataset of MLL-Af4 targets obtained by ChIP-Seq (Lin et al., 2016; Palanichamy et al., 2016). Transcripts modulated by IGF2BP3 were significantly enriched for MLL-Af4-bound genes (Figure 5.1A; Supplemental Figure 1A). Interestingly, IGF2BP3 itself was a direct transcriptional target of MLL-Af4, with binding sites within the first intron and promoter region of IGF2BP3 (Lin et al., 2016). To confirm if IGF2BP3 was a direct transcriptional target of MLL-AF4, we performed ChIP-PCR assays on RS4;11 and SEM cell lines, human B-ALL cell lines that contain the MLL-AF4 translocation, and determined that a region in the first intron of IGF2BP3 is strongly bound by MLL-AF4 (Figure 5.1B; Supplemental Figure 1B) (Wilkinson et al., 2013). This MLL-AF4 binding was abro-

gated when SEM cells were treated with the bromodomain inhibitor, iBET-151 (Supplemental Figure 1C) (Dawson et al., 2011). Furthermore, we observed an MLL-AF4-dose-dependent increase in luciferase reporter activity, using a 950bp promoter region upstream of the transcription start site (TSS) of IGF2BP3 (Figure 5.1C). To confirm that MLL-AF4 not only binds to the IGF2BP3 gene but also promotes its expression, we utilized a retroviral MSCV vector encoding the human MLL fused to the murine Af4 (MLL-Af4)(Lin et al., 2016). In the murine pre-B cell line, 70Z/3, and primary murine bone marrow cells, we found that MLL-Af4 transduction caused an approximately 64-fold upregulation of Igf2bp3 mRNA (Figure 5.1D-E). Concordantly, IGF2BP3 protein was upregulated in MLL-Af4 transduced primary bone marrow cells, after being undetectable in control cells (Figure 5.1F). Taken together, these findings demonstrate that MLL-Af4 drives the expression of Igf2bp3 *in vivo*.

5.3.2 Normal hematopoiesis is maintained in Igf2bp3 KO mice

To test the *in vivo* requirement for IGF2BP3 in leukemogenesis, we generated an Igf2bp3 KO (I3KO) mouse. We initially generated a floxed Igf2bp3 allele (f/f; Supplemental Figure 2A) using CRISPR/Cas9. In the course of mating the mice with the Vav1-Cre mouse strain, we serendipitously generated a germline knockout allele (del), which we isolated and further characterized (Supplemental Figure 2B). This generation of a germline knockout allele is consistent with prior reports for the Vav1-Cre mouse strain, which displays "leaky" Cre expression resulting in germline deletion (Croker et al., 2004; de Boer et al., 2003; Georgiades et al., 2002; Heffner et al., 2012; Joseph et al., 2013). Mendelian inheritance was confirmed, although the distribution of genotypes was marginally skewed (Table S1). Deletion of Igf2bp3 was confirmed at the DNA, RNA, and protein level (Supplemental Figure 2C-E). Thus, Igf2bp3del/del (I3KO) mice were used for the remainder of the study. Immunophenotyping of I3KO mice showed

no significant differences in the numbers of HSPCs in the BM compared to WT (Supplemental Figure 2F). Additionally, I3KO mice showed similar numbers of myeloid-lineage progenitors (including CMPs, GMPs, and MEPs)(Supplemental Figure 5.2G) and normal B-cell development as assessed by the Hardy scheme (Hardy and Hayakawa, 2001) (Supplemental Figure 2H) and normal numbers of mature B-lymphoid, T-lymphoid, and myeloid lineages in the BM and spleen (Supplemental Figure 2I-J). Hence, I3KO mice demonstrate preserved normal, steady-state adult hematopoiesis, although specific differences in other hematopoiesis conditions need further investigation.

5.3.3 Igf2bp3 deletion increases the latency of MLL-Af4 leukemia and survival of mice

After confirmation of preserved baseline hematopoiesis in I3KO mice, we next utilized bone marrow transplantation (BMT) assays to query MLL-Af4 mediated leukemogenesis (Supplemental Figure 3A). Retroviral MLL-Af4 transduction was equivalent between WT and I3KO donor BM, based on DNA copy number analysis (Supplemental Figure 3B) and Western blot analysis for MLL-Af4 (Supplemental Figure 3C). Following transplantation of the transduced HSPCs, we found that the loss of Igf2bp3 significantly increases both the leukemia-free and overall survival of MLL-Af4 mice (Figure 2A-B). The median survival of WT/MLL-Af4 mice was 103 days while I3KO/MLL-Af4 mice had a median survival of greater than 157 days. Complete blood counts of WT/MLL-Af4 mice showed a consistent increase in WBC and absolute myeloid counts over time, which was severely blunted in I3KO/MLL-Af4 mice (Figure 5.2C; Supplemental Figure 5.3D). On average, peripheral blood counts crossed the leukemic threshold much earlier in WT/MLL-Af4 mice compared to I3KO/MLL-Af4 mice (70 days versus 112 days) (Figure 5.2C). Concordantly, peripheral blood smears showed reduced circulating blasts in I3KO/MLL-Af4 mice versus WT/MLL-Af4 mice (Supplemental Figure 3E).

Together, these findings indicated that Igf2bp3 is required for efficient MLL-Af4-mediated leukemogenesis.

5.3.4 Igf2bp3 modulates disease severity in MLL-Af4-driven leukemia

The MLL-Af4 model utilized here causes a highly penetrant, aggressive form of leukemia in mice. To characterize the role of Igf2bp3 in disease severity, we performed detailed immunophenotypic and histopathologic analyses in MLL-Af4-transplanted mice in timed experiments. I3KO/MLL-Af4 transplanted mice showed a highly significant, approximately 4-fold reduction in spleen weights at 14 weeks post-transplant compared to WT/MLL-Af4 transplanted mice (Figure 5.2D). We observed near-total infiltration of the spleen and liver by leukemic cells, obliterating the normal tissue architecture in WT/MLL-Af4 mice, a finding that was much reduced in I3KO/MLL-Af4 mice (Figure 5.2E). In line with this, I3KO/MLL-Af4 transplanted mice showed a significant reduction in CD11b+ cells (Figure 5.2G; Supplemental Figure 3G), which were less proliferative (CD11b+Ki-67+), both in the spleen (approximately 30-fold) and in the BM (approximately 2.5-fold) at 14 weeks (Figure 5.2F; Supplemental Figure 3F). Thus, Igf2bp3 deletion significantly reduces the tumor burden and attenuates disease severity in MLL-Af4 transplanted mice.

5.3.5 Igf2bp3 is required for LIC function in vitro

Several studies highlight the importance of LICs in both human and mouse leukemia. In the MLL-Af4 model, these LICs show expression of CD11b and c-Kit (Lin et al., 2017; Somervaille and Cleary, 2006; Somervaille et al., 2009). Given our findings of delayed initiation and decreased disease severity, we characterized these LICs, and found that I3KO/MLL-Af4 transplanted mice showed a significant 10-fold decrease in the numbers of leukemia-initiating cells (CD11b+c-Kit+) in the spleen and a 5-fold

decrease in the BM at 14 weeks compared to WT/MLL-Af4 mice (Figure 5.3A-B). To further characterize the MLL-Af4 LIC and its dependence on IGF2BP3, we turned to endpoint colony formation assays. Utilizing immortalized HSPCs, denoted as Lin-, from WT/MLL-Af4 and I3KO/MLL-Af4 mice, we confirmed of equal transcript expression levels of MLL-Af4 and deletion of IGF2BP3 at the protein level (Figure 5.3C; Supplemental Figure 5A-B). The deletion of Igf2bp3 resulted in an approximately 2-fold reduction in total colony formation as well as a significant decrease in CFU-GM progenitors (Figure 3D). To confirm our findings regarding LICs, we utilized an orthogonal method to delete Igf2bp3 via CRISPR-Cas9. Briefly, Lin-cells from Cas9-GFP BL/6 mice were collected and transduced with the retroviral MSCV-MLL-Af4 vector. After selection, these MLL-Af4 Cas9-GFP Lin-cells were transduced with a retroviral vector with an mCherry fluorescence marker containing either a non-targeting (NT) sgRNA or a sgRNA targeted against Igf2bp3 (I3sg) (Figure 5.3E). Importantly, Igf2bp3 is deleted after transformation with MLL-Af4, a distinction from the previous method. After confirmation of Igf2bp3 deletion (Figure 5.3F-G), GFP+mCherry+ MLL-Af4 Lin-cells were utilized for colony-forming assays. We confirmed that CRISPR-Cas9 mediated deletion of Igf2bp3 led to a significant reduction in total colony numbers and decreases in the various colony morphologies (Figure 5.3H). The observed differences in overall colony forming capacity between the two systems are most likely a result of the different methodologies being used, but in both systems, IGF2P3 deficiency led to decreased colony formation. Thus, Igf2bp3 is required for MLL-Af4 LIC function *in vitro*.

5.3.6 Igf2bp3 is necessary for the function of MLL-Af4 leukemia-initiating cells in vivo

Since Igf2bp3 deletion causes a reduction in LICs and is required for the function of these LICs, we next wanted to determine if Igf2bp3 specifically affects the capability

of these LICs to initiate MLL-Af4 leukemia *in vivo*. First, to investigate baseline hematopoietic stem cell function in I3KO mice, we completed a competitive repopulation bone marrow transplantation experiment by transplanting lethally irradiated CD45.1 recipient mice with 50% of either WT or I3KO CD45.2 donor BM and 50% CD45.1 donor BM. We found no defect in engraftment over time in transplant recipients of I3KO BM (Supplemental Figure 4A). Moreover, we determined no differences in multilineage hematopoietic reconstitution ability of I3KO donor cells, as immature lineages in the BM and mature B-hematopoietic cells in the periphery were intact (Supplemental Figure 4B-H, respectively). Given that there were no baseline differences in reconstitution by normal HSPCs, we now sought to determine if Igf2bp3 impacted the number of effective LICs in secondary transplant assays. We isolated BM cells from WT/MLL-Af4 and I3KO/MLL-Af4 mice with equivalent disease burdens and transplanted equal numbers (106, 105, and 104) of leukemic BM cells into immunocompetent CD45.1 mice. At 4 weeks post-transplantation, mice that received 106 cells from I3KO/MLL-Af4 mice had significantly reduced donor CD45.2+ cell engraftment (Figure 5.4A). With 105 and 104 transplanted cells, we no longer observed measurable leukemic burden in recipient mice (Figure 4A). This suggests that the active cell frequency of LICs in I3KO/MLL-Af4 mice is lost between 106 and 105 cells (Figure 5.4A)(Brien et al., 2010). Moreover, WBC and splenic weights were significantly decreased in I3KO/MLL-Af4 leukemia transplanted mice (Figure 5.4B-D). Histologically, leukemic infiltration was absent in the spleen and liver of 105 transplanted I3KO/MLL-Af4 mice (Figure 5.4E). Thus, Igf2bp3 deletion greatly attenuated transplantability, in which only 17% of I3KO/MLL-Af4 recipients developed leukemia while 67% of WT/MLL-Af4 recipients developed leukemia at 4 weeks with 106 transplanted cells (Figure 5.4B). These data show that the deletion of Igf2bp3 results in the significant reduction of LICs and reconstitution of MLL-Af4 transplanted mice, suggesting that Igf2bp3 is necessary for the self-renewal capability of LICs *in vivo*.

5.3.7 IGF2BP3 supports oncogenic gene expression networks in LIC-enriched and bulk leukemia cells

To identify differentially expressed transcripts related to the I3KO phenotype, we sequenced RNA from WT/MLL-Af4 and I3KO/MLL-Af4 Lin- and CD11b+ bulk leukemia cells (Figure 5.5A-B, respectively). First, we confirmed expression of MLL and Igf2bp3 in these samples by RT-qPCR and WB (Figure 5.3C; Supplemental Figure 5A-E). Differential expression analysis by DEseq2 revealed hundreds of differentially expressed transcripts (Figure 5.5A-B; Tables S2 and S3) (Love et al., 2014). We observed 208 upregulated and 418 downregulated transcripts in the CD11b+ cells, and 189 upregulated and 172 downregulated transcripts in the Lin-cells. To identify over-represented pathways and gene ontology terms within IGF2BP3 differentially regulated transcripts, we used the Metascape analysis tool (Zhou et al., 2019) and observed a significant enrichment in transcripts associated with the KEGG term related to transcriptional misregulation in cancer in both the Lin- and CD11b+ bulk leukemia dataset (Figure 5.5C-D). Interestingly, there were also distinct oncogenic networks that were regulated in the two datasets, with regulation of discrete signaling pathways noted in the Lin-cells (PI3K/AKT) and in the CD11b+ cells (GTPase, MAPK pathway) (Figure 5.5C-D). This was also confirmed by an independent analysis of differentially expressed data using GSEA, where we noted that the Hallmark KRAS pathway was significantly enriched in the CD11b+ cells (Supplemental Figure 5G) and GO oxidative phosphorylation in the Lin-cells (Supplemental Figure 5H). We used RT-qPCR to validate the RNA-seq data from both Lin- and CD11b+ cells. In Lin-cells, we focused on differentially regulated genes with a known leukemogenic function including Csf2rb, Notch1, Cd69, and the Hoxa cluster of transcripts, including Hoxa9, Hoxa10, Hoxa7. We observed a significant decrease in the steady state mRNA levels for each of these transcripts in the I3KO/MLL-Af4 Lin-cells, confirming our RNA sequencing results (Figure 5.5E). In CD11b+ cells,

we selected transcripts known to play a role in Ras signaling, Ccnd1, Maf, Mafb Itga6, Klf4, and Akt3 (Brundage et al., 2014; Eychène et al., 2008; Lewis et al., 2019; Riverso et al., 2017; Takata et al., 2005; Wagle et al., 2018; Wu et al., 2019). As expected, these transcripts were decreased in I3KO/MLL-Af4 CD11b+ cells by RT-qPCR, confirming the high throughput RNA sequencing findings (Figure 5.5F). Furthermore, we determined that there was a significant decrease in Ras GTPase activity in the I3KO/MLL-Af4 CD11b+ cells compared to WT/MLL-Af4 CD11b+ bulk leukemia cells by ELISA assay (Figure 5G). Together, this data demonstrates that IGF2BP3 plays a major role in amplifying the expression of many cancer-related genes in Lin- and CD11b+ cells

.

5.3.8 eCLIP analysis reveals a putative role for IGF2BP3 in pre-mRNA splicing

To determine how IGF2BP3 modulates gene expression in MLL-Af4 leukemia, we performed eCLIP to identify IGF2BP3 bound transcripts in both Lin- and CD11b+ cells (Figure 5A-B; eCLIP target transcripts denoted in red). Sequencing libraries were prepared from a minimum of two biological replicates with two technical replicates (four immunoprecipitations per cell line) as well as size matched input (smInput) samples from each condition. After filtering out reads that overlap the smInput, reproducible peaks were identified by using CLIPper (Tables S2-3; FS1 column) (Lovci et al., 2013). A significant fraction of the differentially expressed mRNAs are bound by IGF2BP3 ($P< 2.2\times10-16$; Supplemental Figure 6A). Motif analysis revealed an enrichment of CA-rich elements as expected (Supplemental Figure 6B) (Schneider et al., 2019). Although the majority of peaks were present within introns, we observed cell type-specific differences in the locations of IGF2BP3 binding sites within exons. In CD11b+ cells, a greater proportion of exonic peaks were found in 3'UTRs whereas a greater proportion of peaks mapped to internal exons in Lin-cells (Figure 5.6A). The eCLIP data revealed numerous peaks within precursor mRNA (pre-mRNA) in both Lin- and

CD11b+ cells, suggesting a potential role in splicing regulation. To characterize this observation, we utilized MISO (Mixture of Isoforms) analysis to identify differentially spliced transcripts in WT/MLL-Af4 and I3KO/MLL-Af4 cells (Katz et al., 2010). Across both cell lines, we identified hundreds of transcripts with IGF2BP3-dependent changes in alternative splicing, including 97 differential splicing events in Lin- and 261 splicing events in CD11b+ cells (Bayes factor ≥ 10, delta PSI ≥ 0.1, and minimum 20 reads supporting the event) (Supplemental Figure 6C). After merging all replicate eCLIP data for each cell type, we determined the position of eCLIP peaks relative to splice sites for splicing events identified by MISO (Figure 5.6B). Most event types, including skipped exons (SE), alternative first exons (AFE), alternative last exons (ALE), alternative 3' splice sites (A3SS), and alternative 5' splice sites (A5SS) exhibited both increases and decreases in percent spliced in (PSI), however, intron retention (RI) events showed a consistent reduction in splicing in the I3KO/MLL-Af4 cells (Figure 5.6C). A significant fraction of alternatively spliced transcripts contained IGF2BP3 binding sites in proximity of the splicing event ($P<2.2\times10-16$, Supplementary Figure 6D). Across all event types we found that the density of IGF2BP3 binding sites was strongest near the 3'splice site (3'ss), with additional signal near the 5' splice site (5'ss). This pattern was observed for each distinct splicing event class that MISO identified, with retained introns exhibiting the strongest bias towards the 3'ss (Supplemental Figure 6E). This positional bias in the data was noted for some differentially expressed genes, such as Hoxa9, Hoxa7, and Cd69 (Figure 5.6D). Hoxa9 is known to be alternatively spliced, with two well-characterized transcripts, full-length Hoxa9 and truncated (homeobox-less) Hoxa9T (He et al., 2012; Stadler et al., 2014). We designed and validated RT-qPCR primers to measure the two RNA isoforms, finding that I3KO/MLL-Af4 cells showed an alteration in the ratio of the two isoforms (Figure 5.6F). This was in addition to our previous finding that the total transcript (using primers that are isoform-agnostic) was also downregulated

in I3KO/MLL-Af4 cells (Figure 5.5E). Hence, the net effect of IGF2BP3 may be multi-pronged—there is a strong impact on steady state mRNA levels and potentially an impact on splicing. Taken together, these data demonstrate that IGF2BP3 functions in regulation of alternative pre-mRNA splicing in bulk leukemia cells and progenitor cells.

5.4 Discussion

Here, we have generated an Igf2bp3–deficient murine model and queried MLL-Af4 mediated leukemogenesis. We demonstrated that Igf2bp3 is required for the efficient initiation of leukemia and that this regulates the number and frequency of MLL-Af4 LICs. Igf2bp3 regulates the expression of numerous critical transcripts in the Hoxa locus and the Ras signaling pathway, leading to dysregulated gene expression and enhanced downstream signaling, thereby promoting leukemogenesis.

MLL-AF4 driven leukemogenesis is associated with massive transcriptional dysregulation, mediated by the fusion of a histone methyltransferase with a factor involved in the super elongation complex (Smith et al., 2011). We confirm here that Igf2bp3 is a direct transcriptional target of the MLL-AF4 fusion protein. Interestingly, IGF2BP3 itself seems to positively regulate MLL-AF4 transcriptional targets, based on our analysis provided here. Together, these data suggest that IGF2BP3 forms a novel post-transcriptional feed-forward loop that stabilizes and/or enhances the expression of MLL-Af4 transcriptional targets. Because of this unique relationship, and its relatively restricted pattern of expression in MLL-translocated leukemia, it is not clear if IGF2BP3 may play a role in other leukemia subtypes. However, it is worth noting that IGF2BP3 overexpression is noted in a wide range of cancer types and, hence, additional work is required to establish its role in other types of hematologic and non-hematologic malignancy.

In our previous study, we determined that IGF2BP3 is required for B-ALL cell survival, and that overexpression of IGF2BP3 in the bone marrow of mice leads to a pathologic expansion of hematopoietic stem and progenitor cells (Palanichamy et al., 2016). Here, using the MLL-Af4 leukemia model, we found that the deletion of Igf2bp3 caused a striking delay in leukemia development and significantly increased the survival of MLL-Af4 mice. Furthermore, Igf2bp3 deficiency greatly attenuated the aggressiveness of the disease. This was demonstrated by significant decreases in WBC counts, spleen weights, and infiltrating leukemic cells visualized in histopathological analysis of hematopoietic tissues. Although MLL-Af4 drives an acute myeloid leukemia in mice (Lin et al., 2016), it is important to note that MLLr leukemias often show lineage infidelity and plasticity, leading to difficulties in applying targeted therapy (Rayes et al., 2016). While our prior work focused on IGF2BP3 in B-lineage MLLr leukemia, the current work suggests its broader applicability to all MLLr leukemia. Hence, IGF2BP3 may be a constant factor to target across the phenotypic range of MLLr leukemia and may be less subject to change in response to targeted therapy.

We also determined that Igf2bp3 regulates the numbers and function of leukemia-initiating cells (LICs). Importantly, the effect of Igf2bp3 deletion was restricted to LICs and did not significantly impact normal HSC function. Deletion of Igf2bp3 led to a LIC disadvantage *in vivo* and *in vitro*, using both the I3KO mouse and a novel, orthogonal system utilizing CRISPR/Cas9-mediated deletion of Igf2bp3. LICs have been defined as cells that can self-renew and have the capability to produce downstream bulk leukemia cells (Magee et al., 2012). The persistence of these LICs is thought to contribute to relapse after treatment in several different leukemia subtypes (Bao et al., 2006; Chu et al., 2011; Diehn et al., 2009; Merlos-Suárez et al., 2011). In MLLr leukemia, LICs have been shown to have a high frequency in tumors and co-expression of mature

lineage-restricted cell markers, with some excellent work in mouse models (Krivtsov et al., 2006; Somervaille and Cleary, 2006; Somervaille et al., 2009). However, the details of human LICs, particularly in MLL-AF4 leukemia, are less well known (Agraz-Doblas et al., 2019; Bardini et al., 2015; Barrett et al., 2016; Metzler et al., 2006). The role of IGF2BP3 in such cells will be of great interest and is a future direction for our work.

Previously, we discovered that IGF2BP3 interacts primarily with the 3'UTR of its target transcripts via iCLIP-seq (Palanichamy et al., 2016). In this study, we determined that IGF2BP3 targets many transcripts within intronic regions and near splice sites in addition to the 3'UTR, suggesting additional roles in post-transcriptional gene regulation. This difference may be due to the use of the eCLIP technique or the focused application on primary cells as opposed to cell lines. It is not entirely surprising, however, since RBPs are known to regulate gene expression at several steps at the post-transcriptional level through mRNA operons (Keene, 2007; Keene and Lager, 2005). Furthermore, a recent study has shown that IGF2BP3 may regulate alternative splicing in the PKM gene in lung cancer cells (Xueqing et al., 2020). In line with this study, we found dynamic alternative splicing events that reflected various categories of alternative splicing phenomena, including retained Introns, alternative 5' and 3' splice sites, and skipped exons. In this light, it is interesting to note that intron retention has recently been reported to be a mechanism of transcriptome diversification in cancer and, specifically, in leukemia (Dvinge and Bradley, 2015; Wang et al., 2019). With this unexpected, novel discovery and our prior work detailing interactions with the 3'UTR, it is very likely that IGF2BP3 regulates specific mRNA operons through multiple post-transcriptional mechanisms in MLL-Af4 driven leukemia.

As an RBP, the function of IGF2BP3 is intimately connected to the underlying transcriptional program—IGF2BP3 can only act on transcripts that are specifically induced in the cell type where it is expressed. Hence, the finding of unique sets of genes that are bound and regulated by IGF2BP3 in Lin- and CD11b+ cells is not entirely unexpected, given that transcription changes as the leukemia initiating cells differentiate into the bulk leukemic cells. This is similar to what has been observed for miRNAs, which post-transcriptionally regulate distinct gene expression programs in distinct cell types (Lechman et al., 2016). The significant enrichment of IGF2BP3-bound mRNAs in the sets of differentially regulated and differentially spliced transcripts confirms a direct regulatory effect. However, further work is required to confirm functional relationships between the specific transcripts that are regulated and the phenotypic effects driven by IGF2BP3.

Notably, these differentially regulated transcripts showed significant enrichment for the KEGG transcriptional misregulation in cancer term as well as the GO oxidative phosphorylation term. Notable IGF2BP3 targets included critical transcripts in the Hoxa cluster such as Hoxa9, Hoxa10, and Hoxa7. HOXA9 is induced by MLL-AF4, plays a role in normal hematopoiesis, and is required for the survival of MLL-rearranged leukemia (Ernst et al., 2004b; Faber et al., 2009; Imamura et al., 2002; Lawrence et al., 2005; Pineault et al., 2002; Rozovskaia et al., 2001). Furthermore, Hoxa9 is an alternatively spliced gene, with co-expression of a homeodomain-less splice variant, Hoxa9T, together with Hoxa9, shown to be necessary for full leukemogenic transformation (He et al., 2012; Stadler et al., 2014). Hence, Igf2bp3 may act through upregulation of Hoxa9 and Hoxa9T through multiple post-transcriptional mechanisms to promote MLL-Af4 driven leukemogenesis and impact the function of MLL-Af4 LICs. In addition, HOXA9 may play a role in the regulation of oxidative phosphorylation (Lynch et al., 2019), and

it is tempting to speculate that the observed Igf2bp3-dependent impact on LICs is a consequence of dysregulated oxidative phosphorylation, a key pathway that regulates LICs. Importantly, because Igf2bp3 was not required for steady-state hematopoiesis, in contrast to HOXA9, it may represent a more attractive target.

Work from our lab and others have demonstrated that IGF2BP3 targets a wide array of oncogenic transcripts and pathways, including CDK6 and MYC (Palanichamy et al., 2016). Here, we found that IGF2BP3 targets and modulates the expression of many transcripts within the Ras signaling pathway and its downstream effector pathways. RAS proteins control numerous cellular processes such as proliferation and survival, and are amongst the most commonly mutated genes in cancer (Downward, 2003; Schubbert et al., 2007). Interestingly, while MLLr leukemia has a paucity of additional mutations, the mutations that are present are in found mainly in the RAS signaling pathway (Agraz-Doblas et al., 2019; Andersson et al., 2015; Chandra et al., 2010; Emerenciano et al., 2015; Hyrenius-Wittsten et al., 2018; Kerstjens et al., 2017; Lavallée et al., 2015; Trentin et al., 2016). Moreover, several studies have shown selective activity against MLL-r leukemia cell lines and primary samples *in vitro* by MEK inhibitors, suggesting an important role for signaling downstream of RAS mutations in leukemia cell survival and proliferation (Kampen et al., 2014; Kerstjens et al., 2017; Lavallée et al., 2015).

Here, we determined that Igf2bp3 is required for the efficient initiation of MLL-Af4 driven leukemia as well as for the development of and self-renewal capability of MLL-Af4 LICs. Mechanistically, IGF2BP3 binds to hundreds of transcripts and modulates their expression *in vivo* and *in vitro* through multiple, complex post-transcriptional mechanisms. We describe a novel positional bias for IGF2BP3 binding in leukemic cells isolated from an *in vivo* model, a notable advance. In summary, our study demonstrated

that IGF2BP3 is an amplifier of MLLr leukemogenesis by targeting Hoxa transcripts essential for leukemia-initiating cell function and targeting Ras signaling pathway transcripts, thereby controlling multiple critical downstream effector pathways required for disease initiation and severity. Our findings highlight IGF2BP3 as a necessary regulator of MLLr leukemia and a potential therapeutic target for this disease.

5.5 Materials & Methods

5.5.1 ChIP-PCR

RS4;11 and SEM cells were used for ChIP assays as previously described (Janardhan et al., 2017). Primer sequences for the IGF2BP3 promoter region were provided by James Mulloy (University of Cincinnati College of Medicine)(Lin et al., 2016).

5.5.2 Western Blotting and RT-Qpcr

Protein and mRNA extracts were prepared, and Western Blot/RT-qPCR performed as previously described (Fernando et al., 2017). Primers for qPCR and antibodies used for Western blotting are listed in Table S4.

5.5.3 Plasmids

The MSCV-MLL-flag-Af4 plasmid was kindly provided by Michael Thirman (University of Chicago, Department of Medicine) through MTA (Lin et al., 2016). The non-targeting or Igf2bp3 sgRNA was cloned into an in-house MSCV-hU6-sgRNA-EFS-mCherry vector.

5.5.4 Retroviral transduction and bone marrow transplantation

Retroviral transduction and bone marrow transplantation (BMT) were completed as previously described (Fernando et al., 2017; O'Connell et al., 2010; Rao et al., 2010). 5-FU enriched BM and Lin-cells were spin-infected four times with MSCV-MLL-flag-Af4 virus at 30°C for 45 minutes in the presence of polybrene. Cells were selected with

400 μg/ml of G418 for 7 days. For sgRNA-mediated knockout, MLL-Af4 overexpressing Cas9-GFP cells were retrovirally infected with MSCV-hU6-NT/I3sgRNA-EFS-mCherry. The selected cells were then cultured for colony formation assays or injected into lethally irradiated mice.

5.5.5 Mice

C57BL/6J and B6J.129(Cg)-Gt(ROSA)26Sortm1.1(CAG-cas9*,-EGFP)Fezh/J (Cas9-GFP) mice were obtained from Jackson Laboratory. For Igf2bp3 KO mouse generation, the UCI Transgenic Mouse Facility utilized CRISPR/Cas9 to insert loxP sites flanking exon 2 of Igf2bp3 to generate Igf2bp3f/f mice. We originally attempted to generate conditional KO mice by breeding the Igf2bp3f/f mice with Vav1-Cre mice. Consistent with prior reports, we found that this strategy led to "leaky" Cre expression, resulting in germline deletion (Croker et al., 2004; de Boer et al., 2003; Georgiades et al., 2002; Heffner et al., 2012; Joseph et al., 2013). To isolate the floxed and deletion (del) alleles, we back-crossed the mice onto C57BL/6 mice, successfully confirming germline, Mendelian transmission of the del and floxed alleles in two successive generations (Table S1). Mice heterozygous for the del allele were mated together, with the production of a homozygous deletion of Igf2bp3, resulting in the Igf2bp3del/del mice (I3KO) used in this study.

5.5.6 Cell culture

RS4;11, SEM, 70Z/3 and HEK 293T cell lines were cultured as previously described (Fernando et al., 2017). Lin-cells were cultured in IMDM with 15% fetal bovine serum supplemented with SCF, IL-6, FLT3, and TPO. CD11b+ cells were isolated from splenic tumors for positive selection by CD11b antibody and MACS (Miltenyi).

5.5.7 Flow cytometry

Blood, BM, thymus, and spleen were collected from the mice under sterile conditions at the indicated time points and staining performed as previously described (Contreras et al., 2015; Fernando et al., 2017). The list of antibodies used is provided in Table S4. Flow cytometry was performed on a BD FACS LSRII. Analysis was performed using FlowJo software.

5.5.8 Histopathology

Fixation and sectioning has been described previously (O'Connell et al., 2010). Analysis was performed by a board certified hematopathologist (D.S. Rao).

5.5.9 Competitive repopulation assay and secondary leukemia transplantation

Competitive repopulation experiments were completed as previously described (Palanichamy et al., 2016). For leukemia transplantation, BM was collected from WT/MLL-Af4 or I3KO/MLL-Af4 mice that succumbed to leukemia at 10-14 weeks post-transplantation and injected into 8-week-old immunocompetent CD45.1+ female mice.

5.5.10 eCLIP

IGF2BP3 crosslinking-immunoprecipitation studies were carried out from a minimum of two biological replicates with two technical replicates (four immunoprecipitations per cell type) and size matched input (smInput) samples in each cell type using Eclipse BioInnovations eCLIP kit. Briefly, 5×105 cells were crosslinked with 245nm UV radiation at 400mJoules/cm2. Crosslinked cell lysates were treated with RNAse I to fragment RNA and immunoprecipitated with anti-IGF2BP3 antibody (MBL RN009P) coupled to magnetic Protein G beads. Paired-end RNA sequencing was performed on the Illumina HiSeq4000 system at the UCSF Genomics Core Facility. Peaks were

called using CLIPper (Lovci et al., 2013). Peaks were filtered based on appearance in the smInput (FS1). Annotation of the genomic location of the peaks and motif enrichment analysis were performed using HOMER (Heinz et al., 2010) annotatePeaks.pl and findMotifs.pl, respectively. Background for the peaks within differentially expressed genes was simulated using bedtools (Heinz et al., 2010; Quinlan and Hall, 2010) and shuffled 1000 times.

5.5.11 RNA seq

Single-end, strand-specific RNA sequencing was performed on the Illumina HiSeq3000 system for the Lin- and CD11b+ samples, resulting in 15-20 million reads per sample, at the UCLA Technology Center for Genomics & Bioinformatics. Our analysis pipeline has been previously described (Palanichamy et al., 2016). Enrichment analysis for KEGG pathways and Gene Ontology (GO) biological processes terms was completed with the Metascape analysis tool (http://metascape.org)(Zhou et al., 2019). Gene Set Enrichment Analysis (GSEA) was completed using the GSEAPreranked software on both the Lin- and CD11b+ DESeq datasets after calculation of π-value (Mootha et al., 2003; Subramanian et al., 2005; Xiao et al., 2014) to compare to the Hallmark and GO gene sets within the Molecular Signatures Database.

5.5.12 RNA seq data analysis

The RNA seq reads were mapped to the mouse genome assembly mm10 using STAR version 2.7.1a (Dobin et al., 2013). Repeat sequences were masked using Bowtie 2 (Langmead and Salzberg, 2012) and RepeatMasker elements (Tarailo-Graovac and Chen, 2009). Differentially expressed genes were identified using DESeq2 (Love et al., 2014) on the CD11b dataset and fdrtool (Strimmer, 2008a; Strimmer, 2008b) on the Lin-dataset. Multiple testing correction was done using the Benjamini-Hochberg

method. Differentially expressed genes were considered significant if adjusted p-value < 0.1 and log2FC > 1. All data collection and parsing were completed with bash and python2.7. Statistical analyses were performed using R programming language version 3.5.1.

5.5.13 Estimation of alternative splicing

Mixture of Isoforms (MISO) Bayesian Inference model v0.5.4 (Katz et al., 2010) was used to quantify alternative splicing events. The MISO event database for pairwise alternative splicing events for mm10 ("exon-centric annotation") was downloaded from hollywood.mit.edu/burgelab/miso/annotations/. After MISO quantified the percent spliced in (PSI) for each event by counting the number of reads supporting both events and the reads that are unique to each isoform, we calculated delta PSI by subtracting PSI from the WT with the I3KO sample for each alternative event. Finally, we filtered for significant and differential splicing events between wild-type and knockdown samples by requiring that delta PSI > 0.1, the Bayes factor ≥10, and the sum of exclusion and inclusion reads ≥ 10.

5.5.14 Statistics

Data represent mean ±SD for continuous numerical data, unless otherwise noted in the figure legends. One-way ANOVA followed by Bonferroni's multiple comparisons test or 2-tailed Student's t tests were performed using GraphPad Prism software. One-way ANOVA followed by Bonferroni's multiple comparisons test was performed in experiments with more than two groups

Dissertation author contribution:

J.P: performed meta analysis of eCLIP binding sites within miRNAs and performed differential expression analysis on RNAseq data.

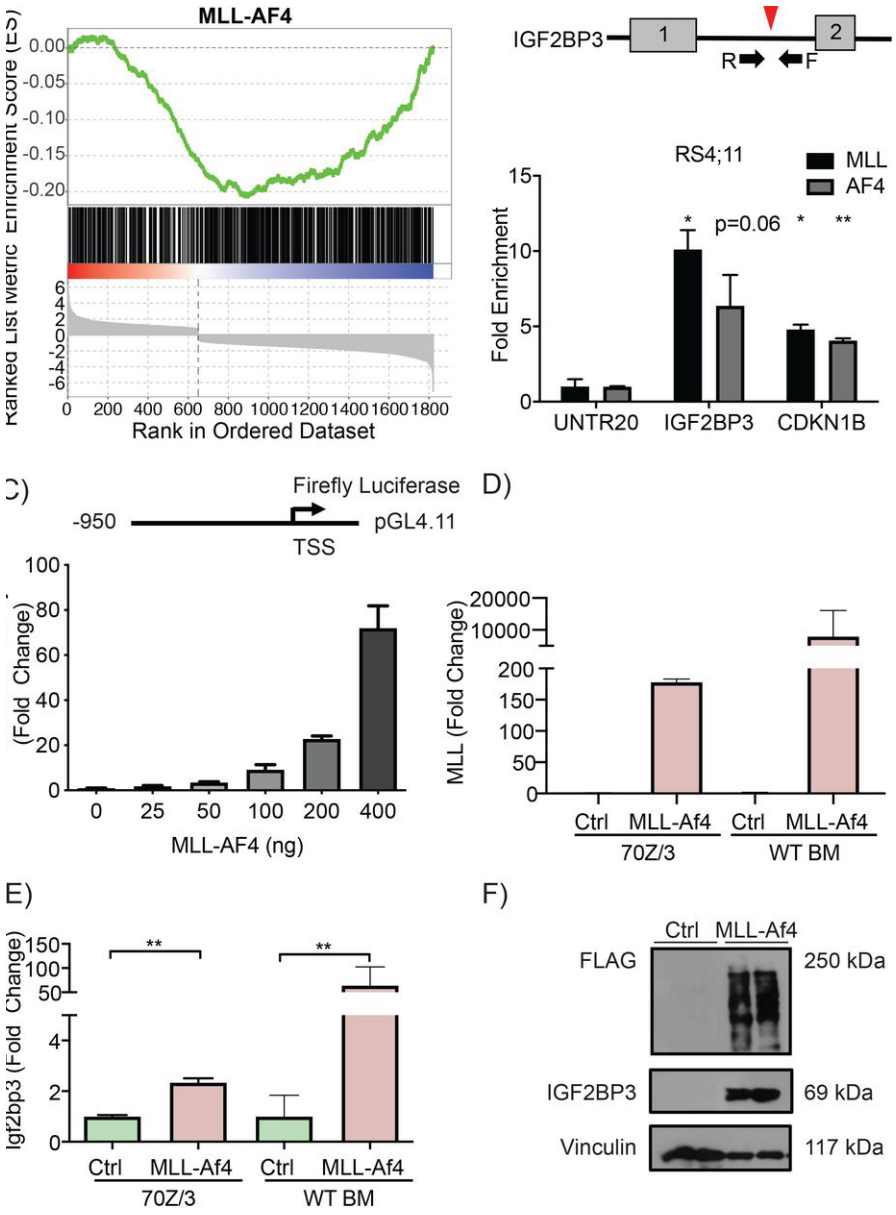

Figure 5.1 MLL-AF4 transcriptionally induces IGF2BP3.
A) GSEA of differentially expressed genes from IGF2BP3 depleted RS4;11 cells shows significant negative enrichment with MLL-AF4 ChIP targets (nominal P value: 0.001, FDR: 0.001, Normalized ES: -1.54)). B) Schematic of MLL-AF4 binding site in intron 1 of IGF2BP3 (top). ChIP-qPCR shows fold enrichment for IGF2BP3 and CDKN1B with MLL and AF4 IP in RS4;11. Normalized to UNTR20, an untranscribed region (t-test; *P < 0.05, **P < 0.01). C) Luciferase assay of the IGF2BP3 promoter shows a dose-dependent response to MLL-AF4. D) Expression of MLL through RT-qPCR of 70Z/3 transduced with either control (Ctrl) or MLL-Af4 vector selected with G418 and MLL expression at the RNA level in the BM of WT recipients transplanted with Ctrl or MLL-Af4 HSPCs. E) Induction of Igf2bp3 at the RNA level in selected 70Z/3 with MLL-Af4 and in the BM of WT recipients transplanted with Ctrl or MLL-Af4 HSPCs (bottom) (t-test; **P < 0.01). F) Induction of Igf2bp3 at the protein level in BM from mice transplanted with MLL-Af4 transduced WT donor HSPCs.

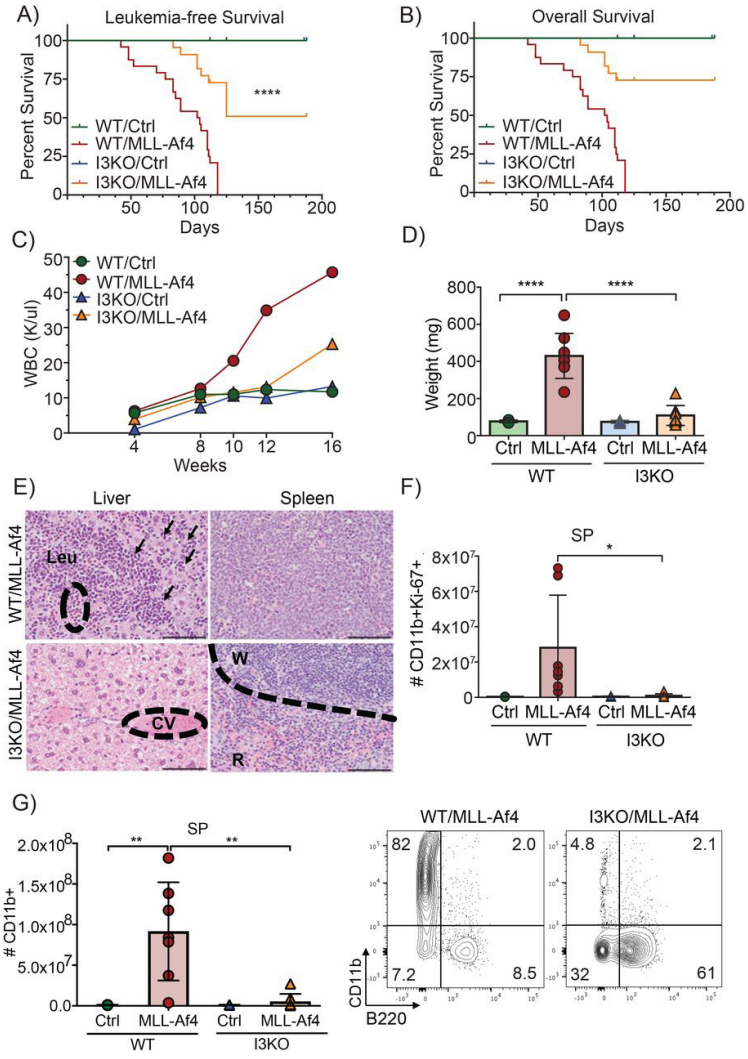

Figure 5.2 Igf2bp3 deletion delays leukemogenesis and reduces disease severity.
A) Leukemia-free survival of mice transplanted with control (Ctrl) or MLL-Af4 transduced HSPCs from WT or Igf2bp3 KO mice (Kaplan-Meier method with Log-rank test; ****P < 0.0001).B) Overall survival of mice transplanted with Ctrl or MLL-Af4 transduced HSPCs from WT or I3KO mice (n=12 WT/Ctrl, n=24 WT/MLL-Af4, n=7 I3KO/Ctrl, n=22 I3KO/MLL-Af4; Kaplan-Meier method with Log-rank test; ****P < 0.0001). C) Time course of WBC in the PB of mice transplanted with Ctrl or MLL-Af4 transduced HSPCs from WT or I3KO mice (Data represented as means of three experiments; n=4 Ctrl, n=8 MLL-Af4 per experiment). D) Spleen weights of mice transplanted with Ctrl or MLL-Af4 transduced HSPCs from WT or I3KO mice at 14 weeks (n= 4 Ctrl, n=8 MLL-Af4; one-way ANOVA followed by Bonferroni's multiple comparisons test; ****P < 0.0001). E) H&E staining of liver and spleen of mice transplanted with mice transplanted with MLL-Af4 transduced HSPCs from WT or I3KO mice at 14 weeks. Scale bar:100 microns; CV=Central vein; W=White pulp; R=Red pulp; Leu= Leukemia; arrows showing infiltration. F) Quantitation of CD11b+Ki67+ cells in the spleen at 14 weeks post-transplantation (n= 4 Ctrl, n=8 MLL-Af4; one-way ANOVA followed by Bonferroni's multiple comparisons test; *P < 0.05). G) (Left) Number of CD11b+ in the SP of recipient mice that received Ctrl or MLL-Af4 transduced HSPCs from WT or I3KO mice at 14 weeks (one-way ANOVA followed by Bonferroni's multiple comparisons test; **P < 0.01). (Right) Corresponding representative FACS plots showing CD11b+ and B220+ cells in the SP.

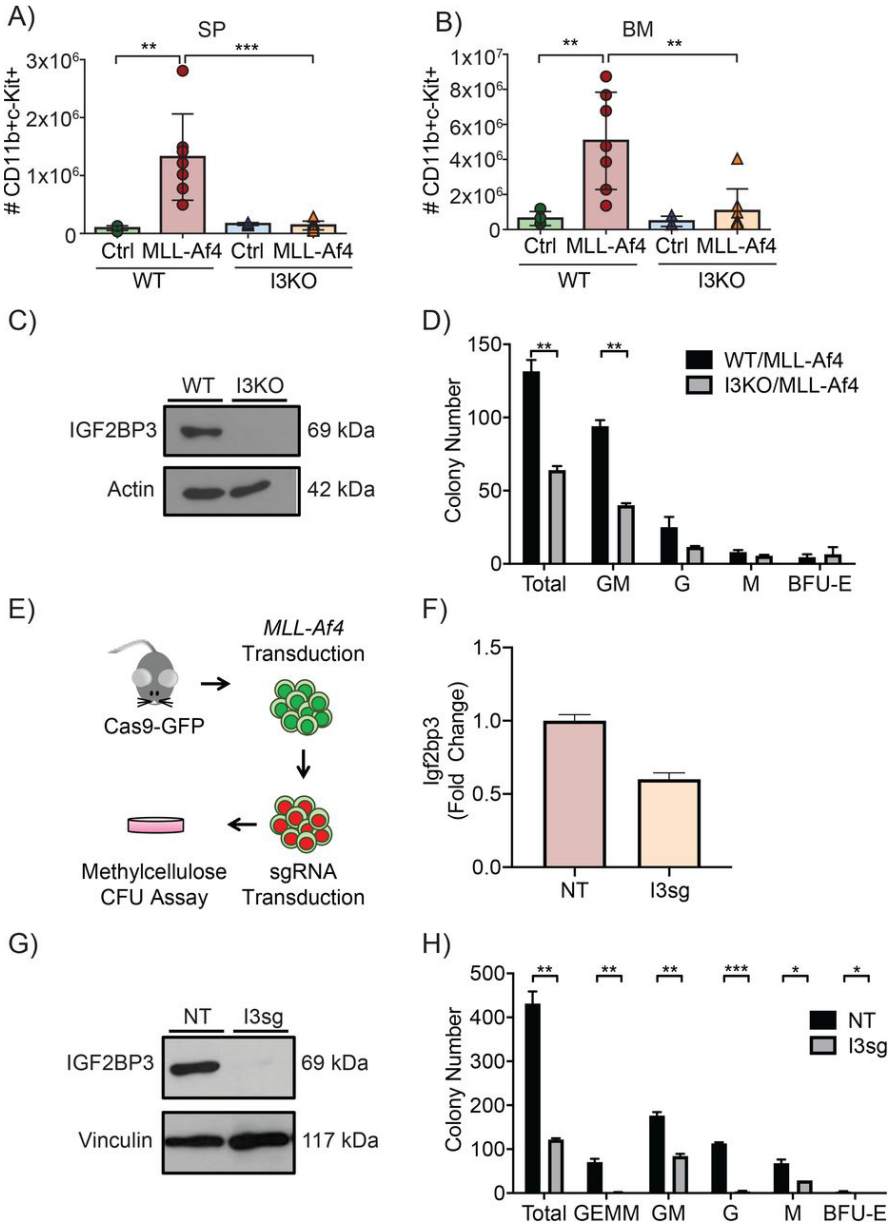

Figure 5.3 Igf2bp3 is required for LIC function in endpoint colony formation assays.
A) Quantification of CD11b+c-Kit+ cells in the spleen of recipient mice at 14 weeks post-transplantation (n= 4 Ctrl, n=8 MLL-Af4; one-way ANOVA followed by Bonferroni's multiple comparisons test; **P < 0.01). B) Quantitation of CD11b+c-Kit+ cells in the BM 14 weeks post-transplantation (n= 4 Ctrl, n=8 MLL-Af4; one-way ANOVA followed by Bonferroni's multiple comparisons test; **P < 0.01, ***P < 0.001). C) Expression of IGF2BP3 of in WT/MLL-Af4 and I3KO/MLL-Af4 immortalized Lin-cells at the protein level. D) Colony formation assay of WT/MLL-Af4 and I3KO/MLL-Af4 immortalized Lin-cells (t-test; **P < 0.01). E) Schematic of collection of Cas9-GFP MLL-Af4 Lin-cells and CRISPR-Cas9 mediated deletion of Igf2bp3. F) Expression of Igf2bp3 in Cas9-GFP MLL-Af4 Lin-cells in non-targeting (NT) and Igf2bp3 deleted (I3sg) cells by RT-qPCR. G) Expression of IGF2BP3 in NT and I3sg Cas9-GFP MLL-Af4 Lin-cells at the protein level. H) Colony formation assay of NT and I3sg deleted Cas9-GFP MLL-Af4 Lin-cells (t-test; *P < 0.05, **P < 0.01, ***P <0.001).

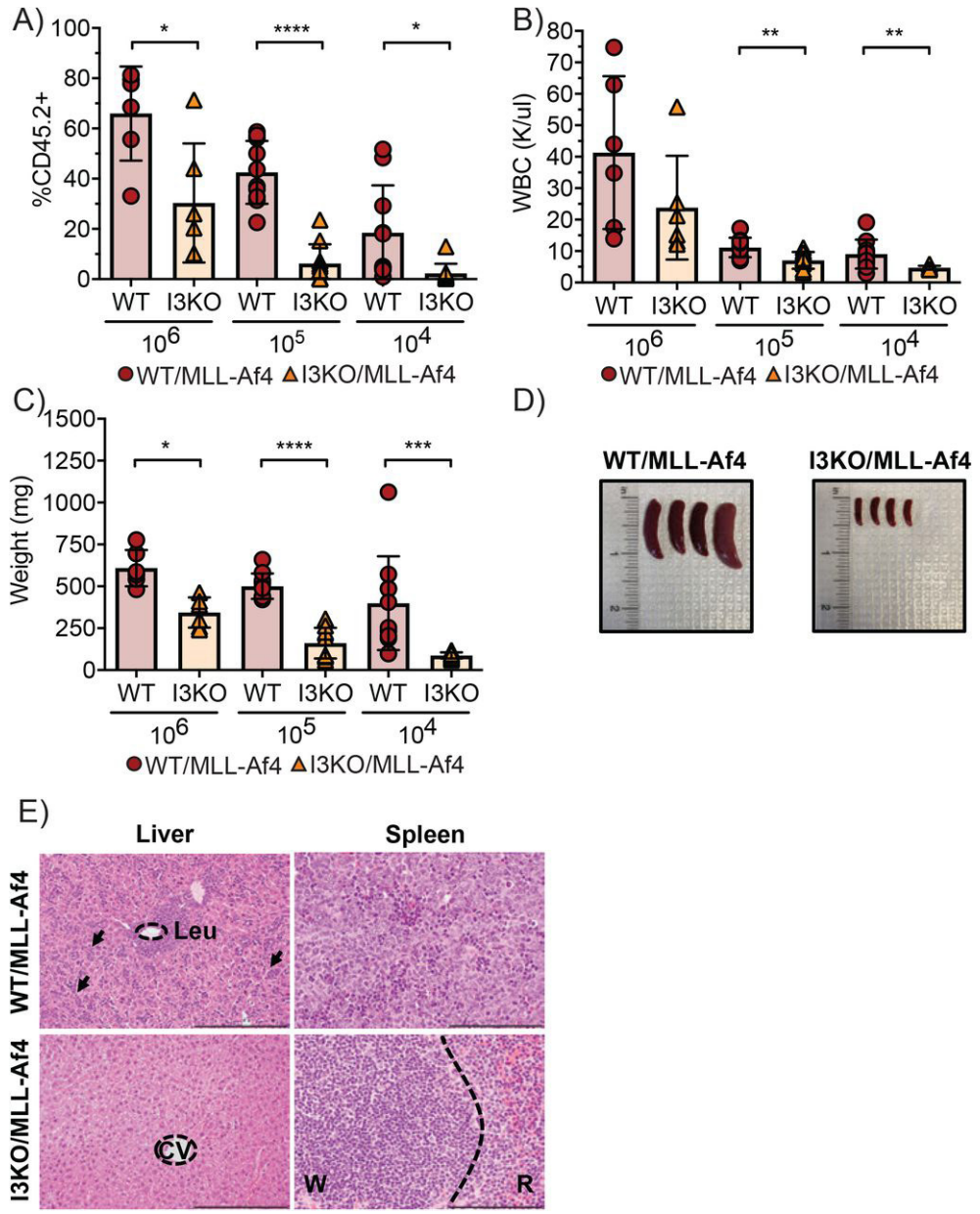

Figure 5.4 Igf2bp3 deletion is necessary for MLL-Af4 leukemia-initiating cells to reconstitute mice in vivo.

A) Percentage of CD45.2+ in the peripheral blood of secondary transplanted mice from leukemic WT/MLL-Af4 or I3KO/MLL-Af4 donor mice at 106, 105, and 104 BM cells at 4 weeks post-transplantation (n= 6 106, n=10 105, n=10 104; t-test; *P < 0.05, ***P < 0.001, ****P < 0.0001). B) WBC from PB of secondary transplanted mice from WT/MLL-Af4 or I3KO/MLL-Af4 BM 3-4 weeks post-transplant (n= 6 106, n=10 105, n=10 104; t-test; **P < 0.01). C) Splenic weights of secondary transplanted mice at 4-5 weeks (n= 6 106, n=10 105, n=10 104; t-test; *P < 0.05, ***P < 0.001, ****P < 0.0001). D) Images of splenic tumors in secondary mice transplanted with 10,000 BM cells from WT/MLL-Af4 mice (left) or I3KO/MLL-Af4 mice (right) at 5 weeks. E) H&E staining of liver and spleen of secondary transplant recipients that received 105 cells at 4 weeks. Scale bar: liver, 200 microns; spleen, 100 microns; CV=Central vein; W=White pulp; R=Red pulp; Leu= Leukemia; arrows showing infiltration.

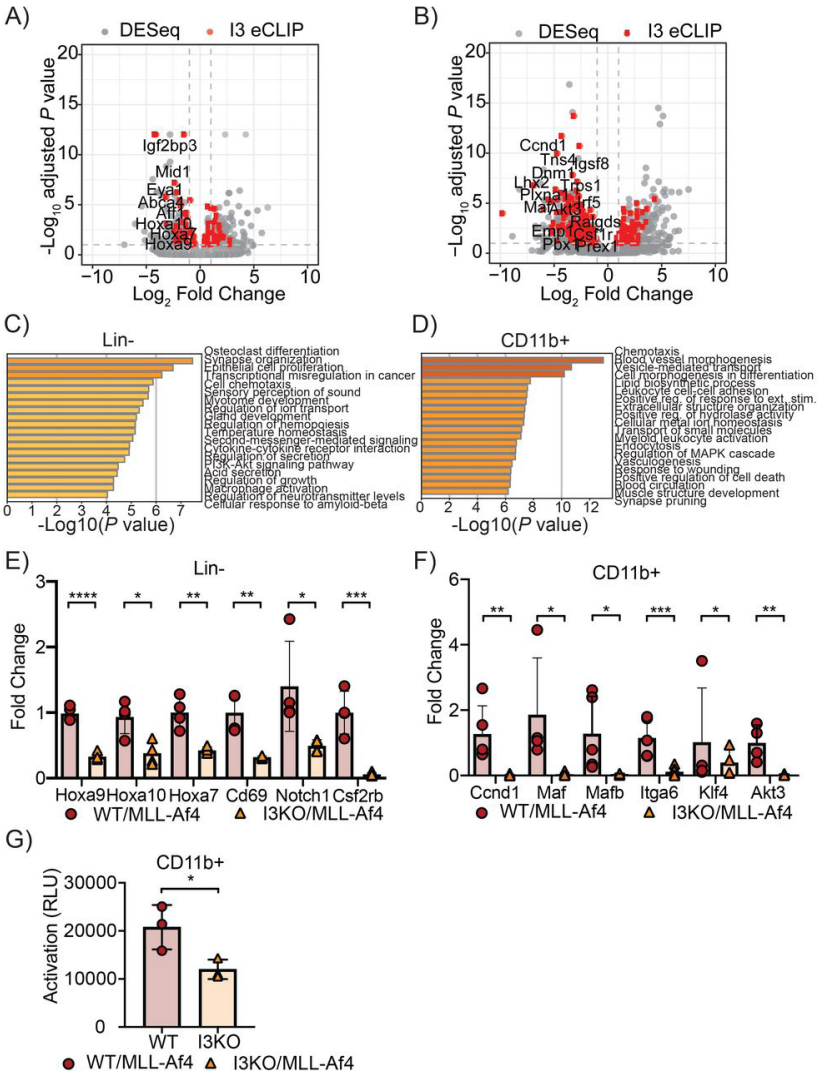

Figure 5.5 IGF2BP3 enhances MLL-Af4 mediated leukemogenesis through targeting transcripts within leukemogenic and Ras signaling pathways.

A) Volcano plot of differentially expressed genes determined using DESeq analysis on RNA-seq samples from WT/MLL-Af4 or I3KO/MLL-Af4 Lin-cells. Dotted lines represent 1.0-fold–change in expression (vertical lines) and adjusted P < 0.1 cutoff (horizontal line). IGF2BP3 eCLIP-seq targets are highlighted in red. B) Volcano plot of differentially expressed transcripts determined using DESeq analysis on RNA-seq samples from WT/MLL-Af4 or I3KO/MLL-Af4 CD11b+ cells. Dotted lines represent 1.0-fold–change in expression (vertical lines) and adjusted P < 0.1 cutoff (horizontal line). IGF2BP3 eCLIP-seq targets are highlighted in red. C) GO Biological Processes and KEGG Pathway enrichment determined utilizing the Metascape enrichment analysis webtool on MLL-Af4 Lin-IGF2BP3 DESeq dataset with an adjusted P < 0.05 cutoff. D) GO Biological Processes and KEGG Pathway enrichment determined utilizing the Metascape enrichment analysis webtool on MLL-Af4 CD11b+ IGF2BP3 DESeq dataset with an adjusted P < 0.05 cutoff. Bar graphs are ranked by P value and overlap of terms within gene list. E) Expression of leukemogenic target genes in WT/MLL-Af4 and I3KO/MLL-Af4 Lin-cells by RT-qPCR (n= 4; t-test; *P < 0.05, **P < 0.01, ****P < 0.0001). F) Expression of Ras signaling pathway genes in WT/MLL-Af4 and I3KO/MLL-Af4 CD11b+ cells by RT-qPCR (n=4; t-test; *P < 0.05, **P < 0.01, ***P < 0.001). G) Ras GTPase activity by ELISA in WT/MLL-Af4 and I3KO/MLL-Af4 CD11b+ cells (n=3; t-test; *P < 0.05).

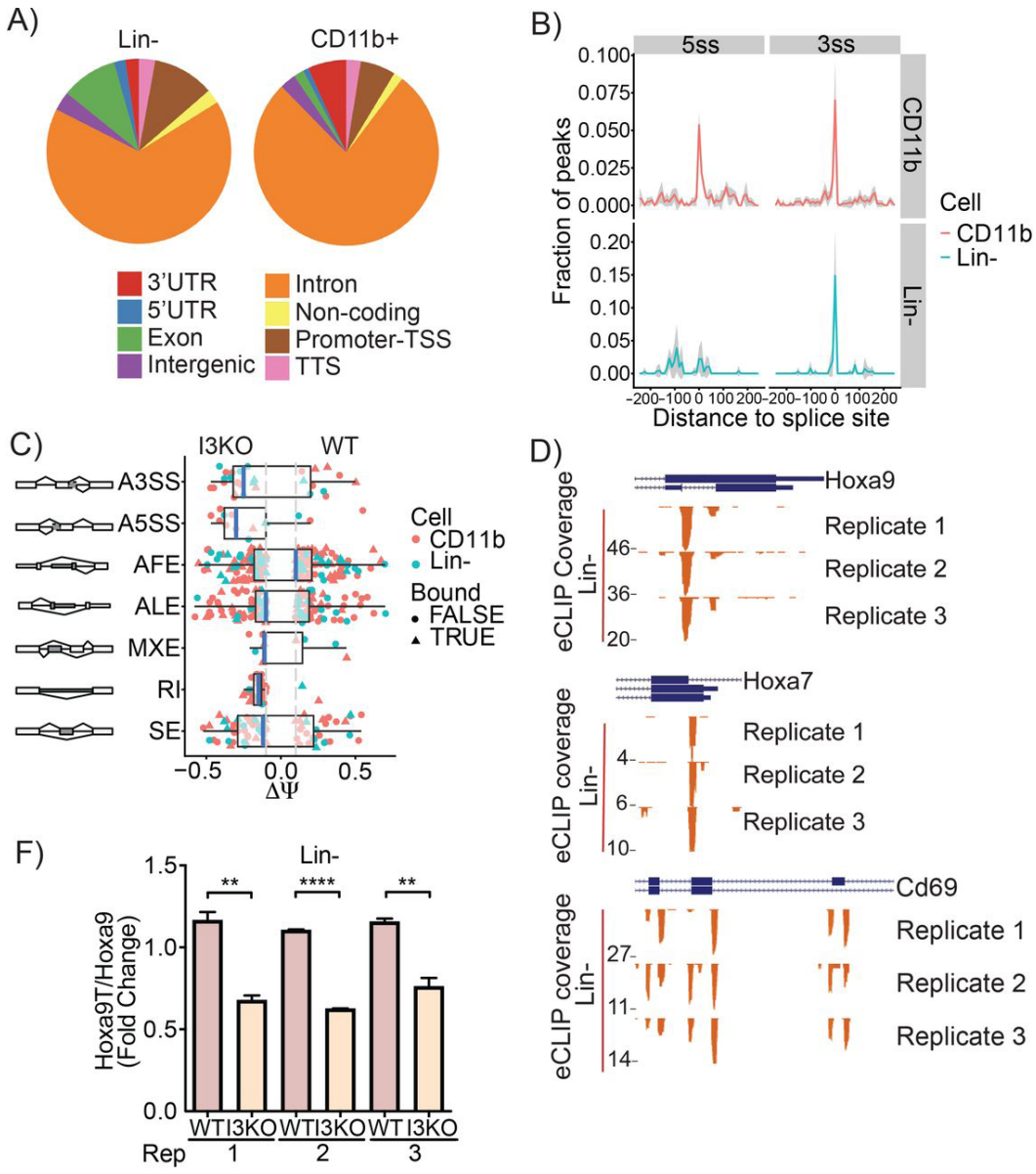

Figure 5.6 eCLIP analysis reveals IGF2BP3 function in regulating alternative pre-mRNA splicing. A) Genomic locations of IGF2BP3 eCLIP peaks in WT/MLL-Af4 Lin-cells and CD11b+ cells. B) Histogram showing normalized IGF2BP3 eCLIP peak counts and distance from IGF2BP3 eCLIP peak of 5' (5ss) and 3' (3ss) splice sites in WT/MLL-Af4 CD11b+ (top) cells and Lin-cells (bottom). C) Distribution of types of alternative splicing patterns for WT/MLL-Af4 or I3KO/MLL-Af4 Lin- and CD11b+ cells using MISO analysis. Delta psi values plotted indicate difference in isoforms.(A3SS, Alternative 3' splice sites; A5SS, Alternative 5' splice sites; AFE, Alternative first exons; ALE, Alternative last exons; MXE, Mutually exclusive exons; RI, Retained introns; SE, Skipped exons; Bound, IGF2BP3 eCLIP target). D) UCSC Genome Browser snapshots of the Hoxa9, Hoxa7, and Cd69 loci. Each panel shows the exon-intron structure of the gene and unique read coverage from 3 eCLIP biological replicates from WT/MLL-Af4 Lin-cells. The maximum number of reads at each position is indicated to the left of each histogram. F) Ratio of Hoxa9T/Hoxa9 isoforms in WT/MLL-Af4 and I3KO/MLL-Af4 Lin-cells by RT-qPCR (t-test; $^{**}P < 0.01$, $^{****}P < 0.0001$).

6. Chapter 6: BiocSwirl - Interactive R Tutorials for Bioinformatics

Lisa Cao[1], Julia Philipp[2], Matthew A. Moss[3], Mariam Arab[1], Jasdeep Singh[4], Sourav Singh[5], Almas Khan[6]

1 Simon Fraser University, Burnaby BC, Canada

2 University of California, Santa Cruz, CA, USA

3 Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, USA

4 Punjab Agricultural University, Ludhiana, India

5 Vishwakarma Institute of Information Technology, Maharashtra, India

6 University of British Columbia, Vancouver BC, Canada

** Corresponding Author*

6.1 Abstract

Bioinformatics has grown in adoption in many non-technical fields, advancing so rapidly that traditional bench scientists are finding it challenging to keep up with gold standard workflows. Complex bioinformatics data analyses become difficult to retain when learned using traditional teaching methods or static formats, which are often not updated or available for feedback. As a continuously maintained R package, BiocSwirl makes learning bioinformatics concepts hands-on through modular, self-paced, interactive course material targeted at beginners. Our courses communicate a wide range of concepts, from first steps in R and data science to high-level analyses and visualizations of biological datasets from within the R console. BiocSwirl's primary feature is the ability to provide real-time feedback on the user's R syntax that promotes good coding practices, open science, and reproducibility in students. Our currently available curriculum includes introductions to RNAseq, scRNAseq, ChIPseq, and Biostatistics as well

as a "Data Science for Life Scientists" course. Ultimately, BiocSwirl removes the barrier to entry into R Data Science in bioinformatics while instilling a foundation of open science practices for students and researchers.

6.2 Introduction

Bioinformatics and data science have grown to be adopted into many traditionally non-computer science fields that lack the time to devote to mastering the skills beyond a surface level understanding. Swirlstats and terminal-based Python courses, such as the browser-based platform Codecademy, are commonly used to teach data science, statistics, and bioinformatics to complete beginners.

More often than not researchers without coding experience or much computational background are expected to analyse their own data and stay on top of state of the art analysis pipelines. Bioconductor and CRAN provide thousands of packages to aid these workflows. The flood of information can be overwhelming and it is almost impossible to identify the current gold standard of bioinformatics analyses from just these packages and their vignettes.

While there are currently many resources available to interdisciplinary researchers interested in learning bioinformatics, including R Bookdown tutorials, in person workshops, Youtube videos, and the popular bioinformatics platform Rosalind, these platforms each have their own shortcomings. Blogpost and video-based tutorials don't provide the ability to troubleshoot, ask questions, or receive feedback. In-person workshops have time limitations and often unfavorable student to teacher ratios.

Here we describe BiocSwirl, an R package that delivers interactive courses on bioinformatics workflows with real life, standardized data, and provides real-time feedback

to the user. BiocSwirl courses currently include an "Introduction to Data Science for Life Scientists", an introduction to statistics, as well as courses covering the analysis and interpretation of of ChIP-seq, scRNA-seq, bulk RNAseq, and will include more courses (including a neuroinformatics suite, methylation array analysis, and a GAPIT pipeline) in future releases (Figure 6.1, current course structure).

These courses are based in R, which is continuously updated and freely available. They are easily configurable and break down complex bioinformatics workflows into simple, easily understandable steps that bridge coding skills and biological interpretation. They contain standardized datasets and place strong emphasis on good coding practices, open science, and reproducibility. The modular structure of the courses can easily be updated to match current best practices and expanded to communicate a wide range of different, and even personalized, workflows.

6.3 Methods

6.3.1 Implementation

All BiocSwirl courses are based on the R package swirl and are constructed using the R package swirlify, which aids in writing swirl courses by providing question and answer templates in YAML syntax. These lessons are loaded into the R environment using the swirl package.

Each course typically contains multiple lessons that can be completed in order, but can be done independently of each other as well. A text file, called the manifest file, within the course will determine the suggested order of the lessons and the start screen of the course once it has been loaded into RStudio.

Each lesson consists of four standard files. The main file is a .yaml file containing the content of the lesson, covering a bioinformatics analysis workflow, in a question and

answer format. Different types of questions, including command questions, text and multiple choice questions, as well as code to display figures and link to urls, can be used as their format is provided by swirlify. The answers entered by the user in the console of R studio will be tested with answer tests also provided by swirlify. Hints are provided for each question to guide the user if they need additional information or get the answer wrong. The dependson.txt file indicates the CRAN packages the lesson depends on, which will be loaded for the user as the lesson starts. Further, the initLesson.R file contains installation and load commands for bioconductor packages if needed as well as loading the data objects that are being used in the lesson. These files will be loaded and available in the environment as the lesson starts. Finally, a customTests.R file provides a basis for including custom answer tests if needed.

In addition to these four files, the course creator can provide .Rdata files containing standardized biological data that can be loaded and manipulated (analyzed) within the lesson,.R files containing code snippets to generate figures, as well as png and html files that can be referenced and opened by the user within the lesson interface.

The final course is packaged into a single .swc file and available on the BiocSwirl github in the appropriate repository or as part of the BiocSwirl R package (https://github.com/biocswirl-dev-team/BiocSwirl).

6.3.2 Operation

The use of the BiocSwirl R package requires a local up-to-date R installation, RStudio for an interactive and visual user interface, and the R package swirl. The data sets are small enough to allow all workflows to be run on a local computer, so no cloud computing is required.

The courses will be delivered to the user in the form of an R package that will guide them through the selection and installation of the desired courses from our github repository. After starting the swirl package with the command 'swirl()', the user can type in a user name and select which course they'd like to work on. Each course typically consists of multiple lessons that can be taken in the suggested order (set in the manifest file) or at will.

The lesson consists of text or multiple-choice questions and code prompts. The user can enter their answer in the console of R studio. If the answer is correct, the lesson will proceed to the next question. If the answer is incorrect, a hint will be provided to the user and they can try again.

Each course contains a short lesson explaining how to load and install CRAN and bioconductor packages, a lesson explaining how to save and export data at the end of an analysis, as well as an introduction to the assay and the biological data that is being covered in the course at hand.

6.4 Results

6.4.1 Use Cases

Current Use Case 1: A student or scientist is completely new to data science and using R. They can download, install, and take the 'Intro to data science for life scientists' course as described on our github. In 5 lessons, they will be guided through basic R syntax, data manipulation and storage, R tidyverse, basic data visualization, and good coding practices. Each of these lessons use precompiled, reproducible data, and real world coding problems. Upon completion of the introductory course, they will be able to take other, more advanced, statistics and bioinformatics courses or work on their own data wrangling.

Current Use Case 2: A wet lab biologist just collected their own bioinformatics datasets and has basic data science experience but doesn't know how to analyze bioinformatics data.They can choose from a variety of bioinformatics classes depending on the purpose, question, and source of their experimental data. They can work through these interactive courses at their own pace and will learn all the necessary steps from processing the data up to high-level analysis, visualization, and interpretation. At the end of the course, they will be able to apply their knowledge of the analysis workflow to their own data and create publication quality figures.

Current Use Case 3: Classroom application with students with various levels of coding experience. Students with various levels of coding, data science, or bioinformatics experience can participate in these courses in a classroom setting. They can proceed through the courses at their own pace and get real-time feedback even in big class rooms with hundreds of students. They will be able to learn the subject independently and interactively. Since the Biocswirl courses are easily adaptable, the lecture material can be integrated and converted to the Biocswirl course format if desired and taught to much bigger classrooms at the same efficiency.

Further lessons will allow for other users to take advantage of our platform, and will continue to make bioinformatics and biological data science more accessible to the broader scientific community.

6.5 Conclusion and next steps

In the future, we plan to expand the collection of courses included with the BiocSwirl package and to include a wider variety of data analysis workflows. We are further aiming to reach out to and collaborate with educators in the data science and bioinformatics community to facilitate adoption of BiocSwirl courses into classrooms and

to gain valuable user feedback. We are also planning on hosting workshops, introducing more users to the BiocSwirl courses and collecting both pre- and post-workshop assessments to monitor the quality of our courses and the progress of the coursetakers.

6.6 Data and software availability

R package repository	https://github.com/biocswirl-dev-team/BiocSwirl
RNAseq course	https://github.com/biocswirl-dev-team/BiocSwirl_RNAseq
scRNAseq course	https://github.com/biocswirl-dev-team/BiocSwirl_scRNA-seq
ChIPseq course	https://github.com/biocswirl-dev-team/BiocSwirl_ChIPseq
Intro to Data Science course	https://github.com/biocswirl-dev-team/BiocSwirl_Intro_to_Data_Science
Intro to Stats course	https://github.com/biocswirl-dev-team/BiocSwirl_Intro_to_Stats

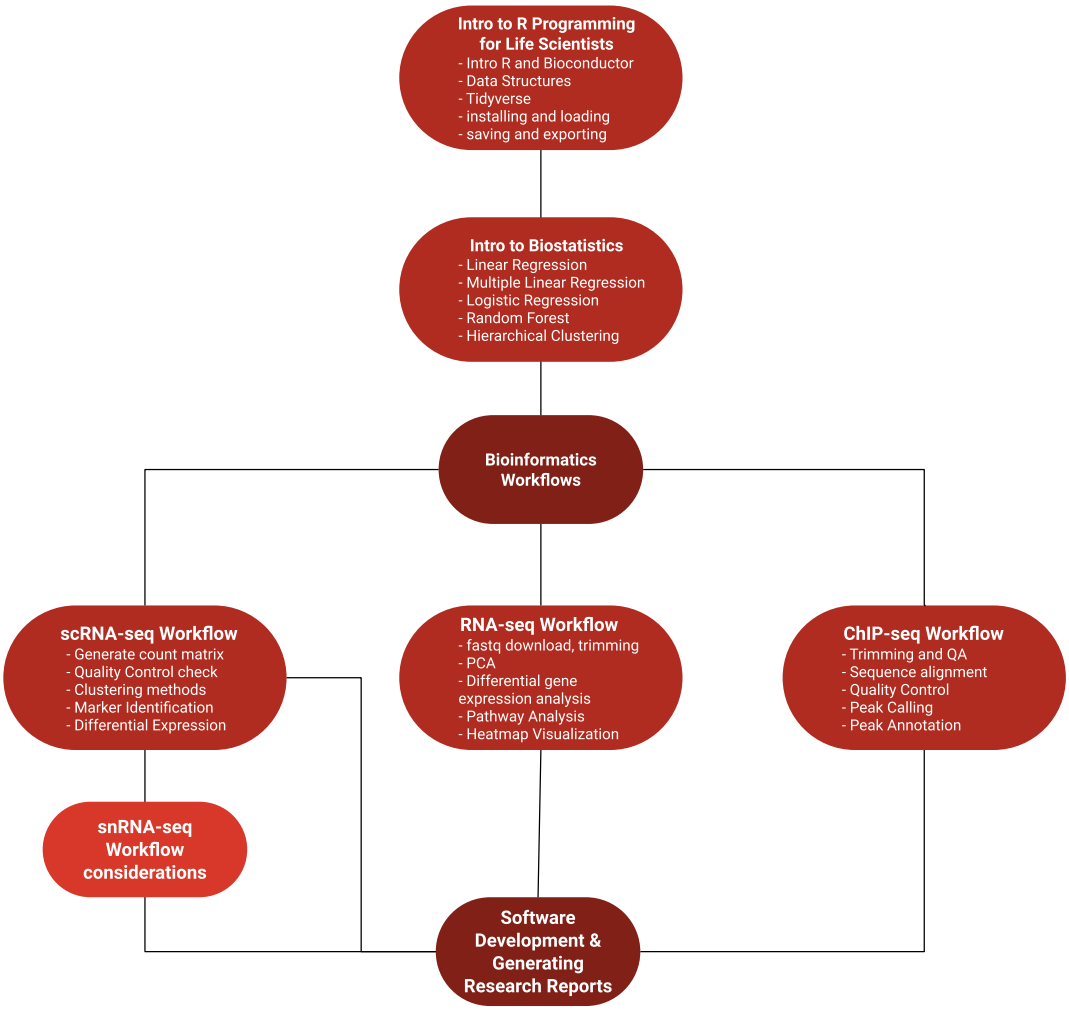

Figure 6.1 BiocSwirl course syllabus

7. Chapter 7: Concluding Remarks and Future Directions

In this dissertation, I have presented the analyses and results of multiple complex transcriptomics datasets, including short-read RNAseq, iCLIP, and eCLIP. I studied the coupling of alternative splicing to translational control in human and primate cell lines and the changes in the cis-regulatory landscape as a potential mechanism of coupling.

In addition, I analyzed multiple iCLIP and eCLIP datasets, both recently collected and publicly available on the ENCODE platform. I examined the interactions between IGF2BP3, a known oncogenic amplifier, and its downstream targets in leukemogenesis. Using these datasets, I was able to elucidate the binding preferences and specificity of IMP3 and identify key players in the IMP3-dependent leukemogenesis. I was further able to show the preferential interaction of different RBPs to different regions of the pre-miRNA (Dargyte et al., 2020), including some binding preferences (e.g., Lin28) that align well with previous literature and shedding light on binding preferences of splicing factors (e.g., SRSF1 and U2AF).

The majority of the projects presented in this dissertation took advantage of high-throughput sequencing data. While these projects have been ongoing, sequencing techniques used to analyze mRNA and small RNA expression and alternative splicing patterns have continuously evolved. This trend will undoubtedly continue and open up new avenues of exploring post-transcriptional regulation of gene expression.

The advent of long-read sequencing techniques such as oxford nanopore (citation) and PacBio (citation) opens up many possibilities for analyzing alternative splicing. Repeating, for example, the Frac-seq analysis of primate iPSCs as presented in Chapter 2 but sequencing the fraction using long-read sequencing technologies would allow for complete detection of alternative splicing patterns and identification of full-length isoforms instead of three-exon alternative splicing events. Using long-read sequencing

in this context would also improve the analysis of cis-regulatory elements drastically. It is well known that cis- and trans-regulatory factors interact with each other to modify their effects, and the interactome will be much better captured and can be predicted more reliably using long-read sequencing.

Overall, cheaper and better sequencing technologies will expand the number of datasets available and will undoubtedly expand our understanding of protein-RNA interactions.

Appendix

List of CLIP experiments available on ENCODE

Accession	Target of assay / Gene Symbol	Biosample
ENCSR356MSW	SSB	K562
ENCSR958FKZ	PABPC4	K562
ENCSR739VVT	APOBEC3C	K562
ENCSR891RIC	NIPBL	K562
ENCSR202HKN	WDR3	K562
ENCSR735HOK	YBX3	HepG2
ENCSR820UYE	PABPN1	HepG2
ENCSR085JPB	WDR43	HepG2
ENCSR038JME	WRN	K562
ENCSR734ZHL	UTP3	K562
ENCSR057DWB	TIA1	K562
ENCSR568DZW	TAF15	K562
ENCSR258QKO	XRCC6	K562
ENCSR121NVA	PRPF8	HepG2
ENCSR916SRV	TBRG4	HepG2
ENCSR543TPH	AGGF1	HepG2
ENCSR987FTF	RBFOX2	HepG2
ENCSR490IEE	UCHL5	HepG2
ENCSR943MHU	SAFB2	K562
ENCSR041NUV	EIF3D	HepG2
ENCSR919HSE	CSTF2T	HepG2
ENCSR887FHF	FASTKD2	K562
ENCSR887LPK	EWSR1	K562
ENCSR061SZV	DGCR8	HepG2
ENCSR685AUR	ZNF800	HepG2
ENCSR050BDZ	SDAD1	HepG2
ENCSR693JWP	EXOSC5	HepG2
ENCSR483NOP	SLBP	K562
ENCSR468FSW	SRSF7	K562
ENCSR438KWZ	ILF3	K562

ENCSR000SSH	SLTM	K562
ENCSR819XBT	AATF	K562
ENCSR907GUB	ZC3H11A	HepG2
ENCSR539ZTS	TROVE2	K562
ENCSR663WES	BUD13	K562
ENCSR145NLR	DDX51	K562
ENCSR970FEW	DDX52	HepG2
ENCSR876EYA	BCLAF1	HepG2
ENCSR979EWD	STAU2	HepG2
ENCSR513NDD	SRSF7	HepG2
ENCSR987NYS	FAM120A	HepG2
ENCSR018ZUE	FKBP4	HepG2
ENCSR576SHT	DDX42	K562
ENCSR721HPX	G3BP1	HepG2
ENCSR194HZU	NOLC1	HepG2
ENCSR484LAB	SAFB	K562
ENCSR586DGV	ZNF800	K562
ENCSR795CAI	HNRNPL	K562
ENCSR485QCG	BCCIP	HepG2
ENCSR308YNT	PUM1	K562
ENCSR606BPV	AQR	K562
ENCSR570WLM	QKI	HepG2
ENCSR961OKA	LARP7	HepG2
ENCSR825SVO	AARS	K562
ENCSR069EVH	FUS	K562
ENCSR361OCV	NIP7	HepG2
ENCSR906ZJF	SDAD1	K562
ENCSR744GEU	IGF2BP1	HepG2
ENCSR773KRC	SRSF9	HepG2
ENCSR456JJQ	RBM22	HepG2
ENCSR861GYE	LIN28B	HepG2
ENCSR993FMY	TROVE2	HepG2
ENCSR128VXC	SND1	K562
ENCSR756CKJ	RBFOX2	K562
ENCSR647HOX	HLTF	HepG2
ENCSR133QEA	SF3B1	K562

ENCSR412NOW	HNRNPM	K562
ENCSR062NNB	IGF2BP2	K562
ENCSR406OOZ	SUB1	HepG2
ENCSR277DEO	NKRF	HepG2
ENCSR766FAC	RPS3	HepG2
ENCSR755TJC	HNRNPUL1	HepG2
ENCSR059CWF	SBDS	K562
ENCSR121GQH	SERBP1	K562
ENCSR774RFN	FXR1	K562
ENCSR196INN	RBM15	K562
ENCSR571ROL	XRCC6	HepG2
ENCSR432XUP	SRSF1	K562
ENCSR658IQB	SMNDC1	K562
ENCSR724RDN	HNRNPL	HepG2
ENCSR339FUY	PCBP2	HepG2
ENCSR322HHA	TIAL1	HepG2
ENCSR964VOX	UTP18	K562
ENCSR903PRV	FTO	HepG2
ENCSR154CSN	DDX52	K562
ENCSR488JKQ	UTP18	HepG2
ENCSR440SUX	MATR3	K562
ENCSR256CHX	PCBP1	HepG2
ENCSR018WPY	AQR	HepG2
ENCSR989SMC	FTO	K562
ENCSR580OFI	SUPV3L1	K562
ENCSR001KKZ	PHF6	K562
ENCSR267UCX	HNRNPM	HepG2
ENCSR977OXG	PRPF4	HepG2
ENCSR023PKW	EIF3G	K562
ENCSR973HOJ	FXR2	HepG2
ENCSR922WJV	PCBP1	K562
ENCSR366DGX	KHSRP	HepG2
ENCSR337XGI	SAFB	HepG2
ENCSR356ZMO	AKAP1	HepG2
ENCSR181NRW	ZC3H8	K562
ENCSR349CMI	WDR43	K562

ENCSR177QQY	AKAP1	K562
ENCSR249ROI	HNRNPC	K562
ENCSR046JHH	CPEB4	K562
ENCSR539BEV	UPF1	HepG2
ENCSR352STY	SSB	HepG2
ENCSR464OSH	FUS	HepG2
ENCSR089BXO	ABCF1	K562
ENCSR040QLV	DDX21	K562
ENCSR999YGP	GRWD1	K562
ENCSR861PAR	NONO	K562
ENCSR534YOI	PRPF8	K562
ENCSR841EQA	TAF15	HepG2
ENCSR373ODC	SMNDC1	HepG2
ENCSR484LTQ	NCBP2	K562
ENCSR661ICQ	PUM2	K562
ENCSR366YOG	QKI	K562
ENCSR862QCH	U2AF1	K562
ENCSR628IDK	KHDRBS1	K562
ENCSR303OQD	METAP2	K562
ENCSR081JYH	NSUN2	K562
ENCSR135VMS	LSM11	HepG2
ENCSR584TCR	TARDBP	K562
ENCSR061EVO	SND1	HepG2
ENCSR828ZID	HNRNPK	HepG2
ENCSR923NKN	DDX55	K562
ENCSR202BFN	U2AF2	HepG2
ENCSR893EFU	DDX6	K562
ENCSR265ZIS	GTF2F1	HepG2
ENCSR520BZQ	HNRNPU	K562
ENCSR532VUB	CPSF6	K562
ENCSR867ZVK	YWHAG	K562
ENCSR893RAV	U2AF2	K562
ENCSR623VEQ	TIA1	HepG2
ENCSR653HQC	DROSHA	K562
ENCSR786TSC	ILF3	HepG2
ENCSR655NZA	XRN2	HepG2

ENCSR384MWO	CSTF2	HepG2
ENCSR001VAC	NOLC1	K562
ENCSR279UJF	SF3B4	HepG2
ENCSR580MFX	SUPV3L1	HepG2
ENCSR845VGB	DDX55	HepG2
ENCSR840DRD	CSTF2T	K562
ENCSR989VIY	SRSF1	HepG2
ENCSR006OEQ	FAM120A	K562
ENCSR668MJX	GRSF1	HepG2
ENCSR120EAR	RPS3	K562
ENCSR307YIW	EIF4G2	K562
ENCSR290VLT	MATR3	HepG2
ENCSR295OKT	RBM22	K562
ENCSR981WKN	PTBP1	K562
ENCSR769UEW	HNRNPA1	HepG2
ENCSR830BSQ	BUD13	HepG2
ENCSR571VHI	HNRNPUL1	K562
ENCSR269AJF	RPS11	K562
ENCSR200DKE	MTPAP	K562
ENCSR834YLD	DROSHA	HepG2
ENCSR820DQJ	NOL12	HepG2
ENCSR527DXF	EFTUD2	HepG2
ENCSR224QWC	FXR2	K562
ENCSR725ARB	AGGF1	K562
ENCSR565DGW	DHX30	HepG2
ENCSR754NDA	RBM15	HepG2
ENCSR712IAG	ZC3H11A	K562
ENCSR805SRN	LARP4	HepG2
ENCSR349KMG	UCHL5	K562
ENCSR550DVK	HNRNPC	HepG2
ENCSR486YGP	FUBP3	HepG2
ENCSR930BZL	DDX3X	K562
ENCSR438GZQ	KHSRP	K562
ENCSR589YHM	HLTF	K562
ENCSR154HRN	HNRNPA1	K562
ENCSR023UHL	FASTKD2	HepG2

ENCSR206RXT	AKAP8L	K562
ENCSR657TZZ	ZNF622	K562
ENCSR921SXC	XPO5	HepG2
ENCSR506OTC	TBRG4	K562
ENCSR844RVX	EFTUD2	K562
ENCSR648LAH	DDX3X	HepG2
ENCSR331VNX	FMR1	K562
ENCSR663NRA	ZRANB2	K562
ENCSR893NWB	GRWD1	HepG2
ENCSR975KIR	IGF2BP1	K562
ENCSR657TZB	XRN2	K562
ENCSR815VVI	CDC40	HepG2
ENCSR331MIC	SF3A3	HepG2
ENCSR197INS	PPIL4	K562
ENCSR965DLL	SFPQ	HepG2
ENCSR456KXI	LARP7	K562
ENCSR999WKT	DDX24	K562
ENCSR993OLA	IGF2BP3	HepG2
ENCSR238CLX	GEMIN5	K562
ENCSR328LLU	U2AF1	HepG2
ENCSR489ABS	RBM5	HepG2
ENCSR529GSJ	DHX30	K562
ENCSR506UPY	SUGP2	HepG2
ENCSR214BZA	DDX59	HepG2
ENCSR365NVO	TRA2A	K562
ENCSR291XPT	PUS1	K562
ENCSR820WHR	POLR2G	HepG2
ENCSR736AAG	GTF2F1	K562
ENCSR916XIV	EIF3H	HepG2
ENCSR301TFY	DKC1	HepG2
ENCSR647CLF	GPKOW	K562
ENCSR141OIM	DDX6	HepG2
ENCSR301UQM	GNL3	K562
ENCSR267OLV	SF3B4	K562
ENCSR351PVI	SLTM	HepG2
ENCSR529FKI	YBX3	K562

ENCSR970NKP	LIN28B	K562
ENCSR097NEE	PPIG	HepG2
ENCSR268ETU	HNRNPK	K562
ENCSR888YTT	LARP4	K562
ENCSR022BVV	LSM11	K562
ENCSR018RVZ	NCBP2	HepG2
ENCSR240MVJ	HNRNPU	HepG2
ENCSR867DSZ	NPM1	K562
ENCSR947JVR	DGCR8	K562
ENCSR456ASB	UPF1	K562
ENCSR013CTQ	EXOSC5	K562
ENCSR384KAN	PTBP1	HepG2
ENCSR314UMJ	TRA2A	HepG2

List of RNAcompete experiments on ENCODE

Gene name	ID	Species	IUPAC
A1CF	RNCMPT00001	Homo_sapiens	WUAAUUR
A2BP1	RNCMPT00123	Drosophila_melanogaster	UGCAUG
RBFOX1	RNCMPT00168	Homo_sapiens	WGCAUGM
An_0265	RNCMPT00265	Aspergillus_nidulans	ACCYMA
ANKHD1	RNCMPT00002	Homo_sapiens	AGACGWW
ARET	RNCMPT00003	Drosophila_melanogaster	UKUKUGU
ARET	RNCMPT00114	Drosophila_melanogaster	UKUKUGU
ARET	RNCMPT00270	Drosophila_melanogaster	UKUGUGU
ASD-1	RNCMPT00180	Caenorhabditis_elegans	WGCAUGH
At_0284	RNCMPT00284	Arabidopsis_thaliana	DGWGUGD
B52	RNCMPT00134	Drosophila_melanogaster	GGASGRV
BRU-3	RNCMPT00122	Drosophila_melanogaster	KUGKUGU
BRUNOL4	RNCMPT00004	Homo_sapiens	KGUGUKK
BRUNOL5	RNCMPT00166	Homo_sapiens	UGUGUKK
BRUNOL6	RNCMPT00187	Homo_sapiens	UGUGDKG
An_0287	RNCMPT00287	Aspergillus_nidulans	UACUAMK
CG11360	RNCMPT00129	Drosophila_melanogaster	GAGUDW
CG14718	RNCMPT00006	Drosophila_melanogaster	CAGAKB
CG17838	RNCMPT00131	Drosophila_melanogaster	BAAAUUD
CG2931	RNCMPT00147	Drosophila_melanogaster	DACUAAG
CG2950	RNCMPT00007	Drosophila_melanogaster	CRACGAV
CG33714	RNCMPT00009	Drosophila_melanogaster	BBGCGUG
CG5213	RNCMPT00010	Drosophila_melanogaster	URCUUU
CG7804	RNCMPT00146	Drosophila_melanogaster	WGKRUGR
CG7903	RNCMPT00144	Drosophila_melanogaster	HWUGCGR
CNOT4	RNCMPT00008	Drosophila_melanogaster	ASACAHW
CNOT4	RNCMPT00156	Homo_sapiens	GACAGA
CPEB2	RNCMPT00012	Homo_sapiens	CHUUUUU
CPEB4	RNCMPT00158	Homo_sapiens	UUUUUU
CPO	RNCMPT00133	Drosophila_melanogaster	WGCACA
Pr_0249	RNCMPT00249	Phytophthora_ramorum	UUGCACD
DAZAP1	RNCMPT00013	Homo_sapiens	UAGKWWR
Ot_0262	RNCMPT00262	Ostreococcus_tauri	UUUUUUK

Ot_0263	RNCMPT00263	Ostreococcus_tauri	AGAACRD
Pp_0206	RNCMPT00206	Physcomitrella_patens	RACCUW
Nv_0278	RNCMPT00278	Nematostella_vectensis	AGAYASA
EIF-2ALPHA	RNCMPT00273	Drosophila_melanogaster	WGCAUG
ELAV	RNCMPT00121	Drosophila_melanogaster	UUUDKUU
ENOX1	RNCMPT00149	Homo_sapiens	HRKACAG
Rbm38	RNCMPT00283	Danio_rerio	KWGUGUG
HNRNPR	RNCMPT00288	Gallus_gallus	MMAAAWY
A1CF	RNCMPT00291	Gallus_gallus	DUAAUUV
RBM47	RNCMPT00279	Gallus_gallus	GAUGAW
Rbm24	RNCMPT00285	Tetraodon_nigroviridis	WGUGUG
Hnrnpr	RNCMPT00289	Xenopus_tropicalis	MAAAAAG
Rbm47	RNCMPT00280	Xenopus_tropicalis	GAUGAWH
Rbm42	RNCMPT00282	Xenopus_tropicalis	WACUAC
Syncrip	RNCMPT00281	Xenopus_tropicalis	MAAAWWD
ESRP2	RNCMPT00150	Homo_sapiens	UGGGRAD
Ng_0261	RNCMPT00261	Naegleria_gruberi	RDUUUUG
Pp_0237	RNCMPT00237	Physcomitrella_patens	WUGGAG
ETR-1	RNCMPT00183	Caenorhabditis_elegans	KKUDUGU
EXC-7	RNCMPT00014	Caenorhabditis_elegans	YUDRGUU
FMR1	RNCMPT00015	Drosophila_melanogaster	AHGGACR
FMR1	RNCMPT00016	Homo_sapiens	KGACARG
FNE	RNCMPT00120	Drosophila_melanogaster	UUKDGUU
FOX-1	RNCMPT00017	Caenorhabditis_elegans	WGCAUGM
FUS	RNCMPT00018	Homo_sapiens	CGCGC
SRSF10	RNCMPT00088	Homo_sapiens	ARAGRRR
SRSF10	RNCMPT00089	Homo_sapiens	AGAGARR
SRSF10	RNCMPT00090	Homo_sapiens	AGAGAVV
SRSF10	RNCMPT00019	Homo_sapiens	AGAGAVM
FXR1	RNCMPT00161	Homo_sapiens	AYGACR
FXR2	RNCMPT00020	Homo_sapiens	DGACRRR
G3BP2	RNCMPT00021	Homo_sapiens	AGGAUDR
Pp_0229	RNCMPT00229	Physcomitrella_patens	UGUUUUD
Pp_0228	RNCMPT00228	Physcomitrella_patens	YUUUUUU
HNRNPA1	RNCMPT00022	Homo_sapiens	DUAGGGW
HNRNPA1L2	RNCMPT00023	Homo_sapiens	DUAGGGW

HNRNPA2B1	RNCMPT00024	Homo_sapiens	DUAGGGW
HNRNPAB	RNCMPT00245	Tetraodon_nigroviridis	WRGWUAG
HNRNPC	RNCMPT00025	Homo_sapiens	HUUUUUK
HNRNPCL1	RNCMPT00167	Homo_sapiens	HUUUUUK
HNRNPH2	RNCMPT00160	Homo_sapiens	GGGAGGG
HNRNPK	RNCMPT00026	Homo_sapiens	CCAWMCC
HNRNPL	RNCMPT00091	Homo_sapiens	ACACRAV
HNRNPL	RNCMPT00027	Homo_sapiens	AMAYAMA
HNRPLL	RNCMPT00178	Homo_sapiens	RCAHACA
HOW	RNCMPT00118	Drosophila_melanogaster	ACUAACV
HRB27C	RNCMPT00093	Drosophila_melanogaster	UAGGWUA
HRB27C	RNCMPT00028	Drosophila_melanogaster	UAGGWWA
HRB87F	RNCMPT00029	Drosophila_melanogaster	GGUAGGG
HRB98DE	RNCMPT00095	Drosophila_melanogaster	GGUAGGG
HRB98DE	RNCMPT00096	Drosophila_melanogaster	RGUAGGG
HRB98DE	RNCMPT00094	Drosophila_melanogaster	GKUAGGK
Hrp1p	RNCMPT00031	Saccharomyces_cerevisiae	UAYRUAV
HuR	RNCMPT00112	Homo_sapiens	UUWGUUU
HuR	RNCMPT00117	Homo_sapiens	UUURKUU
HuR	RNCMPT00274	Homo_sapiens	UUUUUUK
HuR	RNCMPT00032	Homo_sapiens	UUDUUUU
HuR	RNCMPT00136	Homo_sapiens	UUKRUUU
IGF2BP2	RNCMPT00033	Homo_sapiens	VMAHWCA
IGF2BP3	RNCMPT00172	Homo_sapiens	AMAHWCA
KHDRBS1	RNCMPT00169	Homo_sapiens	AUAAAAV
KHDRBS1	RNCMPT00062	Mus_musculus	UAAAAVV
KHDRBS2	RNCMPT00185	Homo_sapiens	RAUAAAM
KHDRBS3	RNCMPT00034	Homo_sapiens	AUAAAV
LARK	RNCMPT00035	Drosophila_melanogaster	DCGCGCG
LARK	RNCMPT00097	Drosophila_melanogaster	DCGCGCG
LARK	RNCMPT00124	Drosophila_melanogaster	SGCGCG
LIN28A	RNCMPT00036	Homo_sapiens	HGGAGWA
LIN28A	RNCMPT00162	Homo_sapiens	YGGAGGR
Lm_0212	RNCMPT00212	Leishmania_major	CAUUUU
Lm_0223	RNCMPT00223	Leishmania_major	KUWCACG
Lm_0254	RNCMPT00254	Leishmania_major	HGAACGM

Lm_0255	RNCMPT00255	Leishmania_major	AAAMAAA
MAL13P1.35	RNCMPT00234	Plasmodium_falciparum	RWUACAC
MAL8P1.40	RNCMPT00197	Plasmodium_falciparum	WUACAUR
MATR3	RNCMPT00037	Homo_sapiens	MAUCUUR
MBNL1	RNCMPT00038	Homo_sapiens	GCUUGC
MEC-8	RNCMPT00181	Caenorhabditis_elegans	DWGCACA
MEX-5	RNCMPT00039	Caenorhabditis_elegans	UAAUAW
MOD	RNCMPT00140	Drosophila_melanogaster	ADUGGAA
MSI1	RNCMPT00041	Homo_sapiens	UAGUWRG
MSI1	RNCMPT00176	Homo_sapiens	UAGKWRG
MSI	RNCMPT00099	Drosophila_melanogaster	AGUAGKD
MSI	RNCMPT00040	Drosophila_melanogaster	WGUAGKD
MSI	RNCMPT00100	Drosophila_melanogaster	AGUAGGD
MUB	RNCMPT00137	Drosophila_melanogaster	WACCCKW
Nab2p	RNCMPT00042	Saccharomyces_cerevisiae	AAAAAAR
NCU02404	RNCMPT00238	Neurospora_crassa	GGWGGAD
NCU08034	RNCMPT00209	Neurospora_crassa	WGCACA
ORB2	RNCMPT00126	Drosophila_melanogaster	KUUUKKK
PABPC1	RNCMPT00155	Homo_sapiens	ARAAAAM
PABPC3	RNCMPT00153	Homo_sapiens	RAAAACM
PABPC4	RNCMPT00043	Homo_sapiens	AAAAAAR
PABPC5	RNCMPT00171	Homo_sapiens	AGAAADU
PABP	RNCMPT00139	Drosophila_melanogaster	GAAAAHV
PABPN1	RNCMPT00157	Homo_sapiens	ARAAGA
PAPI	RNCMPT00011	Drosophila_melanogaster	KGUKUGU
PCBP1	RNCMPT00186	Homo_sapiens	CCWWHCC
PCBP1	RNCMPT00239	Mus_musculus	CYUUCC
Pcbp2	RNCMPT00246	Danio_rerio	WWUCCC
PCBP2	RNCMPT00044	Homo_sapiens	CCYYCCH
PCBP3	RNCMPT00215	Mus_musculus	UUUYCC
PF10_0068	RNCMPT00199	Plasmodium_falciparum	GGWGGA
PF10_0214	RNCMPT00240	Plasmodium_falciparum	WUWCCGA
PF13_0315	RNCMPT00200	Plasmodium_falciparum	KURCAUD
PFF0320c	RNCMPT00235	Plasmodium_falciparum	ACUAAWC
PFI1435w	RNCMPT00202	Plasmodium_falciparum	UKUUUUK
PFI1695c	RNCMPT00203	Plasmodium_falciparum	CYWKCAC

PPRC1	RNCMPT00045	Homo_sapiens	SSGCGCS
SFPQ	RNCMPT00177	Homo_sapiens	KURRUKK
PTBP1	RNCMPT00269	Homo_sapiens	HYUUUYU
PTBP1	RNCMPT00268	Homo_sapiens	HYUUUYU
PUF68	RNCMPT00141	Drosophila_melanogaster	UAWDRGR
PUM	RNCMPT00101	Drosophila_melanogaster	UGUAMAK
PUM	RNCMPT00102	Drosophila_melanogaster	UGUAYAK
PUM	RNCMPT00103	Drosophila_melanogaster	UGUAMRK
PUM	RNCMPT00104	Drosophila_melanogaster	UGUAMAK
PUM	RNCMPT00105	Drosophila_melanogaster	UGUAWDU
PUM	RNCMPT00046	Drosophila_melanogaster	UGUAHAK
QKI	RNCMPT00047	Homo_sapiens	ACUAACV
QKR58E-1	RNCMPT00142	Drosophila_melanogaster	AUAAUWM
RALY	RNCMPT00159	Homo_sapiens	UUUUUUB
RBM24	RNCMPT00184	Homo_sapiens	WGWGUGD
RBM28	RNCMPT00049	Homo_sapiens	GWGUAGD
RBM38	RNCMPT00051	Mus_musculus	KKGUGUK
RBM3	RNCMPT00050	Homo_sapiens	RADACKA
RBM41	RNCMPT00053	Homo_sapiens	WUACWUK
RBM42	RNCMPT00151	Homo_sapiens	AACUAMG
Rbm4.3	RNCMPT00248	Danio_rerio	ACGRCG
RBM45	RNCMPT00241	Homo_sapiens	GACGAMV
RBM46	RNCMPT00054	Homo_sapiens	RAUSAWD
RBM4	RNCMPT00113	Homo_sapiens	GCGCGSG
RBM4	RNCMPT00052	Homo_sapiens	GCGCGSS
RBM5	RNCMPT00154	Homo_sapiens	SAAGGRG
RBM5	RNCMPT00055	Homo_sapiens	GARGGWR
RBM6	RNCMPT00170	Homo_sapiens	HAUCCAR
RBM8A	RNCMPT00056	Homo_sapiens	RYGCGCB
RBMS1	RNCMPT00152	Homo_sapiens	KAUAUAS
RBMS3	RNCMPT00173	Homo_sapiens	HAUAUA
RBMS3	RNCMPT00057	Homo_sapiens	MUAUAKM
RBP1	RNCMPT00058	Drosophila_melanogaster	WCAACRR
RBP1-LIKE	RNCMPT00127	Drosophila_melanogaster	AUCADCR
RBP9	RNCMPT00132	Drosophila_melanogaster	UUDDGUU
REF2	RNCMPT00059	Drosophila_melanogaster	AGAAGRM

RIN	RNCMPT00138	Drosophila_melanogaster	RRWUGAW
RNP4F	RNCMPT00060	Drosophila_melanogaster	AGAKARR
RO3G_00049	RNCMPT00205	Rhizopus_oryzae	RGGGGAA
ROX8	RNCMPT00148	Drosophila_melanogaster	MYAUUUU
RSF1	RNCMPT00061	Drosophila_melanogaster	ACGACGV
SAMD4A	RNCMPT00063	Homo_sapiens	GCKGGHM
SART3	RNCMPT00064	Homo_sapiens	ARAAAAM
SF1	RNCMPT00065	Drosophila_melanogaster	ACUAAYV
SF2	RNCMPT00066	Drosophila_melanogaster	DGGAGGA
SRSF1	RNCMPT00106	Homo_sapiens	GRAGGA
SRSF1	RNCMPT00109	Homo_sapiens	GGASGRV
SRSF1	RNCMPT00107	Homo_sapiens	GGAGGA
SRSF1	RNCMPT00108	Homo_sapiens	GGASGRV
SRSF1	RNCMPT00110	Homo_sapiens	AGGASM
SRSF1	RNCMPT00163	Homo_sapiens	GGRGGAV
Sf3b4	RNCMPT00224	Danio_rerio	CAAAAG
SHEP	RNCMPT00174	Drosophila_melanogaster	WAUWUWD
SHEP	RNCMPT00175	Drosophila_melanogaster	WUAUWWA
SHEP	RNCMPT00068	Drosophila_melanogaster	AUAUUWD
SM	RNCMPT00069	Drosophila_melanogaster	ABACACV
Smp_067420	RNCMPT00232	Schistosoma_mansoni	DWWUUUU
SNF	RNCMPT00145	Drosophila_melanogaster	UWGCAC
SNRNP70	RNCMPT00070	Homo_sapiens	RWUCAAG
SNRNP70K	RNCMPT00143	Drosophila_melanogaster	AUCAHG
SNRPA	RNCMPT00071	Homo_sapiens	WUGCACR
SRP54	RNCMPT00272	Drosophila_melanogaster	KKRGG
SRSF2	RNCMPT00072	Homo_sapiens	GGAGWD
SRSF7	RNCMPT00073	Homo_sapiens	ACGACG
SRSF9	RNCMPT00067	Homo_sapiens	KGRWGSM
SRSF9	RNCMPT00074	Homo_sapiens	AKGAVMR
STAR-PAP	RNCMPT00075	Homo_sapiens	MRAUACU
SUP-12	RNCMPT00179	Caenorhabditis_elegans	WGUGUGD
SUP-26	RNCMPT00182	Caenorhabditis_elegans	AUAUWWR
SXL	RNCMPT00119	Drosophila_melanogaster	UUUUUUU
TARDBP	RNCMPT00076	Homo_sapiens	GAAUGD
Tb_0251	RNCMPT00251	Trypanosoma_brucei	HUUCACR

Tb_0252	RNCMPT00252	Trypanosoma_brucei	WGUAGRW
Tb_0230	RNCMPT00230	Trypanosoma_brucei	GAAGGD
Tb_0216	RNCMPT00216	Trypanosoma_brucei	CAUWGUD
Tb_0217	RNCMPT00217	Trypanosoma_brucei	CUKUUKY
Tb_0218	RNCMPT00218	Trypanosoma_brucei	DUUAUH
Tb_0219	RNCMPT00219	Trypanosoma_brucei	UAUACU
Tb_0220	RNCMPT00220	Trypanosoma_brucei	CUUUCU
Tb_0253	RNCMPT00253	Trypanosoma_brucei	ARAAANA
Tp_0225	RNCMPT00225	Thalassiosira_pseudonana	HACRCGC
TIA1	RNCMPT00077	Homo_sapiens	UUUUUBK
TIA1	RNCMPT00165	Homo_sapiens	WUUUUUB
TIAR-1	RNCMPT00256	Caenorhabditis_elegans	UUUUUU
TIAR-3	RNCMPT00005	Caenorhabditis_elegans	HUUUUUU
TRA2	RNCMPT00078	Drosophila_melanogaster	VAAGAA
Tv_0257	RNCMPT00257	Trichomonas_vaginalis	ADAAAAR
Tv_0258	RNCMPT00258	Trichomonas_vaginalis	UKUUUGD
Tv_0259	RNCMPT00259	Trichomonas_vaginalis	AYCAUGD
Tv_0226	RNCMPT00226	Trichomonas_vaginalis	CAAUAA
Tv_0236	RNCMPT00236	Trichomonas_vaginalis	YUUUUUK
U2AF2	RNCMPT00079	Homo_sapiens	UUUUUYC
U2AF50	RNCMPT00080	Drosophila_melanogaster	UUUUUYY
UNC-75	RNCMPT00081	Caenorhabditis_elegans	UGUUGUD
Vts1p	RNCMPT00111	Saccharomyces_cerevisiae	GCUGGCS
Vts1p	RNCMPT00082	Saccharomyces_cerevisiae	GCUGGYS
YBX1	RNCMPT00116	Homo_sapiens	AACAUCD
YBX1	RNCMPT00083	Homo_sapiens	AACAUC
YBX2	RNCMPT00084	Homo_sapiens	AACAWCD
ZC3H10	RNCMPT00085	Homo_sapiens	SSAGCGM
ZC3H14	RNCMPT00086	Homo_sapiens	UUUDUUU
ZCRB1	RNCMPT00087	Homo_sapiens	GRHUUAA
ZNF638	RNCMPT00164	Homo_sapiens	BGUUSKU

List of RNA Bind-n-Seq experiments on ENCODE

Encode Accession	Target of assay / Gene Name
ENCSR934TDK	A1CF
ENCSR110GHL	AKAP8L
ENCSR472KKU	APOBEC3C
ENCSR497LIF	BOLL
ENCSR992NHR	CELF1
ENCSR806UCE	CNOT4
ENCSR084YCO	CPEB1
ENCSR449VKY	DAZ3
ENCSR005ZRL	DAZAP1
ENCSR488AUU	EIF3D
ENCSR600HIW	EIF4G2
ENCSR171TTH	ELAVL4
ENCSR082AKW	ESRP1
ENCSR063HQO	EWSR1
ENCSR843QMF	FUBP1
ENCSR697VZN	FUBP3
ENCSR936LOF	FUS
ENCSR170PBM	HNRNPA0
ENCSR890PDQ	HNRNPA2B1
ENCSR569UIU	HNRNPC
ENCSR915CDT	HNRNPCL1
ENCSR175OMA	HNRNPD
ENCSR055HDN	HNRNPDL
ENCSR376SUZ	HNRNPF
ENCSR328PGZ	HNRNPH2
ENCSR368NMO	HNRNPK
ENCSR954TYO	HNRNPL
ENCSR928XOW	IGF2BP1
ENCSR588GYZ	IGF2BP2
ENCSR164XGH	IGF2BP3
ENCSR906EKN	ILF2
ENCSR575QYE	KHDRBS2
ENCSR583NVI	KHDRBS3

ENCSR915BDY	KHSRP
ENCSR369RLA	LIN28B
ENCSR006QKZ	MBNL1
ENCSR329RIP	MSI1
ENCSR834CED	NOVA1
ENCSR387CDD	NSUN2
ENCSR102MQN	NUPL2
ENCSR051WAN	PABPC3
ENCSR334QCK	PABPN1L
ENCSR539RTM	PCBP1
ENCSR673FLQ	PCBP2
ENCSR769AEI	PCBP4
ENCSR297UTH	PPP1R10
ENCSR191PTZ	PRR3
ENCSR741ZPT	PTBP3
ENCSR773QCC	PUF60
ENCSR845GNW	PUM1
ENCSR229VBP	RALYL
ENCSR441HLP	RBFOX2
ENCSR421UDF	RBFOX3
ENCSR655NWZ	RBM15B
ENCSR446UHZ	RBM20
ENCSR006TPX	RBM22
ENCSR525PNM	RBM23
ENCSR742AEU	RBM24
ENCSR759QKO	RBM25
ENCSR379HWF	RBM3
ENCSR331BKR	RBM4
ENCSR637HFY	RBM41
ENCSR626INQ	RBM45
ENCSR264RVK	RBM47
ENCSR905BJK	RBM4B
ENCSR345PWR	RBM5
ENCSR548RVM	RBM6
ENCSR492CFG	RBMS2
ENCSR224KSF	RBMS3

ENCSR728SXZ	RC3H1
ENCSR558RBK	SAFB2
ENCSR318HZC	SF1
ENCSR079FDB	SF3B6
ENCSR951YCV	SFPQ
ENCSR167ZZB	SNRPA
ENCSR606JGJ	SNRPB2
ENCSR744POX	SRSF10
ENCSR073DSH	SRSF11
ENCSR275JFN	SRSF2
ENCSR252RIJ	SRSF4
ENCSR914PGB	SRSF5
ENCSR929OLV	SRSF8
ENCSR724HZI	SRSF9
ENCSR474NYR	SUCLG1
ENCSR827QYL	TAF15
ENCSR466JPT	TARDBP
ENCSR456IMV	TDRD10
ENCSR064NOY	TIA1
ENCSR741VUK	TRA2A
ENCSR391FEW	TRA2B
ENCSR419XDN	TRNAU1AP
ENCSR653ZTY	TROVE2
ENCSR497VCL	UNK
ENCSR189MAB	XRCC6
ENCSR605EEO	ZC3H10
ENCSR614KXG	ZC3H18
ENCSR205HMN	ZCRB1
ENCSR315VQD	ZFP36
ENCSR570AIV	ZFP36L1
ENCSR249GVR	ZFP36L2
ENCSR335JQK	ZNF326
ENCSR927QJQ	ZRANB2

References

Agraz-Doblas, A., Bueno, C., Bashford-Rogers, R., Roy, A., Schneider, P., Bardini, M., Ballerini, P., Cazzaniga, G., Moreno, T., Revilla, C., et al. (2019). Unraveling the cellular origin and clinical prognostic markers of infant B-cell acute lymphoblastic leukemia using genome-wide analysis. Haematologica 104, 1176–1188.

Anantharaman, V., Koonin, E.V., and Aravind, L. (2002). Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res. 30, 1427–1464.

Andersson, A.K., for The St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project, Ma, J., Wang, J., Chen, X., Gedman, A.L., Dang, J., Nakitandwe, J., Holmfeldt, L., Parker, M., et al. (2015). The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. Nature Genetics 47, 330–337.

Änkö, M.-L. (2014). Regulation of gene expression programmes by serine–arginine rich splicing factors. Seminars in Cell & Developmental Biology 32, 11–21.

Änkö, M.-L., and Neugebauer, K.M. (2012). RNA–protein interactions *in vivo*: global gets specific. Trends Biochem. Sci. 37, 255–262.

Änkö, M.-L., Müller-McNicoll, M., Brandl, H., Curk, T., Gorup, C., Henry, I., Ule, J., and Neugebauer, K.M. (2012). The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. Genome Biology 13, R17.

Auyeung, V.C., Ulitsky, I., McGeary, S.E., and Bartel, D.P. (2013). Beyond Secondary Structure: Primary-Sequence Determinants License Pri-miRNA Hairpins for Processing. Cell 152, 844–858.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37, W202–W208.

Bao, S., Wu, Q., McLendon, R.E., Hao, Y., Shi, Q., Hjelmeland, A.B., Dewhirst, M.W., Bigner, D.D., and Rich, J.N. (2006). Glioma stem cells promote radioresistance by preferential activation of the DNA damage response. Nature 444, 756–760.

Baralle, F.E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. Nat. Rev. Mol. Cell Biol. 18, 437–451.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., et al. (2012). The evolutionary landscape

of alternative splicing in vertebrate species. Science 338, 1587–1593.

Bardini, M., Woll, P.S., Corral, L., Luc, S., Wittmann, L., Ma, Z., Lo Nigro, L., Basso, G., Biondi, A., Cazzaniga, G., et al. (2015). Clonal variegation and dynamic competition of leukemia-initiating cells in infant acute lymphoblastic leukemia with MLL rearrangement. Leukemia 29, 38–50.

Barrett, N.A., Malouf, C., Kapeni, C., Bacon, W.A., Giotopoulos, G., Jacobsen, S.E.W., Huntly, B.J., and Ottersbach, K. (2016). Mll-AF4 Confers Enhanced Self-Renewal and Lymphoid Potential during a Restricted Window in Development. Cell Rep. 16, 1039–1054.

Boer, J. de, de Boer, J., Williams, A., Skavdis, G., Harker, N., Coles, M., Tolaini, M., Norton, T., Williams, K., Roderick, K., et al. (2003). Transgenic mice with hematopoietic and lymphoid specific expression of Cre. European Journal of Immunology 33, 314–325.

Bolognesi, B., and Lehner, B. (2018). Protein overexpression: reaching the limit. Elife 7, e39804.

Brogna, S., and Wen, J. (2009). Nonsense-mediated mRNA decay (NMD) mechanisms. Nature Structural & Molecular Biology 16, 107–113.

Brundage, M.E., Tandon, P., Eaves, D.W., Williams, J.P., Miller, S.J., Hennigan, R.H., Jegga, A., Cripe, T.P., and Ratner, N. (2014). MAF mediates crosstalk between Ras-MAPK and mTOR signaling in NF1. Oncogene 33, 5626–5636.

Bursen, A., Schwabe, K., Rüster, B., Henschler, R., Ruthardt, M., Dingermann, T., and Marschalek, R. (2010). The AF4·MLL fusion protein is capable of inducing ALL in mice without requirement of MLL·AF4. Blood 115, 3570–3579.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., et al. (2009a). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin. Chem. 55, 611–622.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., et al. (2009b). The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. Clinical Chemistry 55, 611–622.

Cáceres, J.F., Misteli, T., Screaton, G.R., Spector, D.L., and Krainer, A.R. (1997). Role of the Modular Domains of SR Proteins in Subnuclear Localization and Alternative Splicing Specificity. Journal of Cell Biology 138, 225–238.

Caceres, J.F., Screaton, G.R., and Krainer, A.R. (1998). A specific subset of SR proteins shuttles continuously between the nucleus and the cytoplasm. Genes & Development 12, 55–66.

Calarco, J.A., Xing, Y., Cáceres, M., Calarco, J.P., Xiao, X., Pan, Q., Lee, C., Preuss, T.M., and Blencowe, B.J. (2007). Global analysis of alternative splicing differences between humans and chimpanzees. Genes Dev. 21, 2963–2975.

Cazalla, D., Zhu, J., Manche, L., Huber, E., Krainer, A.R., and Cáceres, J.F. (2002). Nuclear Export and Retention Signals in the RS Domain of SR Proteins. Molecular and Cellular Biology 22, 6871–6882.

Chandra, P., Luthra, R., Zuo, Z., Yao, H., Ravandi, F., Reddy, N., Garcia-Manero, G., Kantarjian, H., and Jones, D. (2010). Acute Myeloid Leukemia With t (9; 11)(p21--22; q23) Common Properties of Dysregulated Ras Pathway Signaling and Genomic Progression Characterize De Novo and Therapy-Related Cases. Am. J. Clin. Pathol. 133, 686–693.

Charif, D., and Lobry, J.R. (2007). SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In Structural Approaches to Sequence Evolution: Molecules, Networks, Populations, U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 207–232.

Chen, W., Li, Q., Hudson, W.A., Kumar, A., Kirchhof, N., and Kersey, J.H. (2006). A murine Mll-AF4 knock-in model results in lymphoid and myeloid deregulation and hematologic malignancy. Blood 108, 669–677.

Choudhury, N.R., and Michlewski, G. (2012). Terminal loop-mediated control of microRNA biogenesis. Biochemical Society Transactions 40, 789–793.

Chu, S., McDonald, T., Lin, A., Chakraborty, S., Huang, Q., Snyder, D.S., and Bhatia, R. (2011). Persistence of leukemia stem cells in chronic myelogenous leukemia patients in prolonged remission with imatinib treatment. Blood 118, 5565–5572.

Collins, D.W., and Jukes, T.H. (1994). Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics 20, 386–396.

Contreras, J.R., Palanichamy, J.K., Tran, T.M., Fernando, T.R., Rodriguez-Malave, N.I., Goswami, N., Arboleda, V.A., Casero, D., and Rao, D.S. (2015). MicroRNA-146a modulates B-cell oncogenesis by regulating Egr1. Oncotarget 6, 11023–11037.

Cook, K.B., Vembu, S., Ha, K.C.H., Zheng, H., Laverty, K.U., Hughes, T.R., Ray, D., and Morris, Q.D. (2017). RNAcompete-S: Combined RNA sequence/structure preferences for RNA binding proteins derived from a single-step *in vitro* selection. Methods

126, 18–28.

Corley, M., Burns, M.C., and Yeo, G.W. (2020). How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. Mol. Cell 78, 9–29.

Crick, F. (1970). Central dogma of molecular biology. Nature 227, 561–563.

Croker, B.A., Metcalf, D., Robb, L., Wei, W., Mifsud, S., DiRago, L., Cluse, L.A., Sutherland, K.D., Hartley, L., Williams, E., et al. (2004). SOCS3 is a critical physiological negative regulator of G-CSF signaling and emergency granulopoiesis. Immunity 20, 153–165.

Danan, C., Manickavel, S., and Hafner, M. (2016). PAR-CLIP: A Method for Transcriptome-Wide Identification of RNA Binding Protein Interaction Sites. In Post-Transcriptional Gene Regulation, E. Dassi, ed. (New York, NY: Springer New York), pp. 153–173.

Dargyte, M., Philipp, J., Palka, C.D., Stone, M.D., and Sanford, J.R. (2020). Splicing factor SRSF1 expands the regulatory logic of microRNA expression.

Das, S., and Krainer, A.R. (2014). Emerging Functions of SRSF1, Splicing Factor and Oncoprotein, in RNA Metabolism and Cancer. Molecular Cancer Research 12, 1195–1204.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Research 46, D794–D801.

Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C., and Huang, T.H.-M. (2008). The functional consequences of alternative promoter use in mammalian genomes. Trends Genet. 24, 167–177.

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.

Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. Nat. Biotechnol. 34, 518–524.

Deleault, K.M., Skinner, S.J., and Brooks, S.A. (2008). Tristetraprolin regulates TNF TNF-alpha mRNA stability via a proteasome dependent mechanism involving the combined action of the ERK and p38 pathways. Mol. Immunol. 45, 13–24.

Dezső, Z., Nikolsky, Y., Sviridov, E., Shi, W., Serebriyskaya, T., Dosymbekov, D., Bugrim, A., Rakhmatulin, E., Brennan, R.J., Guryanov, A., et al. (2008). A comprehensive functional analysis of tissue specificity of human gene expression. BMC Biology 6.

(dgt), T.F.C.A.T.R.P.A.C., and The FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014). A promoter-level mammalian expression atlas. Nature 507, 462–470.

Diehn, M., Cho, R.W., Lobo, N.A., Kalisky, T., Dorie, M.J., Kulp, A.N., Qian, D., Lam, J.S., Ailles, L.E., Wong, M., et al. (2009). Association of reactive oxygen species levels and radioresistance in cancer stem cells. Nature 458, 780–783.

Di Giammartino, D.C., Nishida, K., and Manley, J.L. (2011). Mechanisms and consequences of alternative polyadenylation. Mol. Cell 43, 853–866.

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

Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N.J., Van Nostrand, E.L., Pratt, G.A., et al. (2018). Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. Mol. Cell 70, 854–867.e9.

Downward, J. (2003). Targeting RAS signalling pathways in cancer therapy. Nat. Rev. Cancer 3, 11–22.

Du, P., Wang, L., Sliz, P., and Gregory, R.I. (2015). A Biogenesis Step Upstream of Microprocessor Controls miR-17 92 Expression. Cell 162, 885–899.

Dvinge, H., and Bradley, R.K. (2015). Widespread intron retention diversifies most cancer transcriptomes. Genome Medicine 7.

Eickhardt, E., Als, T.D., Grove, J., Boerglum, A.D., and Lescai, F. (2016). Estimating the functional impact of INDELs in transcription factor binding sites: a genome-wide landscape.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138.

Emerenciano, M., da C. Barbosa, T., de Almeida Lopes, B., Meyer, C., Marschalek, R., and Pombo-de-Oliveira, M.S. (2015). Subclonality and prenatal origin ofRAS-mutations inKMT2A (MLL)-rearranged infant acute lymphoblastic leukaemia. British Journal of Haematology 170, 268–271.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Ennajdaoui, H., Howard, J.M., Sterne-Weiler, T., Jahanbani, F., Coyne, D.J., Uren, P.J., Dargyte, M., Katzman, S., Draper, J.M., Wallace, A., et al. (2016). IGF2BP3 Modulates the Interaction of Invasion-Associated Transcripts with RISC. Cell Rep. 15, 1876–1883.

Ernst, P., Fisher, J.K., Avery, W., Wade, S., Foy, D., and Korsmeyer, S.J. (2004a). Definitive Hematopoiesis Requires the Mixed-Lineage Leukemia Gene. Dev. Cell 6, 437–443.

Ernst, P., Mabon, M., Davidson, A.J., Zon, L.I., and Korsmeyer, S.J. (2004b). An Mll-Dependent Hox Program Drives Hematopoietic Progenitor Expansion. Current Biology 14, 2063–2069.

Eychène, A., Rocques, N., and Pouponnot, C. (2008). A new MAFia in cancer. Nature Reviews Cancer 8, 683–693.

Faber, J., Krivtsov, A.V., Stubbs, M.C., Wright, R., Davis, T.N., van den Heuvel-Eibrink, M., Zwaan, C.M., Kung, A.L., and Armstrong, S.A. (2009). HOXA9 is required for survival in human MLL-rearranged acute leukemias. Blood 113, 2375–2385.

Fernando, T.R., Contreras, J.R., Zampini, M., Rodriguez-Malave, N.I., Alberti, M.O., Anguiano, J., Tran, T.M., Palanichamy, J.K., Gajeton, J., Ung, N.M., et al. (2017). The lncRNA CASC15 regulates SOX4 expression in RUNX1-rearranged acute leukemia. Mol. Cancer 16, 126.

Fiddes, I.T., Armstrong, J., Diekhans, M., Nachtweide, S., Kronenberg, Z.N., Underwood, J.G., Gordon, D., Earl, D., Keane, T., Eichler, E.E., et al. (2018). Comparative Annotation Toolkit (CAT)—simultaneous clade and personal genome annotation. Genome Res. 28, 1029–1038.

Field, A.R., Jacobs, F.M.J., Fiddes, I.T., Phillips, A.P.R., Reyes-Ortiz, A.M., LaMontagne, E., Whitehead, L., Meng, V., Rosenkrantz, J.L., Haeussler, M., et al. (2017). Structurally conserved primate lncRNAs are transiently expressed during human cortical differentiation and influence cell type specific genes.

Floor, S.N., and Doudna, J.A. (2016). Tunable protein synthesis by transcript isoforms in human cells. eLife 5.

Foissac, S., and Sammeth, M. (2007). ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. Nucleic Acids Res. 35, W297–W299.

Fried, M.G. (1989). Measurement of protein-DNA interaction parameters by electrophoresis mobility shift assay. Electrophoresis 10, 366–376.

Fried, M., and Crothers, D.M. (1981). Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. Nucleic Acids Res. 9, 6505–6525.

Fried, M.G., and Garner, M.M. (1998). Molecular Biology Methods and Applications.

Friedman, R.C., -H. Farh, K.K., Burge, C.B., and Bartel, D.P. (2008). Most mammalian mRNAs are conserved targets of microRNAs. Genome Research 19, 92–105.

Fu, Y., Dominissini, D., Rechavi, G., and He, C. (2014). Gene expression regulation mediated through reversible m6A RNA methylation. Nat. Rev. Genet. 15, 293–306.

Galante, P.A.F., Sandhu, D., de Sousa Abreu, R., Gradassi, M., Slager, N., Vogel, C., de Souza, S.J., and Penalva, L.O.F. (2009). A comprehensive *in silico* expression analysis of RNA binding proteins in normal and tumor tissue: Identification of potential players in tumor formation. RNA Biol. 6, 426–433.

García-Mauriño, S.M., Rivero-Rodríguez, F., Velázquez-Cruz, A., Hernández-Vellisca, M., Díaz-Quintana, A., De la Rosa, M.A., and Díaz-Moreno, I. (2017). RNA Binding Protein Regulation and Cross-Talk in the Control of AU-rich mRNA Fate. Front Mol Biosci 4, 71.

Garner, M.M., and Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic Acids Res. 9, 3047–3060.

Gebauer, F., Preiss, T., and Hentze, M.W. (2012). From cis-regulatory elements to complex RNPs and back. Cold Spring Harb. Perspect. Biol. 4, a012245.

Gehring, N.H., Kunz, J.B., Neu-Yilik, G., Breit, S., Viegas, M.H., Hentze, M.W., and Kulozik, A.E. (2005). Exon-junction complex components specify distinct routes of nonsense-mediated mRNA decay with differential cofactor requirements. Mol. Cell 20, 65–75.

Gehring, N.H., Wahle, E., and Fischer, U. (2017). Deciphering the mRNP Code: RNA-Bound Determinants of Post-Transcriptional Gene Regulation. Trends Biochem. Sci. 42, 369–382.

Georgiades, P., Ogilvy, S., Duval, H., Licence, D.R., Charnock-Jones, D.S., Smith, S.K., and Print, C.G. (2002). VavCre transgenic mice: a tool for mutagenesis in hematopoietic and endothelial lineages. Genesis 34, 251–256.

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nat. Rev. Genet. 15, 829–845.

Ghosh, S., Marchand, V., Gáspár, I., and Ephrussi, A. (2012). Control of RNP motility and localization by a splicing-dependent structure in oskar mRNA. Nat. Struct. Mol. Biol. 19, 441–449.

Glanzer, J., Miyashiro, K.Y., Sul, J.-Y., Barrett, L., Belt, B., Haydon, P., and Eberwine, J. (2005). RNA splicing capability of live neuronal dendrites. Proc. Natl. Acad. Sci. U. S. A. 102, 16859–16864.

Graveley, B.R., and Maniatis, T. (1998). Arginine/Serine-Rich Domains of SR Proteins Can Function as Activators of Pre-mRNA Splicing. Molecular Cell 1, 765–771.

Guil, S., and Cáceres, J.F. (2007). The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. Nature Structural & Molecular Biology 14, 591–596.

Hachet, O., and Ephrussi, A. (2004). Splicing of oskar RNA in the nucleus is coupled to its cytoplasmic localization. Nature 428, 959–963.

Haddox, C.L., Mangaonkar, A.A., Chen, D., Shi, M., He, R., Oliveira, J.L., Litzow, M.R., Al-Kali, A., Hogan, W.J., and Elliott, M.A. (2017). Blinatumomab-induced lineage switch of B-ALL with t(4:11)(q21;q23) KMT2A/AFF1 into an aggressive AML: pre- and post-switch phenotypic, cytogenetic and molecular analysis. Blood Cancer Journal 7, e607–e607.

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.

Hall, T.M.T. (2005). Multiple modes of RNA recognition by zinc finger proteins. Curr. Opin. Struct. Biol. 15, 367–373.

Hardy, R.R., and Hayakawa, K. (2001). B cell development pathways. Annu. Rev. Immunol. 19, 595–621.

He, M., Chen, P., Arnovitz, S., Li, Y., Huang, H., Neilly, M.B., Wei, M., Rowley, J.D., Chen, J., and Li, Z. (2012). Two isoforms of HOXA9 function differently but work synergistically in human MLL-rearranged leukemia. Blood Cells Mol. Dis. 49, 102–106.

Heffner, C.S., Herbert Pratt, C., Babiuk, R.P., Sharma, Y., Rockwood, S.F., Donahue, L.R., Eppig, J.T., and Murray, S.A. (2012). Supporting conditional mouse mutagenesis with a comprehensive cre characterization resource. Nat. Commun. 3, 1218.

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

Hellman, L.M., and Fried, M.G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nat. Protoc. 2, 1849–1861.

Heo, I., Joo, C., Kim, Y.-K., Ha, M., Yoon, M.-J., Cho, J., Yeom, K.-H., Han, J., and Narry Kim, V. (2009). TUT4 in Concert with Lin28 Suppresses MicroRNA Biogenesis through Pre-MicroRNA Uridylation. Cell 138, 696–708.

Herdy, B., Mayer, C., Varshney, D., Marsico, G., Murat, P., Taylor, C., D'Santos, C., Tannahill, D., and Balasubramanian, S. (2018). Analysis of NRAS RNA G-quadruplex binding proteins reveals DDX3X as a novel interactor of cellular G-quadruplex containing transcripts. Nucleic Acids Res. 46, 11592–11604.

Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J.Y., Yehia, G., and Tian, B. (2012). Analysis of alternative cleavage and polyadenylation by 3 region extraction and deep sequencing. Nat. Methods 10, 133–139.

Howard, J.M., and Sanford, J.R. (2015). The RNAissance family: SR proteins as multifaceted regulators of gene expression. Wiley Interdisciplinary Reviews: RNA 6, 93–110.

Howard, J.M., Lin, H., Wallace, A.J., Kim, G., Draper, J.M., Haeussler, M., Katzman, S., Toloue, M., Liu, Y., and Sanford, J.R. (2018). HNRNPA1 promotes recognition of splice site decoys by U2AF2 *in vivo*. Genome Research 28, 689–698.

Huang, Z.-L., Dai, J., Luo, W.-H., Wang, X.-G., Tan, J.-H., Chen, S.-B., and Huang, Z.-S. (2018). Identification of G-Quadruplex-Binding Protein from the Exploration of RGG Motif/G-Quadruplex Interactions. J. Am. Chem. Soc. 140, 17945–17955.

Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y., Tajnik, M., König, J., and Ule, J. (2014). iCLIP: Protein–RNA interactions at nucleotide resolution. Methods 65, 274–287.

Hyrenius-Wittsten, A., Pilheden, M., Sturesson, H., Hansson, J., Walsh, M.P., Song, G., Kazi, J.U., Liu, J., Ramakrishan, R., Garcia-Ruiz, C., et al. (2018). De novo activating mutations drive clonal evolution and enhance clonal fitness in KMT2A-rearranged leukemia. Nat. Commun. 9, 1770.

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.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. Science 324, 218–223.

Jain, M., Olsen, H.E., Paten, B., and Akeson, M. (2016). Erratum to: The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 17, 256.

Janardhan, H.P., Milstone, Z.J., Shin, M., Lawson, N.D., Keaney, J.F., Jr, and Trivedi, C.M. (2017). Hdac3 regulates lymphovenous and lymphatic valve formation. J. Clin. Invest. 127, 4193–4206.

Jarmoskaite, I., and Russell, R. (2011). DEAD-box proteins as RNA helicases and chaperones. Wiley Interdisciplinary Reviews: RNA 2, 135–152.

Jønson, L., Christiansen, J., Hansen, T.V.O., Vikeså, J., Yamamoto, Y., and Nielsen, F.C. (2014). IMP3 RNP safe houses prevent miRNA-directed HMGA2 mRNA decay in cancer and development. Cell Rep. 7, 539–551.

Joseph, C., Quach, J.M., Walkley, C.R., Lane, S.W., Lo Celso, C., and Purton, L.E. (2013). Deciphering hematopoietic stem cells in their niches: a critical appraisal of genetic models, lineage tracing, and imaging strategies. Cell Stem Cell 13, 520–533.

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.

Jung, H., Lee, D., Lee, J., Park, D., Kim, Y.J., Park, W.-Y., Hong, D., Park, P.J., and Lee, E. (2015). Intron retention is a widespread mechanism of tumor-suppressor inactivation. Nat. Genet. 47, 1242–1248.

Kahles, A., Ong, C.S., Zhong, Y., and Rätsch, G. (2016). SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. Bioinformatics 32, 1840–1847.

Kalnina, Z., Zayakin, P., Silina, K., and Linē, A. (2005). Alterations of pre-mRNA splicing in cancer. Genes Chromosomes Cancer 42, 342–357.

Kampen, K.R., ter Elst, A., Mahmud, H., Scherpen, F.J.G., Diks, S.H., Peppelenbosch, M.P., de Haas, V., Guryev, V., and de Bont, E.S.J.M. (2014). Insights in dynamic kinome reprogramming as a consequence of MEK inhibition in MLL-rearranged AML. Leukemia 28, 589–599.

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

Keene, J.D. (2007). RNA regulons: coordination of post-transcriptional events. Nat. Rev. Genet. 8, 533–543.

Keene, J.D., and Lager, P.J. (2005). Post-transcriptional operons and regulons co-ordinating gene expression. Chromosome Res. 13, 327–337.

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

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, a. D. (2002). The Human Genome Browser at UCSC. Genome Research 12, 996–1006.

Kerstjens, M., Driessen, E.M.C., Willekes, M., Pinhanços, S.S., Schneider, P., Pieters, R., and Stam, R.W. (2017). MEK inhibition is a promising therapeutic strategy for MLL-rearranged infant acute lymphoblastic leukemia patients carrying RAS mutations. Oncotarget 8, 14835–14846.

Khan, Z., Ford, M.J., Cusanovich, D.A., Mitrano, A., Pritchard, J.K., and Gilad, Y. (2013). Primate transcript and protein expression levels evolve under compensatory selection pressures. Science 342, 1100–1104.

Kim, K., Nguyen, T.D., Li, S., and Nguyen, T.A. (2018). SRSF3 recruits DROSHA to the basal junction of primary microRNAs. RNA 24, 892–898.

Köbel, M., Xu, H., Bourne, P.A., Spaulding, B.O., Shih, I.-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.

Koch, L. (2014). New CLIP pipeline improves interactome discovery. Nat. Rev. Genet. 16, 2–2.

Komatsu, K.R., Taya, T., Matsumoto, S., Miyashita, E., Kashida, S., and Saito, H. (2020). RNA structure-wide discovery of functional interactions with multiplexed RNA motif library. Nat. Commun. 11, 6275.

Kooshapur, H., Choudhury, N.R., Simon, B., Mühlbauer, M., Jussupow, A., Fernandez, N., Jones, A.N., Dallmann, A., Gabel, F., Camilloni, C., et al. (2018). Structural basis for terminal loop recognition and stimulation of pri-miRNA-18a processing by

hnRNP A1. Nature Communications 9.

Kosti, I., Jain, N., Aran, D., Butte, A.J., and Sirota, M. (2016). Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. Sci. Rep. 6, 24799.

Krainer, A.R., Conway, G.C., and Kozak, D. (1990). Purification and characterization of pre-mRNA splicing factor SF2 from HeLa cells. Genes & Development 4, 1158–1171.

Krainer, A.R., Mayeda, A., Kozak, D., and Binns, G. (1991). Functional expression of cloned human splicing factor SF2: homology to rna-binding proteins, U1 70K, and drosophila splicing regulators. Cell 66, 383–394.

Krivtsov, A.V., and Armstrong, S.A. (2007). MLL translocations, histone modifications and leukaemia stem-cell development. Nat. Rev. Cancer 7, 823–833.

Krivtsov, A.V., Twomey, D., Feng, Z., Stubbs, M.C., Wang, Y., Faber, J., Levine, J.E., Wang, J., Hahn, W.C., Gary Gilliland, D., et al. (2006). Transformation from committed progenitor to leukaemia stem cell initiated by MLL–AF9. Nature 442, 818–822.

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.

Kroczynska, B., Kaur, S., Katsoulidis, E., Majchrzak-Kita, B., Sassano, A., Kozma, S.C., Fish, E.N., and Platanias, L.C. (2009). Interferon-Dependent Engagement of Eukaryotic Initiation Factor 4B via S6 Kinase (S6K)- and Ribosomal Protein S6K-Mediated Signals. Molecular and Cellular Biology 29, 2865–2875.

Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A., and Burge, C.B. (2014). RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Mol. Cell 54, 887–900.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

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.

Lareau, L.F., and Brenner, S.E. (2015). Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. Mol. Biol. Evol. 32, 1072–1079.

Lavallée, V.-P., Baccelli, I., Krosl, J., Wilhelm, B., Barabé, F., Gendron, P., Boucher,

G., Lemieux, S., Marinier, A., Meloche, S., et al. (2015). The transcriptomic landscape and directed chemical interrogation of MLL-rearranged acute myeloid leukemias. Nat. Genet. 47, 1030–1037.

Lawrence, H.J., Christensen, J., Fong, S., Hu, Y.-L., Weissman, I., Sauvageau, G., Humphries, R.K., and Largman, C. (2005). Loss of expression of the Hoxa-9 homeobox gene impairs the proliferation and repopulating ability of hematopoietic stem cells. Blood 106, 3988–3994.

Lechman, E.R., Gentner, B., Ng, S.W.K., Schoof, E.M., van Galen, P., Kennedy, J.A., Nucera, S., Ciceri, F., Kaufmann, K.B., Takayama, N., et al. (2016). miR-126 Regulates Distinct Self-Renewal Outcomes in Normal and Malignant Hematopoietic Stem Cells. Cancer Cell 29, 602–606.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., et al. (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature 425, 415–419.

Lewis, A., Park, C.S., Puppi, M., and Lacorazza, H.D. (2019). KLF4 controls leukemic stem cell self-renewal in MLL-AF9-induced acute myeloid leukemia. Blood 134, 1231–1231.

Lewis, B.P., Green, R.E., and Brenner, S.E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl. Acad. Sci. U. S. A. 100, 189–192.

Li, X., and Manley, J.L. (2005). Inactivation of the SR Protein Splicing Factor ASF/SF2 Results in Genomic Instability. Cell 122, 365–378.

Li, J.J., Bickel, P.J., and Biggin, M.D. (2014a). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ 2, e270.

Li, X., Song, J., and Yi, C. (2014b). Genome-wide Mapping of Cellular Protein–RNA Interactions Enabled by Chemical Crosslinking. Genomics Proteomics Bioinformatics 12, 72–78.

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.

Lin, C., Smith, E.R., Takahashi, H., Lai, K.C., Martin-Brown, S., Florens, L., Washburn, M.P., Conaway, J.W., Conaway, R.C., and Shilatifard, A. (2010). AFF4, a component of the ELL/P-TEFb elongation complex and a shared subunit of MLL chimeras, can link transcription elongation to leukemia. Mol. Cell 37, 429–437.

Lin, S., Luo, R.T., Ptasinska, A., Kerry, J., Assi, S.A., Wunderlich, M., Imamura, T., Kaberlein, J.J., Rayes, A., Althoff, M.J., et al. (2016). Instructive Role of MLL-Fusion Proteins Revealed by a Model of t(4;11) Pro-B Acute Lymphoblastic Leukemia. Cancer Cell 30, 737–749.

Lin, S., Luo, R.T., Shrestha, M., Thirman, M.J., and Mulloy, J.C. (2017). The full transforming capacity of MLL-Af4 is interlinked with lymphoid lineage commitment. Blood 130, 903–907.

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. Eur. J. Cancer 48, 3405–3413.

Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. 6, 26.

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. Nat. Struct. Mol. Biol. 20, 1434–1442.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550.

Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nat. Rev. Mol. Cell Biol. 8, 479–490.

Lynch, J.R., Salik, B., Connerty, P., Vick, B., Leung, H., Pijning, A., Jeremias, I., Spiekermann, K., Trahair, T., Liu, T., et al. (2019). JMJD1C-mediated metabolic dysregulation contributes to HOXA9-dependent leukemogenesis. Leukemia 33, 1400–1410.

Ma, L., Teruya-Feldstein, J., and Weinberg, R.A. (2007). Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. Nature 449, 682–688.

Magee, J.A., Piskounova, E., and Morrison, S.J. (2012). Cancer stem cells: impact, heterogeneity, and uncertainty. Cancer Cell 21, 283–296.

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

Maris, C., Dominguez, C., and Allain, F.H.-T. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J. 272, 2118–2131.

Markham, N.R., and Zuker, M. (2005). DINAMelt web server for nucleic acid melting prediction. Nucleic Acids Research 33, W577–W581.

Markham, N.R., and Zuker, M. (2008). UNAFold. In Bioinformatics: Structure, Function and Applications, J.M. Keith, ed. (Totowa, NJ: Humana Press), pp. 3–31.

Maslon, M.M., Heras, S.R., Bellora, N., Eyras, E., and Cáceres, J.F. (2014). The translational landscape of the splicing factor SRSF1 and its role in mitosis. Elife e02028.

Mazin, P.V., Jiang, X., Fu, N., Han, D., Guo, M., Gelfand, M.S., and Khaitovich, P. (2018). Conservation, evolution, and regulation of splicing during prefrontal cortex development in humans, chimpanzees, and macaques. RNA 24, 585–596.

McIlwain, D.R., Pan, Q., Reilly, P.T., Elia, A.J., McCracken, S., Wakeham, A.C., Itie-Youten, A., Blencowe, B.J., and Mak, T.W. (2010). Smg1 is required for embryogenesis and regulates diverse genes via alternative splicing coupled to nonsense-mediated mRNA decay. Proc. Natl. Acad. Sci. U. S. A. 107, 12186–12191.

Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., et al. (2015). The human transcriptome across tissues and individuals. Science 348, 660–665.

Menotti, E., Henderson, B.R., and Kühn, L.C. (1998). Translational regulation of mRNAs with distinct IRE sequences by iron regulatory proteins 1 and 2. J. Biol. Chem. 273, 1821–1824.

Merkin, J., Russell, C., Chen, P., and Burge, C.B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science 338, 1593–1599.

Merlos-Suárez, A., Barriga, F.M., Jung, P., Iglesias, M., Céspedes, M.V., Rossell, D., Sevillano, M., Hernando-Momblona, X., da Silva-Diz, V., Muñoz, P., et al. (2011). The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. Cell Stem Cell 8, 511–524.

Metzler, M., Forster, A., Pannell, R., Arends, M.J., Daser, A., Lobato, M.N., and Rabbitts, T.H. (2006). A conditional model of MLL-AF4 B-cell tumourigenesis using invertor technology. Oncogene 25, 3093–3103.

Meyer, C., Burmeister, T., Gröger, D., Tsaur, G., Fechina, L., Renneville, A., Sutton, R., Venn, N.C., Emerenciano, M., Pombo-de-Oliveira, M.S., et al. (2018). The MLL recombinome of acute leukemias in 2017. Leukemia 32, 273–284.

Michlewski, G., and Cáceres, 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., and Cáceres, J.F. (2019). Post-transcriptional control of miRNA biogenesis. RNA 25, 1–16.

Milne, T.A., Kim, J., Wang, G.G., Stadler, S.C., Basrur, V., Whitcomb, S.J., Wang, Z., Ruthenburg, A.J., Elenitoba-Johnson, K.S.J., Roeder, R.G., et al. (2010). Multiple interactions recruit MLL1 and MLL1 fusion proteins to the HOXA9 locus in leukemogenesis. Mol. Cell 38, 853–863.

Mohan, M., Lin, C., Guest, E., and Shilatifard, A. (2010). Licensed to elongate: a molecular mechanism for MLL-based leukaemogenesis. Nat. Rev. Cancer 10, 721–728.

Montes, R., Ayllón, V., Gutierrez-Aranda, I., Prat, I., Carmen Hernández-Lamas, M., Ponce, L., Bresolin, S., te Kronnie, G., Greaves, M., Bueno, C., et al. (2011). Enforced expression of MLL-AF4 fusion in cord blood CD34 cells enhances the hematopoietic repopulating cell function and clonogenic potential but is not sufficient to initiate leukemia. Blood 117, 4746–4758.

Moorman, A.V., Ensor, H.M., Richards, S.M., Chilton, L., Schwab, C., Kinsey, S.E., Vora, A., Mitchell, C.D., and Harrison, C.J. (2010). Prognostic effect of chromosomal abnormalities in childhood B-cell precursor acute lymphoblastic leukaemia: results from the UK Medical Research Council ALL97/99 randomised trial. Lancet Oncol. 11, 429–438.

Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. 34, 267–273.

Mordstein, C., Savisaar, R., Young, R.S., Bazile, J., Talmane, L., Luft, J., Liss, M., Taylor, M.S., Hurst, L.D., and Kudla, G. (2020). Codon Usage and Splicing Jointly Influence mRNA Localization. Cell Syst 10, 351–362.e8.

Mueller, F., Bommer, M., Lacher, U., Ruhland, C., Stagge, V., Adler, G., Gress, T.M., and Seufferlein, T. (2003). KOC is a novel molecular indicator of malignancy. Br. J. Cancer 88, 699–701.

Mueller-Pillasch, F., Pohl, B., Wilda, M., Lacher, U., Beil, M., Wallrapp, C., Hameister, H., Knöchel, W., Adler, G., and Gress, T.M. (1999). Expression of the highly conserved RNA binding protein KOC in embryogenesis. Mechanisms of Development 88, 95–99.

Mukherjee, N., Wessels, H.-H., Lebedeva, S., Sajek, M., Ghanbari, M., Garzia, A., Munteanu, A., Yusuf, D., Farazi, T., Hoell, J.I., et al. (2019). Deciphering human ribonucleoprotein regulatory networks. Nucleic Acids Res. 47, 570–581.

Mustoe, A.M., Corley, M., Laederach, A., and Weeks, K.M. (2018). Messenger RNA Structure Regulates Translation Initiation: A Mechanism Exploited from Bacteria to Humans. Biochemistry 57, 3537–3539.

Nguyen, L.H., Robinton, D.A., Seligson, M.T., Wu, L., Li, L., Rakheja, D., Comerford, S.A., Ramezani, S., Sun, X., Parikh, M.S., et al. (2014). Lin28b is sufficient to drive liver cancer and necessary for its maintenance in murine models. Cancer Cell 26, 248–261.

Nguyen Quang, N., Perret, G., and Ducongé, F. (2016). Applications of High-Throughput Sequencing for In Vitro Selection and Characterization of Aptamers. Pharmaceuticals 9.

Nilsen, T.W., and Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature 463, 457–463.

Noland, C.L., and Doudna, J.A. (2013). Multiple sensors ensure guide strand selection in human RNAi pathways. RNA 19, 639–648.

Nussbacher, J.K., and Yeo, G.W. (2018). Systematic Discovery of RNA Binding Proteins that Regulate MicroRNA Levels. Molecular Cell 69, 1005–1016.e7.

O'Brien, C.A., Kreso, A., and Jamieson, C.H.M. (2010). Cancer stem cells and self-renewal. Clin. Cancer Res. 16, 3113–3120.

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.

Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2020). Biostrings: Efficient manipulation of biological strings. R package version 2.48. 0.

Palanichamy, J.K., Tran, T.M., Howard, J.M., Contreras, J.R., Fernando, T.R., Sterne-Weiler, T., Katzman, S., Toloue, M., Yan, W., Basso, G., et al. (2016). RNA-binding protein IGF2BP3 targeting of oncogenic transcripts promotes hematopoietic progenitor proliferation. J. Clin. Invest. 126, 1495–1511.

Pan, Q., Saltzman, A.L., Kim, Y.K., Misquitta, C., Shai, O., Maquat, L.E., Frey, B.J., and Blencowe, B.J. (2006). Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes Dev. 20, 153–158.

Park, S.-M., Gönen, M., Vu, L., Minuesa, G., Tivnan, P., Barlowe, T.S., Taggart, J., Lu, Y., Deering, R.P., Hacohen, N., et al. (2015). Musashi2 sustains the mixed-lineage leukemia–driven stem cell regulatory program. Journal of Clinical Investigation 125,

1286–1298.

Parker, J.S., Roe, S.M., and Barford, D. (2006). Molecular mechanism of target RNA transcript recognition by Argonaute-guide complexes. Cold Spring Harb. Symp. Quant. Biol. 71, 45–50.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295.

Pfeiffer, B.D., Truman, J.W., and Rubin, G.M. (2012). Using translational enhancers to increase transgene expression in Drosophila. Proc. Natl. Acad. Sci. U. S. A. 109, 6626–6631.

Pineault, N., Helgason, C.D., Jeffrey Lawrence, H., and Keith Humphries, R. (2002). Differential expression of Hox, Meis1, and Pbx1 genes in primitive cells throughout murine hematopoietic ontogeny. Experimental Hematology 30, 49–57.

Pui, C.-H., Carroll, W.L., Meshinchi, S., and Arceci, R.J. (2011). Biology, Risk Stratification, and Therapy of Pediatric Acute Leukemias: An Update. Journal of Clinical Oncology 29, 551–565.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.

Ramos, A., Grünert, S., Adams, J., Micklem, D.R., Proctor, M.R., Freund, S., Bycroft, M., St Johnston, D., and Varani, G. (2000). RNA recognition by a Staufen double-stranded RNA-binding domain. EMBO J. 19, 997–1009.

Ramsköld, D., Wang, E.T., Burge, C.B., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput. Biol. 5, e1000598.

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., Peña 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. Nat. Biotechnol. 27, 667–670.

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature 499, 172–177.

Ray, D., Ha, K.C.H., Nie, K., Zheng, H., Hughes, T.R., and Morris, Q.D. (2017). RNAcompete methodology and application to determine sequence preferences of unconventional RNA-binding proteins. Methods 118-119, 3–15.

Rayes, A., McMasters, R.L., and O'Brien, M.M. (2016). Lineage Switch in MLL-Rearranged Infant Leukemia Following CD19-Directed Therapy. Pediatric Blood & Cancer 63, 1113–1115.

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.

Riverso, M., Montagnani, V., and Stecca, B. (2017). KLF4 is regulated by RAS/RAF/MEK/ERK signaling through E2F1 and promotes melanoma cell growth. Oncogene 36, 3322–3333.

Royce-Tolland, M.E., Andersen, A.A., Koyfman, H.R., Talbot, D.J., Wutz, A., Tonks, I.D., Kay, G.F., and Panning, B. (2010). The A-repeat links ASF/SF2-dependent Xist RNA processing with random choice during X inactivation. Nature Structural & Molecular Biology 17, 948–954.

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.

Ryter, J.M., and Schultz, S.C. (1998). Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. EMBO J. 17, 7505–7513.

Saltzman, A.L., Kim, Y.K., Pan, Q., Fagnani, M.M., Maquat, L.E., and Blencowe, B.J. (2008). Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. Mol. Cell. Biol. 28, 4320–4330.

Sanchez de Groot, N., Armaos, A., Graña-Montes, R., Alriquet, M., Calloni, G., Vabulas, R.M., and Tartaglia, G.G. (2019). RNA structure drives interaction with proteins. Nat. Commun. 10, 3246.

Sanford, J.R., Gray, N.K., Beckmann, K., and Cáceres, J.F. (2004). A novel role for shuttling SR proteins in mRNA translation. Genes Dev. 18, 755–768.

Sanford, J.R., Ellis, J.D., Cazalla, D., and Cáceres, J.F. (2005). Reversible phosphorylation differentially affects nuclear and cytoplasmic functions of splicing factor 2/alternative splicing factor. Proc. Natl. Acad. Sci. U. S. A. 102, 15042–15047.

Sanford, J.R., Coutinho, P., Hackett, J.A., Wang, X., Ranahan, W., and Caceres,

J.F. (2008). Identification of Nuclear and Cytoplasmic mRNA Targets for the Shuttling Protein SF2/ASF. PLoS ONE 3, e3369.

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

Schaeffer, D.F., Owen, D.R., Lim, H.J., Buczkowski, A.K., Chung, S.W., Scudamore, C.H., Huntsman, D.G., Ng, S.S.W., 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.

Schneider, T., Hung, L.-H., Aziz, M., Wilmen, A., Thaum, S., Wagner, J., Janowski, R., Müller, S., Schreiner, S., Friedhoff, P., et al. (2019). Combinatorial recognition of clustered RNA elements by the multidomain RNA-binding protein IMP3. Nature Communications 10.

Schubbert, S., Shannon, K., and Bollag, G. (2007). Hyperactive Ras in developmental disorders and cancer. Nature Reviews Cancer 7, 295–308.

Schuck, P. (2007). Protein Interactions: Biophysical Approaches for the Study of Complex Reversible Systems (Springer Science & Business Media).

Seemann, S.E., Mirza, A.H., Hansen, C., Bang-Berthelsen, C.H., Garde, C., Christensen-Dalsgaard, M., Torarinsson, E., Yao, Z., Workman, C.T., Pociot, F., et al. (2017). The identification and functional annotation of RNA structures conserved in vertebrates. Genome Res. 27, 1371–1383.

Sharma, S. (2008). Isolation of a sequence-specific RNA binding protein, polypyrimidine tract binding protein, using RNA affinity chromatography. Methods Mol. Biol. 488, 1–8.

Shen, H., and Green, M.R. (2006). RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans. Genes & Development 20, 1755–1765.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., et al. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. U. S. A. 100, 15776–15781.

Siepel, A. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15, 1034–1050.

Siepel, A., and Haussler, D. Phylogenetic Hidden Markov Models. Statistical

Methods in Molecular Evolution 325–351.

Smith, E., Lin, C., and Shilatifard, A. (2011). The super elongation complex (SEC) and MLL in development and disease. Genes & Development 25, 661–672.

Soemedi, R., Cygan, K.J., Rhine, C.L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., and Fairbrother, W.G. (2017). Pathogenic variants that alter protein code often disrupt splicing. Nat. Genet. 49, 848–855.

Sola, M., Menon, A.P., Moreno, B., Meraviglia-Crivelli, D., Soldevilla, M.M., Cartón-García, F., and Pastor, F. (2020). Aptamers Against Live Targets: Is In Vivo SELEX Finally Coming to the Edge? Molecular Therapy - Nucleic Acids 21, 192–204.

Somervaille, T.C.P., and Cleary, M.L. (2006). Identification and characterization of leukemia stem cells in murine MLL-AF9 acute myeloid leukemia. Cancer Cell 10, 257–268.

Somervaille, T.C.P., 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.

Sonenberg, N., Rupprecht, K.M., Hecht, S.M., and Shatkin, A.J. (1979). Eukaryotic mRNA cap binding protein: purification by affinity chromatography on sepharose-coupled m7GDP. Proc. Natl. Acad. Sci. U. S. A. 76, 4345–4349.

de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M., and Vogel, C. (2009). Global signatures of protein and mRNA expression levels. Mol. Biosyst. 5, 1512–1526.

Stadler, C.R., Vegi, N., Mulaw, M.A., Edmaier, K.E., Rawat, V.P.S., Dolnik, A., Bullinger, L., Heilmeier, B., Quintanilla-Fend, L., Spiekermann, K., et al. (2014). The leukemogenicity of Hoxa9 depends on alternative splicing. Leukemia 28, 1838–1843.

Stefl, R., Skrisovska, L., and -T. Allain, F.H. (2005). RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. EMBO Reports 6, 33–38.

Sternburg, E.L., and Karginov, F.V. (2020). Global Approaches in Studying RNA-Binding Protein Interaction Networks. Trends Biochem. Sci. 45, 593–603.

Sterne-Weiler, T., Martinez-Nunez, R.T., Howard, J.M., Cvitovik, I., Katzman, S., Tariq, M.A., Pourmand, N., and Sanford, J.R. (2013). Frac-seq reveals isoform-specific recruitment to polyribosomes. Genome Res. 23, 1615–1623.

St Johnston, D., Brown, N.H., Gall, J.G., and Jantsch, M. (1992). A conserved dou-

ble-stranded RNA-binding domain. Proc. Natl. Acad. Sci. U. S. A. 89, 10979–10983.

Stoltenburg, R., Reinemann, C., and Strehlitz, B. (2007). SELEX--a (r)evolutionary method to generate high-affinity nucleic acid ligands. Biomol. Eng. 24, 381–403.

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, Molecules, and Diseases 46, 321–326.

Strimmer, K. (2008a). fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics 24, 1461–1462.

Strimmer, K. (2008b). A unified approach to false discovery rate estimation. BMC Bioinformatics 9.

Su, C.-H., D, D., and Tarn, W.-Y. (2018). Alternative Splicing in Neurogenesis and Brain Development. Front Mol Biosci 5, 12.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102, 15545–15550.

Tahara, S.M., Morgan, M.A., and Shatkin, A.J. (1981). Two forms of purified m7G-cap binding protein with different effects on capped mRNA translation in extracts of uninfected and poliovirus-infected HeLa cells. J. Biol. Chem. 256, 7691–7694.

Takata, M., Goto, Y., Ichii, N., Yamaura, M., Murata, H., Koga, H., Fujimoto, A., and Saida, T. (2005). Constitutive Activation of the Mitogen-Activated Protein Kinase Signaling Pathway in Acral Melanomas. Journal of Investigative Dermatology 125, 318–322.

Taliaferro, J.M., Vidaki, M., Oliveira, R., Olson, S., Zhan, L., Saxena, T., Wang, E.T., Graveley, B.R., Gertler, F.B., Swanson, M.S., et al. (2016). Distal Alternative Last Exons Localize mRNAs to Neural Projections. Mol. Cell 61, 821–833.

Tamai, H., Miyake, K., Takatori, M., Miyake, N., Yamaguchi, H., Dan, K., Shimada, T., and Inokuchi, K. (2011). Activated K-Ras protein accelerates human MLL/AF4-induced leukemo-lymphomogenicity in a transgenic mouse model. Leukemia 25, 888–891.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics 25.

Treiber, T., Treiber, N., Plessmann, U., Harlander, S., Daiß, J.-L., Eichner, N., Leh-

mann, G., Schall, K., Urlaub, H., and Meister, G. (2017). A Compendium of RNA-Binding Proteins that Regulate MicroRNA Biogenesis. Molecular Cell 66, 270–284.e13.

Trentin, L., Bresolin, S., Giarin, E., Bardini, M., Serafin, V., Accordi, B., Fais, F., Tenca, C., De Lorenzo, P., Valsecchi, M.G., et al. (2016). Deciphering KRAS and NRAS mutated clone dynamics in MLL-AF4 paediatric leukaemia by ultra deep sequencing analysis. Scientific Reports 6.

Trincado, J.L., Entizne, J.C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D.J., and Eyras, E. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. 19, 40.

Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Frank Bennett, C., Sharma, A., Bubulya, P.A., et al. (2010). The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation. Molecular Cell 39, 925–938.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249, 505–510.

Tung, K.-F., Pan, C.-Y., Chen, C.-H., and Lin, W.-C. (2020). Top-ranked expressed gene transcripts of human protein-coding genes investigated with GTEx dataset. Sci. Rep. 10, 16245.

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.F., and Smith, A.D. (2012). Site identification in high-throughput RNA–protein interaction data. Bioinformatics 28, 3013–3020.

Valverde, R., Edwards, L., and Regan, L. (2008). Structure and function of KH domains. FEBS J. 275, 2712–2726.

Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods 13, 508–514.

Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. 13, 227–232.

Vogel, C., Abreu, R. de S., Ko, D., Le, S.-Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R., Marcotte, E.M., and Penalva, L.O. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol. Syst. Biol. 6, 400.

Wagle, M.-C., Kirouac, D., Klijn, C., Liu, B., Mahajan, S., Junttila, M., Moffat, J.,

Merchant, M., Huw, L., Wongchenko, M., et al. (2018). A transcriptional MAPK Pathway Activity Score (MPAS) is a clinically relevant biomarker in multiple cancer types. Npj Precision Oncology 2.

Wang, Z., and Kiledjian, M. (2001). Functional link between the mammalian exosome and mRNA decapping. Cell 107, 751–762.

Wang, E., Lu, S.X., Pastore, A., Chen, X., Imig, J., Lee, S.C.-W., Hockemeyer, K., Ghebrechristos, Y.E., Yoshimi, A., Inoue, D., et al. (2019). Targeting an RNA-Binding Protein Network in Acute Myeloid Leukemia. Cancer Cell 35, 369–384.e7.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nature Reviews Genetics 5, 276–287.

Wei, J., Wunderlich, M., Fox, C., Alvarez, S., Cigudosa, J.C., Wilhelm, J.S., Zheng, Y., Cancelas, J.A., Gu, Y., Jansen, M., et al. (2008). Microenvironment Determines Lineage Fate in a Human Model of MLL-AF9 Leukemia. Cancer Cell 13, 483–495.

Wickham, H., and Others (2009). ggplot2: elegant graphics for data analysis (use R!). Springer, New York, Doi 10, 978–970.

Wilkinson, A.C., Ballabio, E., Geng, H., North, P., Tapia, M., Kerry, J., Biswas, D., Roeder, R.G., David Allis, C., Melnick, A., et al. (2013). RUNX1 Is a Key Target in t(4;11) Leukemias that Contributes to Gene Activation through an AF4-MLL Complex Interaction. Cell Reports 3, 116–127.

Wilkinson, K.A., Merino, E.J., and Weeks, K.M. (2006). Selective 2 -hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nature Protocols 1, 1610–1616.

Wolfe, A.L., Singh, K., Zhong, Y., Drewe, P., Rajasekhar, V.K., Sanghvi, V.R., Mavrakis, K.J., Jiang, M., Roderick, J.E., Van der Meulen, J., et al. (2014). RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. Nature 513, 65–70.

Wong, Q.W.-L., Vaz, C., Lee, Q.Y., Zhao, T.Y., Luo, R., Archer, S.K., Preiss, T., Tanavde, V., and Vardy, L.A. (2016). Embryonic Stem Cells Exhibit mRNA Isoform Specific Translational Regulation. PLoS One 11, e0143235.

Wu, H., Sun, S., Tu, K., Gao, Y., Xie, B., Krainer, A.R., and Zhu, J. (2010). A Splicing-Independent Function of SF2/ASF in MicroRNA Processing. Molecular Cell 38, 67–77.

Wu, Y., Tan, X., Liu, P., Yang, Y., Huang, Y., Liu, X., Meng, X., Yu, B., Wu, M., and Jin, H. (2019). ITGA6 and RPSA synergistically promote pancreatic cancer invasion and metastasis via PI3K and MAPK signaling pathways. Experimental Cell Research 379, 30–47.

Xiao, Y., Hsiao, T.-H., Suresh, U., Chen, H.-I.H., Wu, X., Wolf, S.E., and Chen, Y. (2014). A novel significance score for gene selection and ranking. Bioinformatics 30, 801–807.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science 347, 1254806.

Xueqing, H., Jun, Z., Yueqiang, J., Xin, L., Liya, H., Yuanyuan, F., Yuting, Z., Hao, Z., Hua, W., Jian, L., et al. (2020). IGF2BP3 May Contributes to Lung Tumorigenesis by Regulating the Alternative Splicing of PKM. Frontiers in Bioengineering and Biotechnology 8.

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

Zahler, A., Neugebauer, K., Lane, W., and Roth, M. (1993). Distinct functions of SR proteins in alternative pre-mRNA splicing. Science 260, 219–222.

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

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol. Cell 40, 939–953.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature Communications 10.

Zhu, J. (2000). Pre-mRNA splicing in the absence of an SR protein RS domain. Genes & Development 14, 3166–3178.

Zhuo, Z., Yu, Y., Wang, M., Li, J., Zhang, Z., Liu, J., Wu, X., Lu, A., Zhang, G., and Zhang, B. (2017). Recent Advances in SELEX Technology and Aptamer Applications in Biomedicine. Int. J. Mol. Sci. 18.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research 31, 3406–3415.