Biochemical and Functional Characterization of Human RNA Binding Proteins

by

Peter Freese

A.B., Harvard University (2012)

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

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2018

Author .....
Computational and Systems Biology
December 22, 2017

Certified by .....
Christopher B. Burge
Professor of Biology and Biological Engineering
Thesis Supervisor

Accepted by .....
Christopher B. Burge
Director, Computational and Systems Biology Graduate Program

Biochemical and Functional Characterization of Human RNA Binding Proteins

by

Peter Freese

Submitted to the Program in Computational and Systems Biology
on December 22, 2017, in partial fulfillment of the
requirements for the degree of Doctor of Philosophy

Abstract

RNA not only shuttles information between DNA and proteins but also carries out many other essential cellular functions. Nearly all steps of an RNA's life cycle are controlled by approximately one thousand RNA binding proteins (RBPs) that direct RNA splicing, cleavage and polyadenylation, localization, translation, and degradation. Despite the central role of RBPs in RNA processing and gene expression, they have been less well studied than DNA binding proteins, in part due to the historical dearth of technologies to probe RBP binding and activity in a high-throughput, comprehensive manner. In this thesis, I describe the affinity landscapes of the largest set of human RBPs to date elucidated through a high-throughput version of RNA Bind-N-Seq (RBNS), an unbiased *in vitro* assay that determines the primary sequence, secondary structure, and contextual preferences of an RBP. The 78 RBPs bound an unexpectedly low diversity of RNA motifs, implying convergence of binding specificity toward a small set of RNA motifs characterized by low compositional complexity. Offsetting the low diversity of sequence motifs, extensive preferences for contextual features beyond short linear motifs were observed, including bipartite motifs, flanking nucleotide content, and preference for or against RNA structure. These features likely refine which motif occurrences are selected in cells, enabling RBPs that bind the same linear motif to act on distinct subsets of transcripts. Additionally, RBNS data is integrated with complementary *in vivo* binding sites from enhanced crosslinking and immunoprecipitation (eCLIP) and functional (RNAi/RNA-seq) data produced through collaborative efforts with the ENCODE consortium. These data enable creation of "RNA maps" of RBP activity in pre-mRNA splicing and gene expression levels, either with (eCLIP) or without (RBNS) crosslinking-based assays. The mapping and characterization of RNA elements recognized by over 200 human RBPs is also presented in two human cell lines, K562 and HepG2 cells. Together, these novel data augment the catalog of functional elements encoded in the human genome to include those that act at the RNA level and provide a basis for how RBPs select their RNA targets, a fundamental requirement in more fully understanding RNA processing mechanisms and outcomes.

Thesis Supervisor: Christopher B. Burge
Title: Professor of Biology and Biological Engineering

Acknowledgments

I would like to thank my advisor, Prof. Chris Burge, for providing an incredibly rich environment in which to conduct research and for teaching me to be a rigorous computational biologist grounded in fundamental biological questions. Having never performed research on RNA or touched high-throughput sequencing data sets, the lab has been a wonderful place to learn new skills, formulate and test hypotheses, and receive feedback that has fueled my intellectual growth. A particular thanks goes out to Burge lab postdoc Daniel Dominguez and fellow CSB graduate student Maria Alexis, who have been phenomenal RBNS collaborators over the past years and central to this work as well as my overall scientific development; without either, I'm sure the RBNS project would have been nowhere near as successful or impactful as it has turned out to be. Additionally, I would like to thank all other past and present Burge lab members for their feedback and insights over the past five years, as well as their friendship and great company at our lab social events and beer hours, creating a multitude of memorable experiences outside of the lab. In particular, I'd like to thank Athma for being an always pleasant and engaging baymate, Alex and Nicole for being early RBNS collaborators who whetted my appetite for studying protein-RNA interactions, Jennifer and Yevgenia for our lighthearted coffee breaks, Matt for always being encouraging and asking great biological questions, Bridget for planning great graduate student social events and group exercise class outings, Canadian Peter for his always outspoken and thought-provoking discussions, Marvin, Ana, Emma, and Kayla for their smiling faces and fresh perspectives, Genny for being a friendly face who always made me feel welcome in the Burge lab, Cassie, Myles, Amanda and Tsultrim for being amazing technicians who helped my understanding of the technical complexities of the RBNS assay, and Jason for being generous in taking me under his wing during my rotation despite being on paternity leave.

My Thesis Advisory Committee members, Phil Sharp and Manolis Kellis, have been very generous in providing their guidance and expertise at my committee meetings as well as facilitating my professional development. The connections I have made with and through them have been and will continue to be incredibly valuable. I would also like to thank Melissa Moore for agreeing to serve as my external committee member and taking time out

of her incredibly busy schedule to support my scientific endeavors.

I would like to thank my ENCODE RBP collaborators, who have been an instrumental source of feedback on my work and incredibly generous in sharing their time, data, and expertise in our group efforts. Through our work, I have learned a tremendous amount not only about RBP-RNA interactions and regulation but also the complexities and promise of studying biological problems through massive high-throughput integrative functional genomic assays. In particular, Brent Graveley's overall leadership of the RBP project has been instrumental in its success, Xintao Wei has always been a patient and responsive collaborator who made the logistics of working on a high-throughput consortium project very smooth, and Gene Yeo and his lab members Eric Van Nostrand and Gabe Pratt provided excellent scientific guidance and insights into RNA-RBP biology through their improved experimental and computational CLIP methods.

My undergraduate research advisor Irene Chen and postdoc Kirill Korolev initially sparked my interest in conducting scientific research and showed me the challenges and rewards seeing a research project through from start to finish. I would likely not have pursued graduate studies without these formative academic role models and their support of my ongoing scientific development.

I would like to acknowledge the amazing friendship of my CSB 2012 classmates Vincent, Colette, Mandy, Mariana, Nezar, and Rotem, with whom I've shared over 5 years of incredible outings throughout MIT, Boston, and the Northeast. We've celebrated countless birthdays, taken summer trips to explore and relax, and supported each other through the ups and downs of a graduate career in a way that I couldn't have imagined when we started this program together. It has been a pleasure to befriend the many other CSBs in years above and below me as we gather at retreats and other CSB events to share our research, successes, and struggles. Additionally, CSB administrator Jacquie Carota has done a phenomenal job of making the graduate program run smoothly and has always been a warm, smiling face to run into and chat with on the 2nd floor of Building 68.

Finally, I would like to thank my family members, who have been incredibly patient and generous in their support of my PhD endeavors and broader educational upbringing that has allowed me to be where I am today. My mom, dad, and sister Kelly have been endless

supporters of my scientific passions and research, and always provided me with a down-to-earth perspective of the incredible fortunes I have been lucky enough to receive through graduate education at MIT. Going home to Minnesota to visit them and my extended family as well as our phenomenal vacations have helped me contextualize my research and career within a broader perspective. This work and my intellectual growth as a graduate student would not have been possible without their unflagging support.

Contents

List of Figures

Chapter 1

Introduction

The central dogma of molecular biology, first stated nearly 60 years ago (Crick [1958]), detailed that genetic information does not transfer from protein to DNA but instead typically from DNA → RNA → protein. With proteins effecting most functions in the cell and DNA being the central molecule of heredity passed from generation to generation, RNA was originally viewed as a somewhat less important intermediary between information and action. However, work over the past decades has revealed that RNA is a highly dynamic and regulated molecule, subject to a wide range of RNA processing mechanisms in eukaryotes (Mitchell and Parker [2014]). For messenger RNA (mRNA) molecules that encode protein sequences, these processes include 5' capping and 3' polyadenylation of the newly transcribed mRNA; constitutive and alternative pre-mRNA splicing; RNA editing; export from the nucleus into the cytoplasm; subcellular localization within different parts of the cytoplasm; regulation of translational efficiency; RNA surveillance and quality control; and regulation of mRNA stability and eventual degradation of the RNA. RNA binding proteins (RBPs) play critical roles in these post-transcriptional pathways, with each of the 1,000+ RBPs in humans having unique RNA binding activity and protein-protein interaction partners to produce a diverse assortment of highly regulated RNA molecules from the relatively modest ∼20,000 genes encoded in the genome.

1.1 RNA binding proteins

In eukaryotic cells, each mRNA is bound by a dynamic repertoire of RNA binding proteins (RBPs) such that it exists as an mRNA-protein complex (messenger ribonucleoprotein, mRNP, Singh et al. [2015], Rissland [2017]). The proper pre-mRNA splicing, processing, nuclear export, subcellular localization, and stability and degradation of mRNAs critically depend on these RBP-RNA interactions. Some mRNP components are members of large macromolecular machines, such as the spliceosome or ribosome, that bind mRNA in a coordinated manner to direct processes such as splicing, nuclear export, translation, and mRNA decay. These RBPs typically are deposited on mRNAs according to earlier RNA processing events or through interaction with mRNA landmarks such as the 5' cap, pre-mRNA splice sites, or the poly(A) tail. In addition to these members of common machineries, other RBPs interact with sequence-specific features of individual mRNAs. These proteins often bind mRNAs concurrently with core machineries to regulate specific steps in RNA processing, such as splicing factors binding introns or exons to influence alternative splicing or AU-rich element binding proteins binding 3' UTRs to influence mRNA stability or translation. However, not all RBPs fall into these extreme categories of high sequence specificity or general machineries but instead often operate in the middle ground of a specificity continuum from promiscuous to selective (Mitchell and Parker [2014]). For instance, Pumilio domain-containing proteins bind eight to ten RNA bases with high specificity at one extreme (Zamore et al. [1997]) while DEAD-box helicases have shown little dependence on RNA sequence to rearrange their mRNA substrates (Linder and Jankowsky [2011]). In between, SR (Serine/Arginine-rich) proteins and HNRNPs (hetergeneous nuclear ribonucleoproteins) exhibit discernible sequence preferences but are able to bind a wide range of targets to effect their transcriptome-wide splicing outcomes (Goren et al. [2006], Geuens et al. [2016]).

1.1.1 RNA binding domains (RBDs) and sequence-specific recognition of RNA

The sequence and/or structural specificity of an RBP for RNA targets is typically mediated through one or more well-defined RNA binding domains (**Fig.** 1-1). Among the ∼600

structurally distinct RBD classes catalogued by Gerstberger et al. [2014], just 20 have more than ten members, with most having just one or two members. The three most prevalent sequence-specific RNA binding domains in humans are the RNA Recognition Motif (RRM, ∼278 human RBPs), various types of Zinc Finger domains (ZF, ∼90 human RBPs though also present hundreds of DNA binding proteins), and the hnRNPK homology domain (KH, ∼63 human RBPs) (Gerstberger et al. [2014]). The RNA binding properties of other prevalent RBDs, including those that bind double-stranded RNA, is less well understood and/or thought to occur largely in a sequence-independent manner (see below).

A common feature of the most abundant RBD classes in mRNA binding proteins is their frequent occurrence in multiple copies and/or in combination with different RBD types (Gerstberger et al. [2014]), with such modular design providing greater specificity and affinity to permit the diverse biological functions employed by eukaryotic RBPs (Lunde et al. [2007]). Interestingly, the average number of RBDs within a protein is inversely correlated with the number of nucleotides that RBD type commonly binds, ranging from ZFs binding an average of ∼3 nt with more than 3 ZFs per protein to the Pumilio Homology Domain (PUM-HD) binding ∼8 nt with just one RBD per protein on average (Mitchell and Parker [2014]).

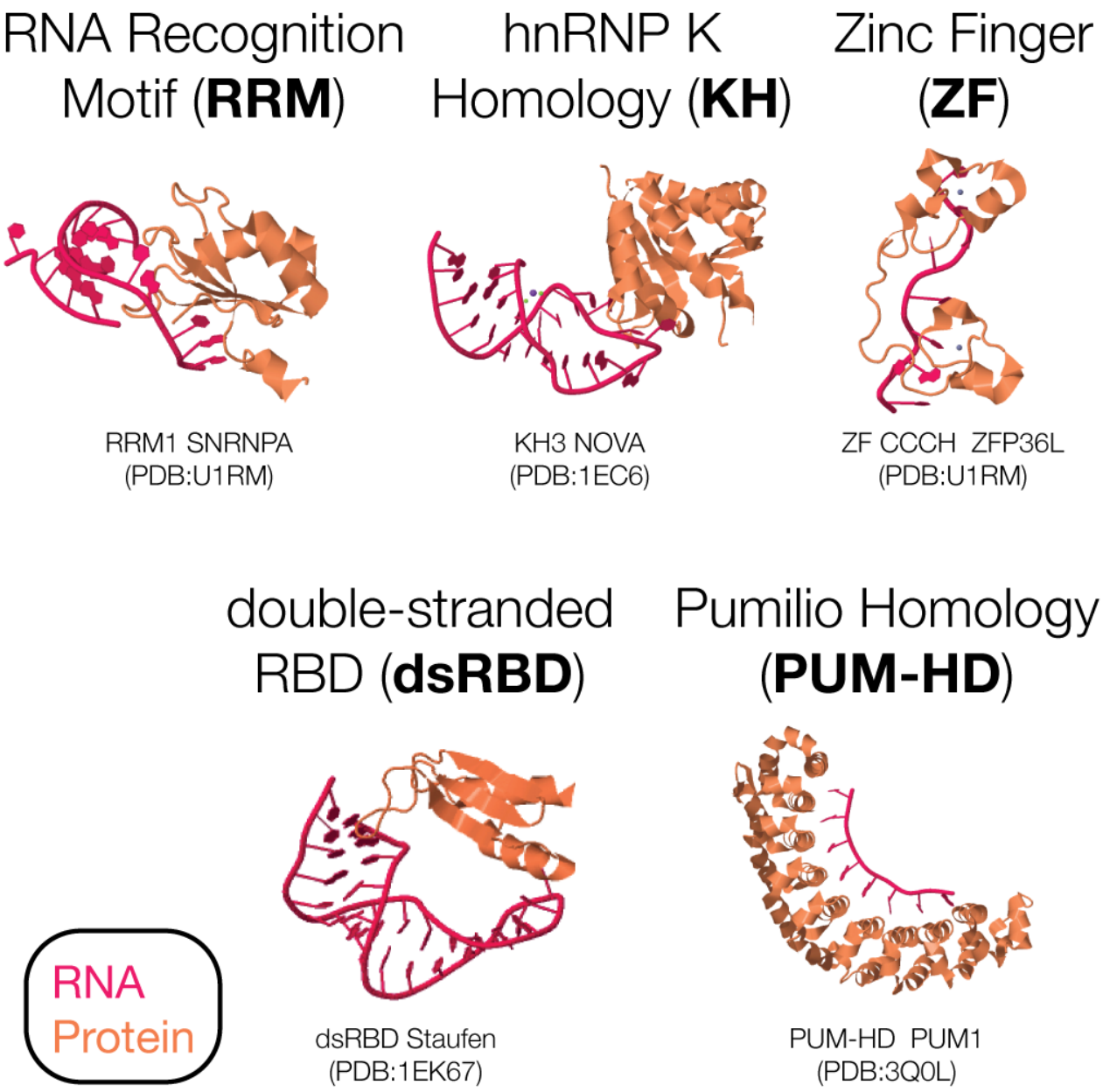

Figure 1-1: **Common RNA Binding Domain (RBD) types**
Among the most prevalent RNA binding domains in eukaryotic proteomes include the RNA Recognition Motif (RRM, present in ∼278 human RBPs); hnRNPK Homology Domain (KH, present in ∼63 human RBPs); and Zinc Finger Domain (ZF, present in ∼961 human proteins, though most bind DNA with only dozens currently implicated in RNA binding). Other well-defined RBDs include the double-stranded RBD (dsRBD, present in 26 human RBPs) and Pumilio Homology Domain (PUM-HD, present in 3 human RBPs). All numbers taken from Uniprot human entries with example structures from PDB shown.

RNA Recognition Motif (RRM) domain

The RRM is the most abundant RBD in higher eukaryotes, occurring in nearly 300 human RBPs (Gerstberger et al. [2014]). Discovered in 1989 as the RNA binding domain of U1-70K (Query et al. [1989]), it is present in all kingdoms of life including prokaryotes and viruses with RRM-containing proteins involved in most post-transcriptional gene regulatory pathways (Afroz et al. [2015]). RRMs are typically ∼90 amino acids in length with a characteristic $\beta_1 - \alpha_1 - \beta_2 - \beta_3 - \alpha_2 - \beta_4$ arrangement that folds into a four-stranded antiparallel β-sheet packed against two α-helices. The loops between the secondary structural elements can vary in length and are typically disordered in their free form, and variability is also occasionally seen in the secondary structural elements themselves (for example, one α-helix in U2AF1 is three times longer than usual) or extensions at the RRM extremities (Afroz et al. [2015]).

RRMs interact with anywhere from two to eight RNA nucleotides, though three to four is most common. They typically recognize single-stranded RNA with the β-sheet surface contacting the RNA bases spread on the protein surface across the β-sheet from $\beta4$ to $\beta2$. RRMs are characterized by two consensus sequences: ribonucleoprotein (RNP) 1 and 2, which are 8 and 6 amino acids long on the $\beta3$ and $\beta1$ strands (Lys/Arg-Gly-**Phe**/**Tyr**-Gly/Ala-**Phe**/**Tyr**-Val/Ile/Leu- X-Phe/Tyr and Ile/Val/Leu-**Phe**/**Tyr**-Ile/Val/Leu-X-Asn-Leu, respectively, X = any amino acid, Afroz et al. [2015]). Three key aromatic side chains in RNP1 and RNP2 (bolded in the sequences above) recognize two nucleotides to provide affinity, though these interactions do not explain the sequence-specificity of different RRMs. Some nucleotide biases are observed at certain positions contacted by the RRM but all four nucleotides are found in all five of the most commonly recognized positions, making the RRM an incredibly plastic RNA binding domain. Though a K_d in the nanomolar range has been observed for a few RRMs, most bind RNA targets with micromolar affinity (Afroz et al. [2015]).

hnRNPK homology (KH) domain

First identified in its namesake hnRNPK protein in 1993 (Siomi et al. [1993]), the KH domain occurs in about 63 human RBPs and is present in diverse archaea, bacteria, and eukaryotes.

KH domains are contained within proteins involved in diverse biological processes, including transcription regulation, splicing regulation, and translational control (Valverde et al. [2008]). The domain consists of a core $\beta_1\alpha_1\alpha_2\beta_2$ motif with a highly conserved GXXG loop between the two α-helices. Flanking the $\beta_1\alpha_1\alpha_2\beta_2$ motif is another α-helix and β-strand, though they can come after (type I) or before (type II) the core motif, separating KH domains into two folds: type I ($\beta_1\alpha_1\alpha_2\beta_2\beta'\alpha'$, more common in eukaryotes) and type II ($\alpha'\beta'\beta_1\alpha_1\alpha_2\beta_2$, more common in prokaryotes). Though the two folds have have different three-dimensional structures, both have the three α-helices packed onto the surface of an anti-parallel β-sheet. In addition to binding unpaired RNA, KH domains can bind single-stranded DNA, with recognition occurring via backbone interactions between four nucleic acid bases and the KH domain near the GXXG loop. A cleft in the protein structure near the GXXG loop allows protein hydrophobic interactions as well as mainchain and sidechain hydrogen bonds to mediate recognition of the four nucleobases. Due to two hydrogen bonds made with the nucleobases, adenine or cytosine are typically at positions 2 and 3 of the RNA tetramer, though one exception to this rule has been observed in a solution structure (Nicastro et al. [2015], Nicastro et al. [2012]).

KH domains are often found in multiple copies within eukaryotic proteins, including up to 14 in vigilin. For protein families whose members contain multiple KH domains, the first KH domain (KH1) is typically more similar to other KH1 domains in different proteins than it is to its other KH domains (KH2, KH3, etc.), with similar relationships seen for other KH domains (Valverde et al. [2008]). Individual KH domains bind RNA and single-stranded DNA with low-to-intermediate affinity in the micromolar range, though increased affinity and specificity are achieved through use of multiple domains which can be structurally decoupled or form a contiguous extended RNA binding surface. Similar longer RNA sequences can be recognized by dimerization of RBPs that contain a single KH domain, as is crucial for the biological activity of STAR (Signal Transduction and Activation of RNA) family proteins (Nicastro et al. [2015]).

Zinc Finger (ZF) domain

In addition to their more classically defined roles in binding double-stranded DNA, zinc finger-containing proteins can act as RNA binding modules (Font and MacKay [2010]). Indeed, the first ZFs discovered, in the transcription factor TFIIIA, were identified through their binding to double-stranded 5S rRNA in *Xenopus* oocytes (Miller et al. [1985]). ZF domains typically coordinate a zinc ion with pairs of cysteine and histidine residues, though the arrangement of these residues within the ∼30 amino acid domain composed of a β-hairpin and α-helix can vary, lending the most prevalent types of these domains to be commonly characterized as C2H2, CCHC, and CCCH/C3H1 (another smaller class of ZFs, RanBP2-type, is named for the RanBP2 protein in which it was discovered and is defined by a W-X-C-X_{2-4}-C-X_3-N-X_6-C-X_2-C motif, with such ZFs included in the FET family of FUS, EWSR1, and TAF15 RBPs).

While DNA recognition by ZFs is known to occur via major groove contacts with 3 base pairs of DNA, the reported binding of ZFs to single- or double-stranded RNA is much more varied and less well understood. For example, three of the nine ZFs in TFIIIA are important for RNA binding activity, with two of these (fingers 4 and 6) making base-specific contacts with 'flipped-out' RNA bases while finger 5 makes exclusively non-sequence-specific interactions with the RNA phosphate backbone via amino acid basic side chains (Font and MacKay [2010]). It has been proposed, based on conclusions from crystallography studies, that ZFs can recognize complex structures comprising internal loops and double helicies both through specific contacts with individual bases that are exposed for access out of a rigid RNA structure as well as sequence independent binding to the regularly folded portion of an RNA double-helix (Lu et al. [2003]). It has also been proposed that a common mode of RNA recognition of ZFs might be to bind structured RNAs. This is based on, among other studies of ZFs binding dsRNA or stem-loops with much greater affinity than ssRNA (Font and MacKay [2010]), the characterization of Wilm's tumor 1 (WT1) *in vitro*-selected RNA aptamers as requiring a hairpin loop, with tolerance for compensatory mutations that maintain proper base-pairing of the stem (Zhai et al. [2001]). Yet in contrast to this paradigm, numerous CCCH zinc fingers have been shown to recognize single-stranded A/U-rich elements and

promote degradation of their mRNAs (e.g., ZFP36, Lai et al. [1999], and Tis11d through backbone interactions with the Watson-Crick edges of A and U bases, Hudson et al. [2004]), underscoring the complexity of ZF RNA binding modes. Although CCCH and CCHC zinc finger motifs have classically been known to bind RNA, an mRNA interactome study (see Section 1.2.2) also identified a significant enrichment of AKAP95 and HC5HC2H-type zinc fingers, making them likely bona-fide RNA binders (this study identified 69 total ZF containing proteins bound to poly(A)-selected RNA in HeLa cells, Castello et al. [2012]). Mirroring their diverse RNA recognition modes, ZFs that bind to RNA are involved in many functional processes including mRNA trafficking, stability, and transcriptional and translational regulation (Wai et al. [2016]).

Other RNA binding domains and low-complexity domains

While a few other eukaryotic RBD types such as the PUM-HD bind RNA in a sequence-specific manner, most others bind in a predominantly sequence-independent manner, recognizing secondary structure or being guided to target RNAs by other protein cofactors, and/or they have not been studied in detail (Gerstberger et al. [2014]). RBDs containing a Asp-Glu-Ala-Asp (DEAD) motif, present in ∼62 human RBPs that are typically RNA helicases, recognize five nucleotides exclusively through interactions with the sugar phosphate backbone of the RNA in a characteristic bent conformation (Linder and Jankowsky [2011]). The double-stranded RNA binding domain (dsRBD), present in ∼26 human RBPs, consists of 65-70 amino acids that adopt an $\alpha\beta\beta\beta\alpha$ fold to recognize A-form double-stranded RNA through contacts to bases and ribose sugars in two successive minor grooves as well as the phosphate backbone in the intervening major groove. Traditionally thought to interact with RNA without any sequence specificity, recent structural information shows that dsRBDs can recognize additional sequence features beyond the A-form RNA helix (Masliah et al. [2013]).

In addition to well-defined RNA binding domains containing characteristic amino acids that recognize specific RNA sequences and/or structures, some low-complexity domains may also play roles in contacting RNA. The second most common RBD type in the human genome after the RRM, the RG/RGG motif, is characterized by a region rich in arginines and glycines and is present ∼80-100 human RBPs depending on domain definition. These domains display

degenerate binding specificity yet still display different degrees of preference for RNA with some domains achieving affinity approaching that of their full-length protein counterparts (Thandapani et al. [2013], Ozdilek et al. [2017]). The RS (arginine/serine-rich) domains of SR and other (e.g., U2AF) proteins are classically thought to contact one another to directly mediate protein-protein interactions (Busch and Hertel [2012]), though work from Michael Green's lab has shown that these domains intermittently make direct contacts with RNA sequences important for splicing, including the branchpoint and 5' splice site (Shen et al. [2004], Shen and Green [2004]). Changes in the phosphorylation state of serines within RS repeats could potentially alter the RS domain interaction mode between making protein-protein and making protein-RNA contacts (Hertel and Graveley [2005]).

Altogether, many of these other RNA binding and low-complexity domain types are less well studied than RRM, ZF, and KH domains, and further investigation into how they impart RNA specificity and/or affinity is warranted.

1.1.2 RBP-mediated regulation of pre-mRNA splicing, RNA stability, and RNA translation

The following sections contain an overview of the processes of pre-mRNA splicing and RNA stability & translation, particularly noting steps in which RBPs are known to play regulatory roles. Additionally, examples of RBPs regulating these processes are provided, often highlighting the context-dependent manner in which an RBP can have different effects on the same post-transcriptional regulatory pathway depending on the cell state or type, binding location, and competing or cooperating RBP(s).

Pre-mRNA splicing and *cis*-Splicing Regulatory Elements (SREs)

Metazoan protein-coding genes contain multiple exons split up by intervening introns which must be removed from the pre-mRNA to produce a mature mRNA ready for nuclear export and translation. This process of splicing is carried out by the complex macromolecular machine known as the spliceosome, which contains five small nuclear ribonucleoproteins (snRNPs) and ∼170 auxiliary proteins that enter and exit the spliceosome during various

stages of the splicing reaction (Wahl et al. [2009]). Through the alternative inclusion or exclusion of exons (or parts thereof), different isoforms of mature mRNAs are produced from the same pre-mRNA through a process known as alternative splicing, creating messages with different coding-potential or regulatory capacity.

Each splicing reaction requires three relatively short core RNA sequence elements in the pre-mRNA: the 5' splice site (5'ss), the 3' splice site (3'ss), and the branchpoint sequence (BPS). The 5'ss, at the end of the upstream exon and beginning of the intron, has human consensus sequence CAG|GUAAGU (|=exon/intron boundary, Roca et al. [2013]) and is recognized via complementarity to the 5' portion of the U1 snRNA. At the other end of the intron, the 3' splice site consists of a ∼25 nt sequence rich in cytosine and uridine known as the polypyrimidine tract, which is preceded upstream by a branchpoint sequence (BPS) and downstream by an intron-terminal AG dinucleotide. The branchpoint sequence is conserved in yeast with consensus UACUA**A**C but is quite degenerate in humans with consensus YUN**A**Y, Y = pyrimidine (Gao et al. [2008]). The polypyrimidine tract is initially recognized the larger subunit of the U2 auxiliary factor (U2AF) heterodimer composed of U2AF65 (also known as U2AF2) and U2AF35 (also known as U2AF1). The ternary complex of U2AF65 with U2AF35 and splicing factor 1 (SF1) recognizes the surrounding BPS and 3'ss AG, with U2AF65 recruiting the U2 snRNP to replace it in the active spliceosome (Agrawal et al. [2016]). The branchpoint adenosine (bolded in previous sequences) is bulged out via complementarity of the flanking RNA sequence with a 'GUAGUA' sequence in U2 snRNA to act as the nucleophile in the first of two transesterification reactions during splicing. In the first reaction, the 2'-OH of the branchpoint A attacks the conserved guanine at the 5' end of the intron to produce a 2'-5' phosphodiester RNA lariat structure and a free 3'-OH at the upstream exon. In the second reaction, the 3'-OH of the upstream exon attacks the phosphodiester bond of the guanosine at the 3' end of the intron to ligate the two exons, resulting in the exons being spliced together and the intron being released as a lariat structure.

In addition to the ∼170 core spliceosomal proteins that partake in all splicing reactions, alternative splicing is regulated by the binding of *trans*-acting RBPs (splicing factors) to short *cis*-sequences in the pre-mRNA to enhance or inhibit the use of adjacent splice

sites. These splicing regulatory elements (SREs) are broken down based on whether they are located in the **E**xon or **I**ntron and whether they **E**nhance or **S**ilence splicing from that location (ESE/ESS enhancing or silencing from exons, ISE/ISS from introns, **Fig.** 1-2). Although regulatory elements can in principle exert their action from anywhere within the pre-mRNA, most studies have focused on regulatory sequences within the exon or proximal flanking introns (∼200-300 nt adjacent to splice sites). Though some splicing factors are ubiquitously expressed, many act in a tissue- or developmental-specific manner to execute alternative splicing programs central to tissue and organ development (Baralle and Giudice [2017]). Thus, one challenge in identifying potential SREs and their corresponding RBPs is that the same pre-mRNA sequences can be recognized differently in different cell types, partially due other RBPs expressed or the different RBP:RNA stoichiometries in each cell type. Additionally, the same *cis*-sequences, *trans*-factors, or combinations thereof can have different effects on splicing outcomes depending on the sequence context and their position within the intron or exon (Fu and Ares Jr [2014]). For example, G-runs can enhance splicing from intronic locations (McCullough and Berget [1997]) or repress splicing from exonic locations (Chen et al. [1999]), and RBFOX2 and Nova typically suppress cassette exon inclusion from upstream introns but enhance it from downstream introns (Yeo et al. [2009], Ule et al. [2006]). As it has been estimated that the three core human splice site motifs (5'ss, 3'ss, and BPS) only contain about half of the information needed to accurately define intron/exon boundaries (Lim and Burge [2001]), these *cis*-sequences and their *trans*-factors likely play a large role in ensuring the high fidelity of splicing.

Two key protein families that regulate pre-mRNA splicing and subsequent aspects of RNA metabolism are the SR (serine/arginine-rich) and HNRNP (heterogeneous nuclear ribonucleoprotein) proteins which often function by interacting with ESEs and ESSs/ISSs, respectively. The SR proteins (∼12 in human, depending on definition) are characterized by N-terminal RRM(s) which typically bind ESEs and C-terminal RS domains that participate in protein-protein interactions and facilitate spliceosome assembly (Busch and Hertel [2012], Graveley and Maniatis [1998]), though the RS domains have also been shown to directly contact RNA splicing signals (see "Other RNA binding domains and low-complexity domains" above). The study of SR proteins originates in *Drosophila* screens that identified splicing

factors containing protein domains rich in arginine and serine dipeptides (Chou et al. [1987], Moretti et al. [1987], Amrein et al. [1988]), with subsequent identification of human splicing factors SRSF1 and SRSF2 as proteins that also contain such RS-rich domains in addition to their RRM(s) (Ge and Manley [1990], Krainer et al. [1990], Fu and Maniatis [1992]). The diverse HNRNPs (∼37 in human, first detailed by the Dreyfuss lab, Piñol Roma et al. [1988]) contain one or more RBDs (typically RRMs, though five contain KH domains) and often RGG boxes (repeats of Arg-Gly-Gly tripeptides) and glycine-rich, acidic or, proline-rich domains that mediate protein-protein interactions and influence splicing outcomes through a wide variety of mechanisms (Busch and Hertel [2012], Geuens et al. [2016]).

A major goal of the field is the development of a 'splicing code' which could predict the splicing outcomes of any transcript from its primary sequence. One central feature in such a splicing code is a 'parts list' of splicing factors and the motif(s) they bind. Indeed, position- and tissue-specific effects of sequence elements that match the motifs of the FOX, NOVA, MBNL, CELF, TIA, PTB, and QKI protein(s) arose in an early inferred splicing code (Barash et al. [2010]). An updated splicing code model detected 2080 significant correlations between RNA-seq Ψ values of 10,689 exons and densities of 98 *in vitro* RBP binding motifs (Ray et al. [2013]) in six intronic or exonic regions (Xiong et al. [2015]), and undoubtedly more will arise with an expanded catalog of RBP binding motifs. Such splicing codes and related efforts will not only provide greater mechanistic insight into the interplay between *trans*-factors and their *cis*-regulatory elements in splicing regulation but also may provide opportunities for understanding genetic determinants of human disease and support for casual variants acting at the level of pre-mRNA splicing.

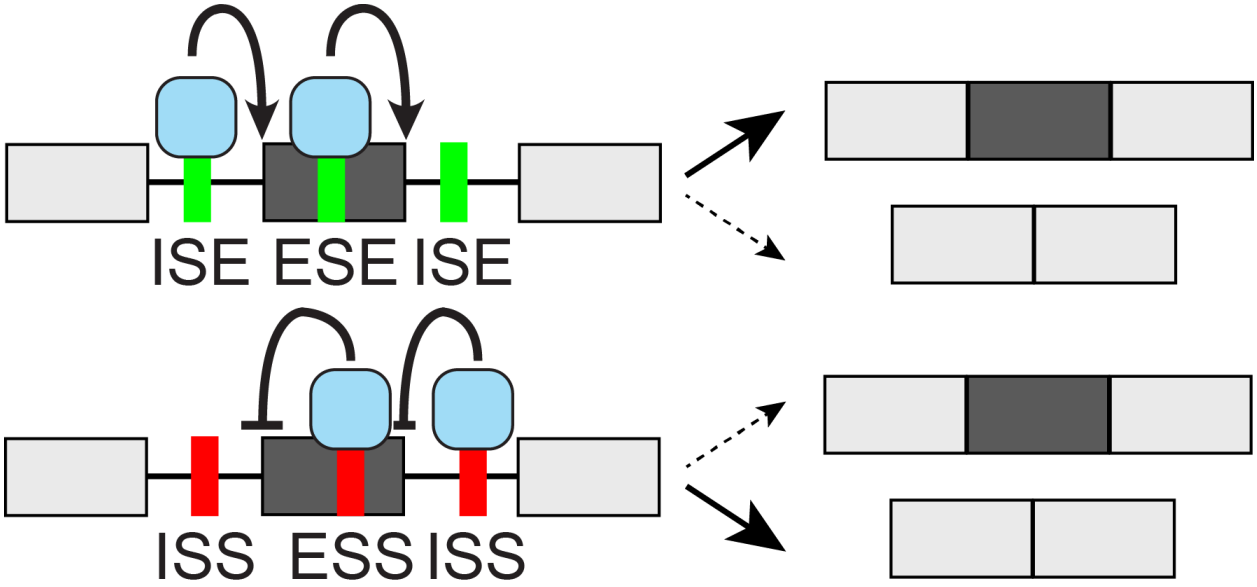

Figure 1-2: ***cis*-acting splicing regulatory elements and *trans*-acting splicing factors** *cis*-acting splicing regulatory sequences in the pre-mRNA include ESE/ESS (**E**xonic **S**plicing **E**nhancer/**S**ilencer) and ISE/ISS (**I**ntronic **S**plicing **E**nhancer/**S**ilencer) elements. Typically 4-6 nt in length, they are recognized by *trans*-acting proteins known as splicing factors, which promote or inhibit productive assembly of a catalytically active spliceosome that carries out the splicing reaction. The 'context-dependent' nature of *cis*-regulatory splicing sequences makes their activity highly reliant on location within the exon and flanking introns as well as the presence of other synergistic or antagonistic elements nearby. Additionally, different sets of splicing factors are expressed in different tissues and cell states, resulting in highly tissue-specific and dynamic alternative splicing programs.

3' UTR *cis*-elements and regulation of mRNA stability and translation

3' untranslated regions (3' UTRs) are the noncoding parts of mRNAs following stop codons. Their length expansion in humans compared to yeast, as well as the increased prevalence of alternative 3' UTR isoforms, suggests an important role for 3' UTRs in the regulation of genes of higher organisms (Mayr [2017]). Best known to regulate the degradation, translation, and localization of mRNAs, 3' UTRs function by recruiting RBPs that bind to *cis*-elements and recruit effector proteins such as deadenylases (Rissland [2017]). Though most 3' UTR functions are carried out in the cytoplasm, some RBPs are loaded onto the mRNA in the nucleus and are exported with the message as an mRNP while others are added locally in the cytoplasm (Mayr [2017]). Co-transcriptional loading of RBPs at promoters can result in their deposition onto mRNAs with their remained association allowing them to play 3' UTR regulatory roles in the cytoplasm, enabling crosstalk between the seemingly unrelated processes of mRNA synthesis in the nucleus and translation or decay in the cytoplasm (Bregman et al. [2011], Moore and Proudfoot [2009]).

3' UTRs determine protein output by regulating mRNA stability and translation primarily through the activity of AU-rich elements and miRNAs. Though about half as conserved as coding sequences on average, 3' UTRs often contain 'islands' that are conserved similarly to coding sequences and frequently contain binding sites for miRNAs or RBPs (Xie et al. [2005], Friedman et al. [2009]). AU-rich elements (AREs), characterized by variants of the pentamer "AUUUA" occurring in variable length repetitions, were one of the first motifs discovered in 3' UTRs, preferentially found in genes subject to tight expression level regulation such as immune-regulatory factors, cytokines, and proto-oncogenes (Barreau et al. [2006]). The mRNA half-lives of these genes are shorter than a half hour compared to a median $\sim$7 hour half-life across the entire mammalian transcriptome (Sharova et al. [2009]). Effects of AREs and miRNAs on protein abundance in cell lines is more modest (Baek et al. [2008], Selbach et al. [2008], Spies et al. [2013]), possibly due to 3' UTRs not substantially regulating protein abundance under steady-state growth conditions in cell culture with miRNAs and RBP-mediated repression more playing important roles in select biological contexts (Mayr [2017]). Recognized by proteins known as ARE-binding proteins (ARE-BPs), AREs can

have different effects on mRNA depending on the protein(s) bound to them. Competition between ARE-BPs and which one(s) ultimately bind to messages can result in either mRNA stabilization and translational enhancement (commonly observed for the Hu/ELAV family RBPs) or in mRNA destabilization and translational repression (e.g., HNRNPD, ZFP36, and TIA1). Consistent with 3' UTR repressive elements being more prevalent than activating elements, 3' UTR length is typically inversely correlated with mRNA stability and gene expression levels. Furthermore, among genes with alternative polyadenylation sites, those highly expressed typically prefer the proximal poly(A) site to produce a shorter 3' UTR while genes with lower expression levels often use the distal poly(A) site to produce a longer 3' UTR (Matoulkova et al. [2012]).

In addition to AU-rich elements, other 3' UTR *cis*-elements are crucial for post-transcriptional gene regulation through their interplay with RBPs. GU-rich elements in arrangements of 2-5 overlapping pentamers are contained in at least 5% of human mRNAs. They are present in the 3' UTRs of short-lived mRNAs in T-lymphocytes and contribute to additional post-transcriptional pathways such as deadenylation, mRNA decay, and mRNA splicing (Halees et al. [2011]). The CELF family of six RBPs have been identified as recognizing GU-rich elements, with two members almost identical in their RBDs having opposing effects on post-transcriptional regulation (CELF1 has destabilizing effects on mRNAs with subsequent increased translational efficiency, while CELF2 has stabilizing effects and inhibits translation, Vlasova et al. [2008]). CA-rich elements, with A/C being the most common dinucleotide repeat found in the human genome, are located in both coding and noncoding regions. Thought to be predominantly recognized by HNRNPL, they typically exert stabilizing effects on mRNA (Hui et al. [2003]). Other 3' UTR *cis*-elements include CU-rich elements, iron responsive elements, and selenocysteine insertion sequence elements (Matoulkova et al. [2012]).

In addition to being platforms for regulating mRNA stability, 3' UTRs play a major role in regulating the translation of an mRNA molecule into protein, and indeed the mechanisms by which *trans*-acting factors regulate mRNA stability and translation can be coupled. Translation is initiated when eukaryotic translation initiation factors (eIFs) recruit the small ribosomal subunit to the 5' end of the mRNA. The assembly of the eIF4F complex, com-

posed of the cap-binding protein eIF4E, the scaffold protein eIF4G, and the RNA helicase eIF4A, is rate-limiting in this process of translation initiation. eIF4G has binding sites for eIF4E, eIF4A, eIF3, and PABP (poly(A)-binding proteins), making it a hub for regulation of translation. The eIF4G/PABP interaction stimulates the formation of a closed-loop mRNA structure, activating cap-dependent translation and facilitating ribosome recruitment to the mRNA. mRNA translation is thus regulated by the formation of the eIF4F complex and mRNA circularization induced by eIF4G/PABP, with RBPs and miRNAs promoting or inhibiting these processes to affect translational efficiency. Once the small ribosomal subunit is recruited to the mRNA, this 40S subunit and its associated factors then scan the 5' UTR, recognize the start codon, and the large ribosomal subunit finally joins to form a full ribosome competent for elongation (Fukao and Fujiwara [2017]).

Many proteins, such as ARE-BPs, play roles in regulating translation via 3' UTR or 5' UTR binding in addition to possible other roles in regulating mRNA metabolism. For example, ZFP36 (also known as TTP) bound to AREs not only directly binds the deadenylase complex and recruits it to the mRNA (Fabian et al. [2013]), but ZFP36 also inhibits the translation of target mRNAs by directly interacting with a specific isoform of eIF4E, eIF4E2, that likely disrupts the assembly of the eIF4F complex (Tao and Gao [2015]). CPEB, the cytoplasmic polyadenylation element binding protein, binds a U-rich sequence in target 3' UTRs and regulates the translation of maternally deposited mRNA during oocyte embryogenesis as well as later cytoplasmic polyadenylation and activation of translation. CPEB maintains maternal mRNAs in a dormant state via binding of the protein Maskin, which contains an eIF4E-binding domain; the CPEB-Maskin complex thus competes with eIF4G for binding to eIF4E (Stebbins-Boaz et al. [1999]). The Hu family of proteins, composed of four highly conserved ARE-BPs in vertebrates with one (HuR/ELAVL1) ubiquitously expressed while the other three (HuB/ELAVL2, HuC/ELAVL3, HuD/ELAVL4) are primarily expressed in neurons, regulate numerous aspects of RNA metabolism including mRNA stability, poly(A)-tail length, and mRNA translation (Fukao and Fujiwara [2017]). Among the known translational regulatory roles of these proteins are HuR upregulating p53 protein levels after UV irradiation by binding the p53 3' UTR (Mazan-Mamczarz et al. [2003]); HuR and HuD inhibiting p27 translation by binding to an internal ribosome entry site in the 5'

UTR (Kullmann et al. [2002]); HuD inhibiting translation of *Ins2* mRNA in pancreatic β cells by binding to a 22 nt sequence in its 5' UTR (Lee et al. [2012]); and HuD enhancing cap-dependent translation by binding to eIF4A in a poly(A)-dependent manner that is required for neurite outgrowth in PC12 cells (Fukao et al. [2009]). Together, reminiscent of the context-dependent RBP regulation of alternative splicing and mRNA stability, these findings underscore that Hu proteins can regulate translation in a positive or negative manner that partially depends on which messages and where within the mRNA (5' or 3' UTR) they are bound (Fukao and Fujiwara [2017]).

Although not RBPs themselves, miRNAs recruit RBPs such as Argonaute, the RISC complex, and deadenylases and TNRC6 to mediate mRNA degradation (discussed above) as well as an independent role in translational inhibition of initiation through displacement of PABP from the translation initiation complex to destroy the closed-loop structure (Moretti et al. [2012]). Yet other RBPs regulate translation at late-initiation or post-initiation steps (Gebauer and Hentze [2004]). For example, HNRNPK and HNRNPE1 inhibit the translation of *LOX* mRNA by binding a CU-rich element in 3' UTRs known as the differentiation-control element (DICE), targeting initiation factors that prevent the large ribosomal subunit from joining the small subunit at the initiation codon (Ostareck et al. [2001]).

In sum, 3' UTRs regulate gene expression, translation, and protein levels through a complex interplay of RBPs binding to UTR *cis*-elements to mediate functions via the recruitment of effector proteins in a dynamic cell state- and context-specific manner.

1.2 Approaches for the study of RNA binding proteins

The following sections contain a brief overview of approaches commonly used to profile RBP-RNA interactions at the biochemical level *in vitro* as well as at the systems level *in vivo*. Genetic studies of RNA profiling after RBP perturbation as well as the two most common structural methods to study RBPs and RBP-RNA interactions (X-ray crystallography and NMR spectroscopy) are also discussed.

1.2.1 *In vitro* RNA-RBP profiling techniques

Among the most commonly utilized assays that have been developed to characterize RBP-RNA biochemical interactions *in vitro* are:

- Systematic evolution of ligands by exponential selection (SELEX) identifies high-affinity ligands for a protein of interest through sequential cycles of ligand selection from a pool of variant sequences and amplification of the bound sequences (Tuerk and Gold [1990]). Multiple rounds of enrichment selection result in the exponential increase of high-affinity ligands, which can then be clonally isolated and characterized through electrophoresis and Sanger sequencing. While SELEX typically identifies one or a few consensus sequences of an RNA binding protein *de novo*, it is not quantitative and doesn't provide information about an RBP's lower affinity sites.

- The RNA electrophoretic mobility shift (EMSA) or 'gel-shift' assay allows for the rapid detection, visualization, and quantification of protein-RNA interactions. In a gel-shift experiment, unlabeled protein is incubated with *in vitro*-generated RNA 5' end-labeled with [γ-^{32}P] ATP. Protein-RNA complexes are separated from unbound (free) RNA by native, nondenaturing polyacrylamide gel electrophoresis (PAGE). The amount of bound RNA in the complex as well as the free RNA is measured via phosphorimaging, with the fraction of bound RNA plotted as a function of protein concentration. From this curve, the apparent equilibrium binding constant (K_d), defined as the concentration of protein at which 50% of the RNA is bound, can be derived as a measure of the affinity that the protein has for the particular RNA assayed. An advantage of the gel-shift assay is that it provides an absolute K_d for the protein-RNA interaction, though previous knowledge of a potential RNA substrate for the RBP of interest is required (Yakhnin et al. [2012]).

- Surface plasmon resonance (SPR) is a real-time, label-free optical biosensing technology that provides kinetic information about the rates of association and dissociation of an RBP for an RNA ligand of interest (Katsamba et al. [2002]). The RNA is immobilized to a gold sensor surface and a solution containing the RBP is flowed over

the surface while a light source shines on the sensor chip and is reflected to a detector. As the RBP solution is injected into the flow cell and binds to the RNA ligand, a change in the refractive index causes some of the light to be reflected at a different angle, with measurement of this index throughout RBP injection and wash out over the course of minutes at multiple different protein concentrations allowing inference of the association (k_A) and dissociation rates (k_D), and thus the dissociation constant $K_D = \frac{k_D}{k_A}$. SPR was originally used to study two RRM-containing RBPs and mutants and individual RBDs thereof (Katsamba et al. [2002]), and has more recently been utilized to provide absolute dissociation constants for RBFOX2 in a previous RNA Bind-n-Seq study (Lambert et al. [2014]). While SPR is a powerful method for measuring intermolecular interactions in real time, it requires specific instrumentation and expensive consumables, making it impractical for profiling hundreds of RBPs.

- In RNAcompete, a purified epitope-tagged RBP selects RNA sequences from an RNA pool of ∼240,000 designed mostly unstructured sequences up to 41 nt in length. Bound RNAs are identified via microarray hybridization and the 7mer binding profile of an RBP of interest is determined computationally (Ray et al. [2017]). Originally applied to nine yeast and human RBPs (Ray et al. [2009]), it has subsequently been applied to 205 RBPs from 24 diverse eukaryotic species (Ray et al. [2013]), and a more recent adaptation using a sequencing-based approach produced "Sequence and Structure Models" (SSMs) derived from 40mers for seven yeast and human RBPs performed at a single protein concentration (RNAcompeteS, Cook et al. [2017]).

- In RNA Bind-n-Seq (RBNS), a purified epitope-tagged RBP (consisting minimally of the RNA binding domains plus 50 flanking amino acids on the N- and C-terminal ends) is incubated with a pool of random 20 or 40 nt oligonucleotides, and the pulled down RBP-bound RNA is subjected to high-throughput sequencing (Lambert et al. [2014]). Typically five separate incubation reactions are performed with differing quantities of the tagged RBP (5 - 1300 nM), with each of these five libraries as well as the input RNA sequenced to a depth of ∼15-20 million reads. Computational analysis of the pulldown reads compared to the input reads provides the full spectrum of bound motifs (including

high and moderate affinity RNA sequences) as well as their secondary structure and context preferences. Importantly, because the RBNS oligo pool is random in contrast to the designed pool of ∼250,000 oligos used in RNAcompete, the RBP is presented with motifs in a wide variety of sequence and secondary structure contexts, enabling the fine-tuned dissection of an RBP's specificity and affinity landscape.

- Other *in vitro* techniques that have been used to profile the specificity and/or affinity of one or a few RBPs include: SEQRS (**S**election, high-throughput s**eq**uencing of **R**NA and **S**SLs (sequence specificity landscapes), Campbell et al. [2012]); RNA-MaP (quantitative analysis of RNA on a **Ma**ssively **P**arallel array, Buenrostro et al. [2014]); HiTS-RAP (**Hi**gh-**T**hroughput **S**equencing - **R**NA **A**ffinity **P**rofiling, Tome et al. [2014]); and RNA-MITOMI (RNA-**M**echanically **I**nduced **T**rapping **O**f **M**olecular **I**nteractions, Martin et al. [2012]).

1.2.2 *In vivo* RNA-RBP profiling techniques

While *in vitro* techniques reveal the intrinsic specificity an RBP has for RNA sequence(s), complementary *in vivo* techniques have been developed to study RBP-RNA interactions in their endogenous cellular environments, including:

- The first genome-wide studies to profile the set of RNAs bound to an RBP of interest employed **R**NA **I**mmuno**P**recipitation followed by microarray analysis (RIP-chip, Tenenbaum et al. [2000]) or later high-throughput sequencing (RIP-seq, Zhao et al. [2010]). These and other initial studies on dozens of RBPs revealed that RBP-RNA interactions are many-to-many; that is, each RBP typically binds hundreds to thousands of genes while each gene is typically bound by numerous different RBPs (Hogan et al. [2008]).

- **C**ross-**L**inking and **I**mmuno**P**recipitation followed by high-throughput sequencing (CLIP-seq) is the state-of-the-art method to determine an RBP's RNA targets and specific binding sites throughout the transcriptome. Though several CLIP-seq variants have been developed over the past years (see below), most share a general workflow of: stabilization of protein-RNA interactions via UV crosslinking to create covalent bonds

between amino acid residues and RNA bases in close proximity; RNA shearing; immunoprecipitation of the RBP of interest; RNA adapter ligation; reverse transcription (RT); PCR amplification; high-throughput sequencing; and mapping of reads to the transcriptome and calling of significant regions of binding. Compared with previous methods to identify transcriptome-wide protein-RNA interactions such as RIP-chip or RIP-seq, the crosslinking step of CLIP enables more stringent purification of protein-RNA complexes, and the RNase digestion step provides binding-site resolution by creating short RBP-protected fragments of length 20-70 nucleotides (Wheeler et al. [2017]). Though UV treatment at 254 nm is common due to its simplicity and ability to crosslink unmodified cells and tissues, this step does introduce known biases including: pyrimidines are more photoactivatable than purines (uracil most highly, Sugimoto et al. [2012], Hauer et al. [2015]); Cys, Lys, Phe, Trp, and Tyr residues crosslink more efficiently than other amino acids; and RBPs interacting with double-stranded RNA crosslink poorly due to the deep and narrow groove of A-form RNA helices making amino acid residues inaccessible to the nucleotides (Wheeler et al. [2017]).

Multiple variants of the CLIP-seq assay have been developed over the past decade, the most commonly used being:

- Photoactivatable Ribonucleoside CLIP (PAR-CLIP): Metabolic labeling in cell culture incorporates UV radioactive nucleoside analogs (4-thiouridine, 4sU, or 6-thioguanosine, 6sG) into RNA, which is subsequently crosslinked at 365 nm UV irradiation (Maatz et al. [2017]). RNA yield is increased due to the high reactivity compared to traditional UV crosslinking, and the mutation of T to C at the crosslinking site of 4sU after reverse transcription in up to 70% of reads provides confidence in identified interaction sites (Hafner et al. [2010]). However, the method is only applicable to cell culture systems in which the modified nucleoside can be introduced, and the modified nucleoside may introduce stress responses (Burger et al. [2013]) and favor RNAs with short half-lives as the nucleoside is present in a higher proportion of those messages.
- Individual nucleotide CLIP (iCLIP): In first generation CLIP protocols, both 5'

and 3' RNA adapters were ligated to the immunoprecipitated RNA fragments prior to reverse transcription. As the UV-induced amino acid-RNA covalent adducts often terminate reverse transcriptase, up to 80% of the cDNA products did not contain the 5' adapter and were not amplified. iCLIP addressed this by introducing a single RT primer that contained two cleavable adapter regions, with circularization after RT followed by digestion producing a linear cDNA molecule containing both sequencing adapters (König et al. [2010]). Because a number of reads terminate at the RT stop, a portion of the mapped iCLIP reads mark crosslink sites with individual nucleotide resolution. However, the circular ligation is very inefficient and has known biases of preferred nucleotides at the fragment ends, calling into question the quantitative nature of the assay (Baran-Gale et al. [2015]).

– enhanced CLIP (eCLIP): modifications to the iCLIP protocol and inclusion of input controls in the eCLIP protocol has enabled large-scale, robust profiling of hundreds of diverse RBPs (**Fig.** 1-3, Van Nostrand et al. [2016]). These improvements include omission of radiolabeling and autoradiographic visualization steps; improved enzymatic reaction efficiencies; ligation of a second 3' adapter to the single-stranded DNA after reverse transcription instead of circular ligation to increase capture efficiency; inclusion of an in-line 5-10 nt randomer in the second adapter to distinguish unique molecules from PCR duplicates, making the number of reads more quantitative and enabling a higher percentage of usable, nonduplicated reads; and sequencing of a 'size-matched' input pre-IP sample that has the same crosslinking, fragmentation, ligation, and amplification biases as the IP sample to control for nonspecific background and inherent capture biases. These improvements reduce the PCR amplification needed by ∼1000-fold (e.g., 16 eCLIP vs. 28 iCLIP PCR cycles needed for RBFOX2), requiring fewer starting cells (less than a million) and resulting in a sequenced library with a higher percentage of usable reads and greater library complexity.

The eCLIP method has recently been performed at scale for over 100 different

RNA binding proteins in two human cell lines through the Encyclopedia of DNA Elements (ENCODE) project, allowing a consistent comparison of the *in vivo* binding preferences of diverse RBPs that is explored in Chapter 3.

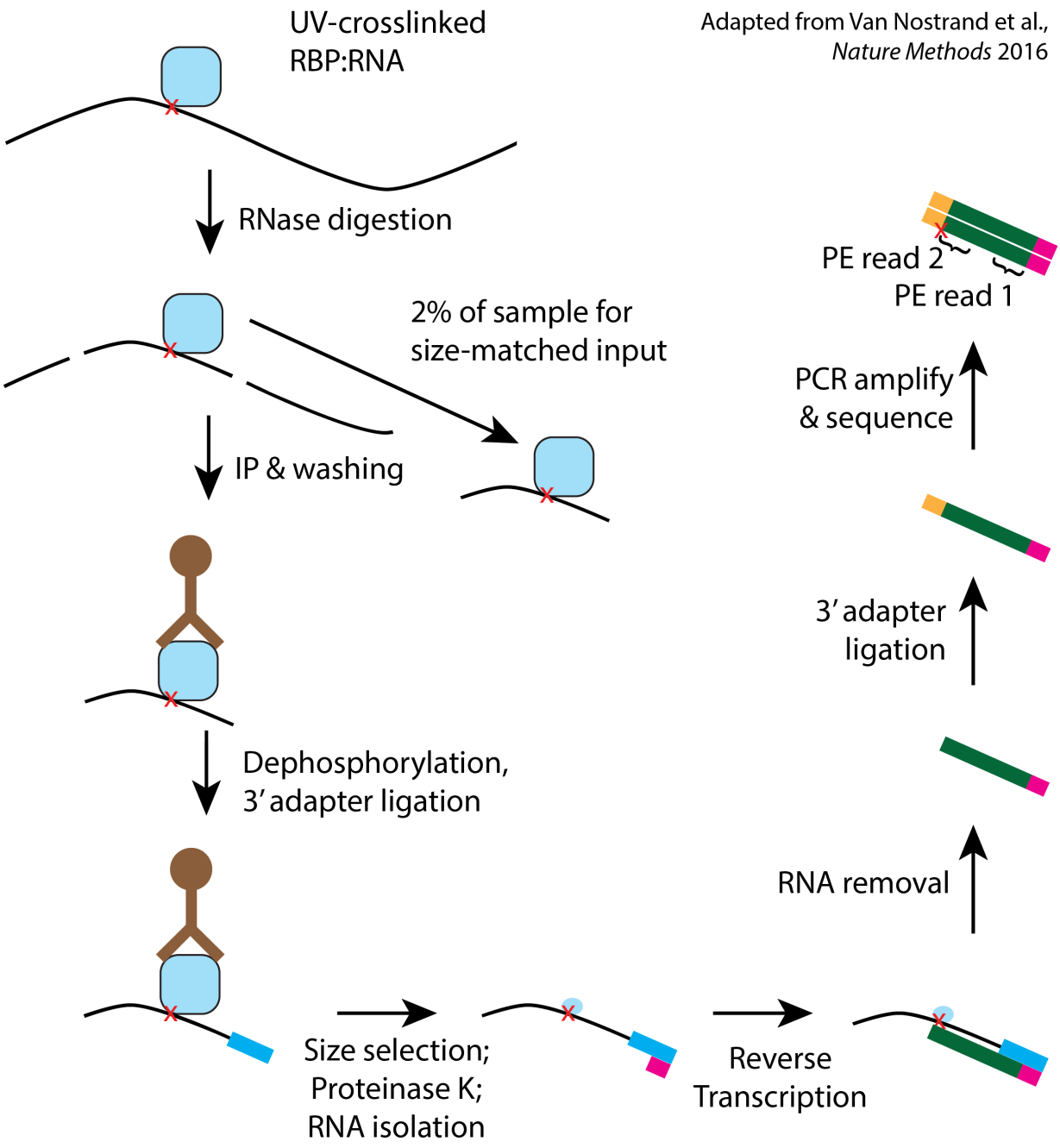

Figure 1-3: **The enhanced CLIP (eCLIP) assay**
RBP-RNA interactions are covalently stabilized by UV-crosslinking, followed by RNase I digestion and IP with a validated antibody. After stringent washes, the 3' adapter with an in-line random barcode is ligated to RNA, and a region 75 kDa (∼220 nt of RNA) above the protein size is excised and treated with proteinase K to isolate RNA. After RT and the second 3' adapter ligation (this time to cDNA), a library is prepared for high-throughput sequencing. Reads that were truncated at the RT position result in sequencing reads with read 2 of the paired-end (PE) sequence beginning just 3' of the crosslink site (Van Nostrand et al. [2016]).

- mRNA interactome capture

To assay the global scope of protein-RNA interactions in mammalian cells, 'mRNA interactome capture' methods have been developed that combine UV-crosslinking with highly stringent oligo(dT) affinity purification to enrich for proteins associated with polyadenylated RNA (Kastelic and Landthaler [2017]). These methods have uncovered vast repertoires of RBPs with hundreds of novel RBP candidates that were not predicted to bind RNA based on presence of well-defined, annotated RBDs or previous RNA-related roles reported in the literature. One pioneering study in HEK cells identified close to 800 mRNA-bound proteins, nearly one-third of which were previously unannotated as binding mRNA and 15% of which were not computationally predicted to interact with RNA (Baltz et al. [2012]). Another study in HeLa cells identified 860 mRNA-bound proteins, only 233 of which contained a classical RBD with intrinsically disordered regions being highly enriched in the mRNA-bound set (Castello et al. [2012]). These interactome capture methods will allow changes in protein-mRNA interactions in response to stimuli and disease to be studied in addition to the identification of the RNA-bound proteome in diverse cell lines and organisms.

Variants of mRNA interactome capture to identify the proteins bound by a specific RNA of interest include ChIRP-MS (**C**omprehensive **I**dentification of **R**NA binding **P**roteins by **M**ass **S**pectrometry, which utilizes cells grown in standard medium, Chu et al. [2015]) and RAP-MS (**R**NA **A**ntisense **P**urification-**M**ass **S**pectrometry, which utilizes cells grown in SILAC medium, McHugh et al. [2015]), both of which were developed to determine *Xist*-interacting proteins that mediate X-chromosome inactivation. After protein-RNA crosslinking, 20-90 nt long biotinylated DNAs complementary to the RNA of interest are used to capture RNA-protein complexes, and eluted proteins are identified through mass spectrometry. The high overlap of 9 out of 10 RAP-MS-identified RBPs also being pulled down in ChIRP-MS highlights the former's specificity, with these two techniques enabling identification of the proteins that interact with any specific RNA sequence *in vivo*.

1.2.3 Genetic studies of RNA profiling after RBP perturbation

To better understand the role(s) than an RBP plays in RNA homeostasis, the RBP can be perturbed through genetic or other means with resulting effects on RNA processing measured in a systematic and genome-wide manner through any number of sequencing-based assays. RBP perturbation can be achieved through transient siRNA-, stably integrated shRNA-, or more recent CRISPR-Cas13a-mediated (Abudayyeh et al. [2017]) RBP knockdown (KD); CRISPR-mediated RBP knockout (KO, if the RBP is not essential) or deletion of particular protein domain(s); or RBP overexpression. The most common assay to profile changes in the transcriptome after RBP perturbation is traditional steady-state RNA sequencing, which can reveal changes in alternative splicing and promoter usage; gene expression levels (from which RBP regulation of mRNA stability can be inferred); and RNA editing in the RBP-perturbed cells compared to control cells. Alternative assays to probe different aspects of RNA metabolism after RBP perturbation include ribosome profiling to measure changes in translation (Ingolia et al. [2009]); RNA-seq on different subcellular fractions to measure changes in mRNA localization (Lefebvre et al. [2017]); techniques such as 3P-seq (Jan et al. [2011]) to measure changes in poly(A) site usage and TAIL-seq (Chang et al. [2014]) or PAL-seq (Subtelny et al. [2014]) to measure changes in poly(A) tail length; and metabolic labeling or transcriptional inhibition followed by time course measurements to more directly measure mRNA half-lives (Tani and Akimitsu [2012]). A limitation of RBP perturbation studies is that it is often difficult to disambiguate direct from indirect changes (e.g., whether the RBP in question directly affects splice site choice or its perturbation changes the level of a splicing factor that is responsible for an observed splicing change), though confidence in identifying directly regulated events can be achieved by integrating binding (e.g., *in vivo*-based CLIP or *in vitro*-based RBNS) data and considering events that have evidence of direct RBP interaction.

1.2.4 Structural studies of RBPs and RBP-RNA interactions

The two most common methods for determining high-resolution atomic structures of proteins and protein-RNA complexes are nuclear magnetic resonance (NMR) spectroscopy and

X-ray crystallography. Since the first protein-RNA complex was solved using X-ray crystallography (Chen et al. [1989]), improvements in instrumentation and computational modeling techniques have led to these structural studies being incredibly influential in revealing information about RBP-RNA interactions. Overviews of the two techniques and and differences between them include:

- NMR structures represent an average over semi-random oriented molecules tumbling in solution over a nanosecond to second time scale (Brünger [1997]). Proteins $\leq$ 30 kDa ($\sim$ 270 amino acids) are amenable to NMR spectroscopy. As they occur in solution, NMR studies of proteins are closer to their physiological state and allow identification of flexible portions of the protein including interdomain linker sequences commonly observed in multi-domain RBPs. NMR studies also allow conformational changes to be observed, such as those that may occur upon RNA binding to a multi-RBD protein (Afroz et al. [2015], Mackereth et al. [2011]).

- X-ray crystallography structures represent an average over molecules arranged in a regular crystal lattice over a seconds to hours time scale (Brünger [1997]). X-ray crystallography is also able to provide the position of water molecules in the structure, allowing prediction of water-mediated hydrogen bonds which may be important in protein-RNA specificity (Afroz et al. [2015]). It can be applied to proteins or complexes $>$ 100 kDa, though obtaining a crystal is not guaranteed and is often time-consuming even when possible.

  As of October 2017, there were 230 structures of human RRMs in the Protein Data Bank (140 solution NMR and 90 X-ray crystallography; 174 RRM(s) alone and 56 in complex with RNA). There were 58 structures of human KH domains in the PDB, 27 solution NMR and 31 X-ray crystallography, 13 of which are in complex with RNA or DNA.

1.3 RNA secondary structure in RBP-RNA interactions and regulation

Deciphering the structures of complex three-dimensional biomolecules is essential to fully understand their regulatory capacity and biological function. In contrast to double-stranded DNA, RNA is single-stranded, permitting it to fold into complex secondary and tertiary structures that can directly influence RNA regulatory capacity or alter the ability of RBPs to bind *cis*-sequence elements. Single nucleotide polymorphisms (SNPs) in UTRs that induce RNA conformational changes have been associated with six genetic diseases, underscoring the importance of considering secondary structure in understanding RNA function (Halvorsen et al. [2010]).

RNA structures can be revealed through three complementary approaches: *in silico* folding as well as *in vitro* or *in vivo* profiling studies. *In silico*-folded structures are typically predicted from dynamic programming algorithms that efficiently search the set of all possible structures, except for those containing pseudoknots (Eddy [2004]). The prediction(s) for an RNA can include the minimum free energy (MFE) structure, which is the most probable structure at equilibrium, or a set of structures with associated probabilities based on the partition function that describes the ensemble thermodynamic properties of the system (**Fig.** 1-4). Commonalities at particular RNA bases over the latter set of structures can provide an estimate of the quality of the prediction and identify highly probable base pairs (Mathews et al. [2010]). Benchmarking of *in silico* RNA structure prediction from sequence alone for rRNAs and other well-studied catalytically active RNAs yielded an ∼70% base pair accuracy (Hajiaghayi et al. [2012]), with the accuracy dropping for sequences 1 kb or longer such as full-length mRNAs (Doshi et al. [2004]).

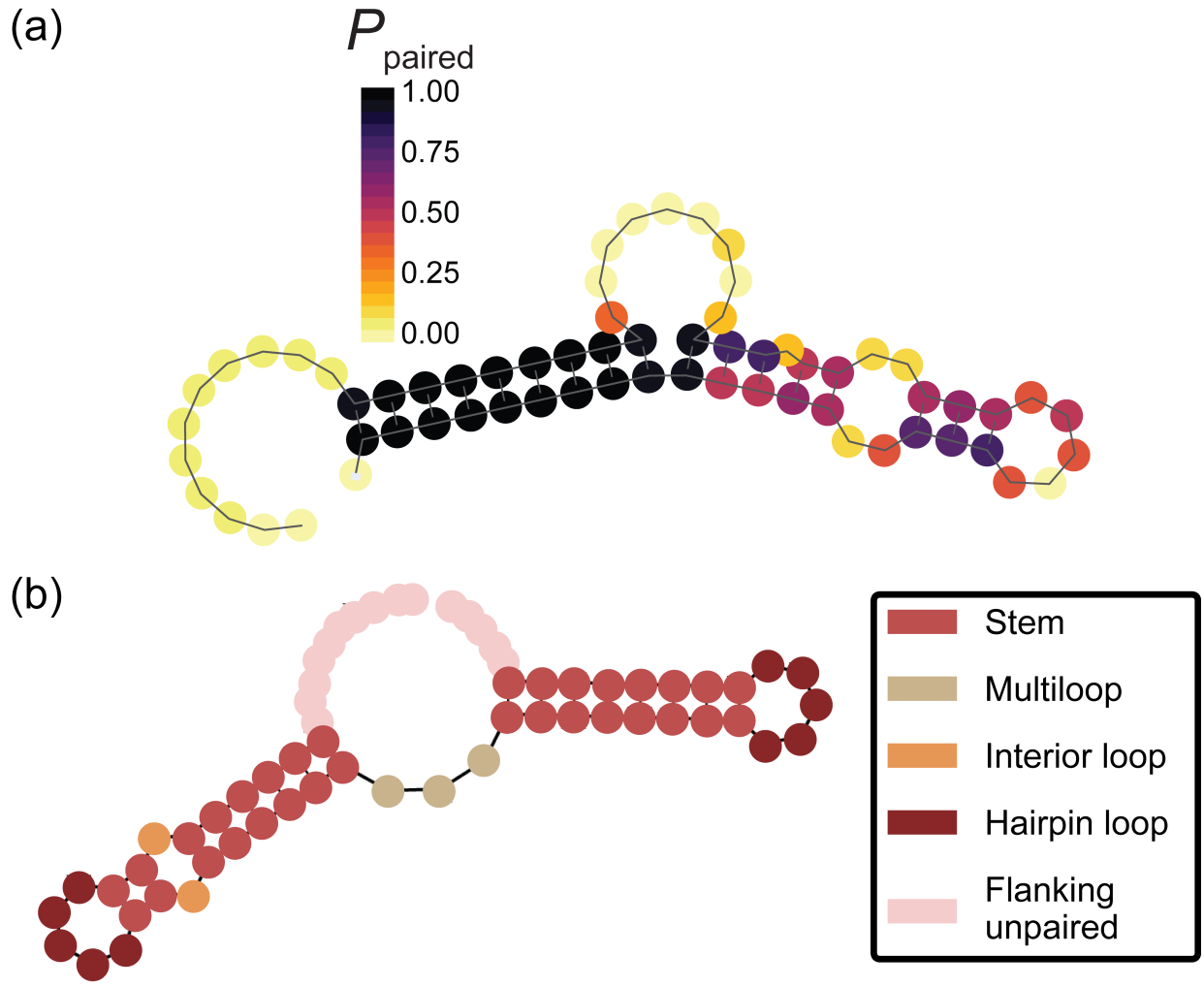

Figure 1-4: **Examples of *in silico*-folded RNA oligos**
(**a**) Example of an RNA oligo folded *in silico* with RNAfold (Lorenz et al. [2011]), with the probability of each RNA base being paired (P_{paired}) in the ensemble of predicted structures according to their thermodynamic energies.
(**b**) Example of the RNA bases of the minimum free energy structure of an RNA oligo classified into five secondary structural elements.

Improvements to *in silico*-predicted structures can be made by considering secondary structures common to homologous RNA sequences. As RNAs whose biological function depends on their secondary structure (e.g., rRNAs, tRNAs) should be structurally conserved across related RNAs from different species, covariation with compensatory base pair changes that preserve not primary sequence but rather secondary structure (e.g., G-C mutating to A-U) is often observed in such RNAs and is generally more conserved than primary sequence (Seetin and Mathews [2012]). While covariation sequence analysis is the most accurate measure of secondary structure prediction with >95% of predicted pairs correct (Gutell et al. [2002]), multiple homologous sequences, as well as a high quality alignment of them, are required, which is not always possible for rare RNAs or RNAs with high sequence variability (Mathews et al. [2010]). Comparative sequence analysis and identification of covarying sites is often a manual undertaking, though a number of software packages are available to generate hypotheses, with methods that predict the conserved structure for all homologous sequences together typically being most accurate (Seetin and Mathews [2012]).

Experimentally, RNA structures and folding have predominantly been studied in dilute *in vitro* conditions, leading to fundamental insights. However, *in vitro* experiments are typically lacking the biological ions, ligands, proteins, and crowding that affect folding and function *in vivo*. Up to 40% of the cytosol is taken up by macromolecules, making crowding a critical but poorly understood contributor to RNA folding *in vivo*. While *in vivo* studies that utilize structure probing agents are thus desirable due to their biological relevance, they generally only measure the ensemble structure of each RNA species, do not provide any information about the kinetics or thermodynamics of folding, and do not disambiguate RNA structure formed from pairing with itself and inaccessibility due to protein-RNA interactions (Leamy et al. [2016]).

Major themes from the study of thermodynamic and kinetic studies *in vitro* over the past decades include that RNAs fold on rugged pathways in which pathway intermediates are populated and RNA can form misfolded structures before populating the native secondary structure state (Leamy et al. [2016]). Additionally, secondary structures form before tertiary contacts, at least in the context of tRNAs, ribozymes, and riboswitches (Leamy et al. [2016]). For these RNAs, secondary structures form on a relatively quick timescale

(on the order of μs-ms) followed by the slower folding of the tertiary structure (Crothers et al. [1974]). Quantitative analyses of the melting temperatures of short RNA duplexes led to improved experimental parameters for RNA base pairs, helices with loops, unpaired terminal nucleotides, and hydrogen bonds known as the “Turner Rules" (Turner et al. [1987]). This study also demonstrated that the two major fundamental interactions in RNA, base stacking and hydrogen bonding, contribute roughly similarly to free-energy changes in oligoribonucleotide association such that both are likely to be important in three-dimensional RNA structure prediction and interpretation of RNA-RNA associations. Canonical A-U and G-C base pairs and the wobble G-U pair engage the canonical Watson-Crick base edges to form two, three, and two hydrogen bonds, respectively, with the Turner rules and likewise ‘nearest-neighbor’ models assuming the stability of a base pair or other RNA structure is dependent only on the identity of the adjacent base pairs.

Structure probing chemical and enzymatic methods *in vitro* and *in vivo* followed by high-throughput sequencing readouts have recently been utilized to elucidate the structure of RNAs with nucleotide resolution. Among the popular chemical probes used to attack and modify the solvent-accesible RNA bases or sugars to cause reverse transcriptase dropoffs are DMS (modifies unpaired adenines and cytosines; Structure-seq, Ding et al. [2014]; DMS-seq, Rouskin et al. [2014]; and Mod-seq, Talkish et al. [2014]) and SHAPE (modifies the 2’-hydroxyl of all unpaired bases, Lucks et al. [2011]). Commonly used enzymatic probes are the RNases S1 (cleaves single-stranded RNA) and V1 (cleaves double-stranded RNA) in combination (PARS, Underwood et al. [2010]) or RNase P1 (cleaves single-stranded RNA) alone along with an untreated sample (FragSeq, Kertesz et al. [2010]). Among the findings of recent studies using the aforementioned structure probing methods include that there is significantly more structure in coding regions than in UTRs of yeast RNAs (Kertesz et al. [2010]) in contrast to humans in which the CDS is slightly more single-stranded than UTRs (Wan et al. [2014]), and there is less structure at the start and stop codons than at other transcript areas, facilitating read-through of the ribosome (Ding et al. [2014], Wan et al. [2014]).

RNA has been shown to be less structured *in vivo* than *in vitro* or folded *in silico*, likely due to the translocating ribosome, RNA helicases, and other RBPs unwinding the

RNA (Rouskin et al. [2014]). Upon *in vivo* ATP depletion, yeast mRNAs become more structured, implying that ATP-dependent processes unfold RNAs in cells. Integration of a click-chemistry based SHAPE experiment (icSHAPE) with RBFOX2 iCLIP binding sites in mouse embryonic stem cells revealed that *in vitro* vs. *in vivo* differential icSHAPE signal over RBFOX2's UGCAUG motif matched the key RNA residues involved in the RBFOX2-RNA interaction (Spitale et al. [2015]). Similar analysis identified peaks of structural arrangement at iCLIP-bound HuR (ELAVL1) binding sites and enabled reasonably accurate HuR binding site prediction from icSHAPE data alone. Combining structure probing data with an RBP's motif occurrences in the transcriptome may thus collectively boost prediction accuracy of RBP binding sites in the cell type in which the structure probing was performed and provide insight into RNA structural rearrangements upon RBP binding.

The secondary structure and accessibility of mRNA sites has been shown to affect RBP binding and post-transcriptional regulation at a transcriptome-wide scale. miRNAs mediate less repression in the middle of 3' UTRs than near the end of 3' UTRs, presumably due to the decreased accessibility of the middle sites due to occlusive structures (Grimson et al. [2007]). RNA structure affects recognition of ESEs by SR proteins in the fibronectin EDA exon and contributes to differential regulation of this exon between human and mouse (Buratti et al. [2004]), and this finding was generalized in an analysis of experimentally identified SREs showing they are significantly enriched in single-stranded pre-mRNA regions (Hiller et al. [2007]). Additionally, a discrepancy arose when examining the effects of 8mer 3' UTR regulatory elements determined through a high-throughput cell-based screen versus their effect when tested within the context of a larger 500 nt endogenous sequence through luciferase reporters (Wissink et al. [2016]). Five of the seven investigated *cis*-elements (including both stabilizing and destabilizing elements) had opposing effects among the 3-5 endogenous 3' UTR sequences tested, highlighting the importance of sequence context on regulatory element function. One likely explanation for the discrepancy is that the endogenous sequence context puts the regulatory elements within a new local secondary structure that prevents them from being accessible to RBPs or changes the folding of the isolated 8mers (Mayr [2017]).

In addition to the general accessibility of RNA sites modulating RBP binding, specific

RNA structures play an important role in recruiting or inhibiting certain RBPs to effect biological functions. For example, SLBP (stem-loop binding protein) binds the characteristic stem-loop structure near the end of replication-dependent histone pre-mRNAs necessary for efficient 3'-end processing of these messages (Wang et al. [1996]). The selenocysteine insertion element is an ∼60 nt specific stem-loop structural motif in the 3' UTRs of particular mRNAs that causes the UGA stop codon to be translated as selenocysteine (Walczak et al. [1996]). Long-range RNA duplexes in 3' UTRs are bound by STAU1 to regulate message stability (Sugimoto et al. [2015]), and ADAR proteins specifically deaminate adenosines in duplex RNA to carry out A-to-I editing while leaving other adenosines unmodified (Li et al. [2009]). Alternative splicing, too, can be modulated through highly regulated and conserved structures; long-range RNA-RNA base-pairing interactions are necessary for some RBFOX binding events in distal introns to be recruited to alternative exons to productively enhance spliceosomal activity (Lovci et al. [2013]). In summary, RNA secondary structure can greatly influence RBP binding and its consideration is essential in our efforts to more fully understand RBP-mediated regulation of RNA processing.

1.4 Overview of the thesis

This thesis details the computational methods developed and insights learned from high-throughput sequencing-based assays applied to *in vitro* and *in vivo* RNA-protein interactions. In Chapter 2, I characterize the RNA binding specificities of 78 diverse human RNA binding proteins assayed through a high-throughput technique, RNA Bind-n-Seq (RBNS), developed in the Burge laboratory. A resource of the primary sequence, secondary structure, and other RNA contextual preferences of the RBPs is provided, along with novel integration of these motifs with functional perturbation data (RBP knockdown followed by RNA-seq) to infer context-specific maps of RBP regulation of alternative splicing and mRNA stability. Additionally, a framework that incorporates these contextual features layered on top of primary sequence motifs to more fully characterize RBP specificity is presented with a quantitative comparison of these contextual preferences across RBPs. In Chapter 3, I present methods for integrating *in vitro* RBP specificities with the enhanced crosslinking

and immunoprecipitation (eCLIP) assay that maps RBP binding sites in cells. Further integrative analysis of five assays, each focused on a distinct aspect of RBP activity (eCLIP; RBP knockdown/RNA-seq; RBNS; RBP subcellular localization; and RBP association with chromatin), is presented to expand the catalog of functional RNA elements encoded in the human genome and understand their processing and regulation by RBPs in cells.

Chapter 2

Sequence, Structure, and Context Preferences of Human RNA Binding Proteins

Under review, posted to bioRxiv on 10/12/17:

D Dominguez[§], P Freese[§], MS Alexis[§], A Su, M Hochman, T Palden, C Bazile, NJ Lambert, EL Van Nostrand, GA Pratt, GW Yeo, B Graveley, CB Burge. "Sequence, Structure and Context Preferences of Human RNA Binding Proteins".
https://doi.org/10.1101/201996

My contributions:

Development of RBNS computational pipeline for assay quality control and generation of enrichment values and sequence motif logos (**Fig.** 2-1, **Fig.** 2-S1**A**, **Fig.** 2-2**A**); comparison to RNAcompete in conjunction with DID (**Fig.** 2-S1**C, D**); analysis of RNA motif properties (**Fig.** 2-2**C, E, F**; **Fig.** 2-S2); generation of RNA maps from RBNS and knockdown/RNA-seq data and comparison with *in vivo* binding (**Fig.** 2-3**C, F**; **Fig.** 2-S3**A-C, E**); overlap with splicing and stability regulatory 6mers in conjunction with DID and MSA (**Fig.** 2-3**A, B, D, E**); RNA secondary structure analyses (**Fig.** 2-4; **Fig.** 2-S1**B**; **Fig.** 2-S4; **Fig.** 2-S6**E,**

H, I); comparison of contributions of context features and context feature preferences *in vivo* in conjunction with MSA (**Fig.** 2-6; **Fig.** 2-S6**A-D**); tissue specificity of RBP expression (**Fig.** 2-S6**F**); binding similarity among RBPs with similar vs. different RBD types (**Fig.** 2-S6**G**); writing of text in conjunction with DID, MSA, and CBB.

2.1 Abstract

RNA binding proteins (RBPs) orchestrate every step of the production, processing, and function of mRNAs. Here we present the affinity landscapes of 78 human RBPs using an unbiased assay that determines the sequence, structure, and context preferences of an RBP *in vitro* by deep sequencing of bound RNAs. These data enable construction of "RNA maps" of RBP activity without requiring crosslinking-based assays. We observed an unexpectedly low diversity of RNA motifs, implying frequent convergence of binding specificity toward a relatively small set of RNA motifs, many with low compositional complexity. Offsetting this trend, we observed extensive preferences for contextual features distinct from short linear RNA motifs, including spaced 'bipartite' motifs, biased flanking nucleotide composition, and bias away from or towards RNA structure. Our results emphasize the importance of these contextual features in RNA recognition, which likely enable targeting of distinct subsets of transcripts by different RBPs that recognize the same linear motif.

2.2 Introduction

RNA binding proteins (RBPs) control the production, maturation, localization, modification, translation, and degradation of cellular RNAs. Many RBPs contain well-defined RNA binding domains (RBDs) that engage RNA in a sequence- and/or structure-specific manner. The human genome encodes at least 1500 RBPs that contain established RBDs, the most prevalent of which include RNA recognition motifs (RRM, ∼240 RBPs), hnRNP K-homology domains (KH, ∼60 RBPs) and C3H1 zinc-finger domains (ZNFs, ∼50 RBPs) (reviewed by Gerstberger et al. [2014]). While RBPs containing RRM (Query et al. [1989]) or KH domains (Siomi et al. [1993]) were first described over two decades ago, the repertoires of RNA sequences and cellular targets bound by different members of these and other classes of RBPs are still largely unknown.

Structural studies have identified conserved residues that enable canonical RBP-RNA interactions but have also uncovered non-canonical binding modes, making it difficult to infer RNA target preferences from amino acid sequence alone (reviewed by Cléry and Allain [2011], Valverde et al. [2008]). For example, RRMs adopt a structure with an antiparallel four-stranded beta sheet packed onto two alpha helices, with the two central strands (RNP1 and RNP2) typically mediating interactions with RNA (reviewed by Afroz et al. [2015]). However, crystallography and NMR studies have shown that certain RBPs bind RNA via the linker regions, loops, or the C- or N- terminal extremities of their RRMs rather than the canonical RNP1 and RNP2 strands (reviewed by Daubner et al. [2013]). Similarly, KH domains form a hydrophobic binding cleft that is generally thought to accommodate a pyrimidine-rich tetranucleotide motif, but specificity is often modulated by hydrogen bonding or additional interactions with the protein backbone (Grishin [2001],Valverde et al. [2008]). These variable RNA binding mechanisms in combination with the presence of multiple RBDs in most RBPs (Lunde et al. [2007]) have motivated efforts to experimentally interrogate the specificity of individual RBPs (reviewed by Cléry and Allain [2011]).

Several methods exist for determining RBP binding sites *in vivo*, most notably RNA immunoprecipitation (RIP, Gilbert and Svejstrup [2006]) and UV crosslinking followed by immunoprecipitation (CLIP) and sequencing (Ule et al. [2003]). While such techniques cap-

ture RBP-RNA interactions in their cellular contexts, it is often difficult to derive motifs from these experiments due to interactions with protein cofactors, high levels of non-specific background (Friedersdorf and Keene [2014]), and non-random genomic composition. Quantitative *in vitro* assays such as electrophoretic mobility shift assay (EMSA), surface plasmon resonance (SPR), and isothermal calorimetry (ITC) must be guided by a priori knowledge of putative RNA substrates, making them unsuitable for high-throughput motif discovery. Methods such as SELEX (systematic evolution of ligands by exponential selection) typically select a few high-affinity 'winner' sequences, but generally do not reveal the full spectrum of RNA targets or their associated affinities (reviewed by Cook et al. [2015]). RNAcompete is a high-throughput *in vitro* binding assay that captures a more complete specificity profile by quantifying the relative affinity of an RBP for a pre-defined set of ∼250,000 RNA molecules (Ray et al. [2013]). One limitation of this approach is that the designed RNAs present motifs in a relatively small range of predominantly unstructured contexts, restricting the analysis to short, mostly unpaired motifs. More recent approaches such as RNA Bind-n-Seq (RBNS) (Lambert et al. [2014]) and RNAcompeteS (Cook et al. [2017]) perform high-throughput sequencing of bound RNAs selected from a random pool, yielding a more comprehensive profile of the sequence and RNA secondary structural specificity of an RBP.

The RNA binding specificity of ∼100 human RBPs has been assessed using various unbiased (*de novo*) methods (Giudice et al. [2016]), though the diversity of techniques employed hampers comparison of their specificities. To systematically explore the binding specificity spectrum of humans RBPs at high resolution, we performed RBNS on a diverse set of 78 human RBPs, half of which had previously uncharacterized specificities. RBNS comprehensively and quantitatively maps the RNA binding specificity landscape of an RBP through a one-step *in vitro* binding reaction using recombinant RBP incubated with a random pool of RNA oligonucleotides (Lambert et al. [2014]). The assay was typically carried out for each RBP at five protein concentrations totaling 400 binding assays. The depth of sequencing yielded over 6 billion protein-associated reads, enabling detection not only of simple sequence motifs but also of preferred structural and contextual features (**Fig.** 2-1**A**). Analysis of these data revealed a pattern in which many proteins bind to similar motifs, but differ in their preferences for additional binding features such as RNA secondary structure, flanking nu-

cleotide composition, and bipartite motifs, expanding their ability to distinguish regulatory targets.

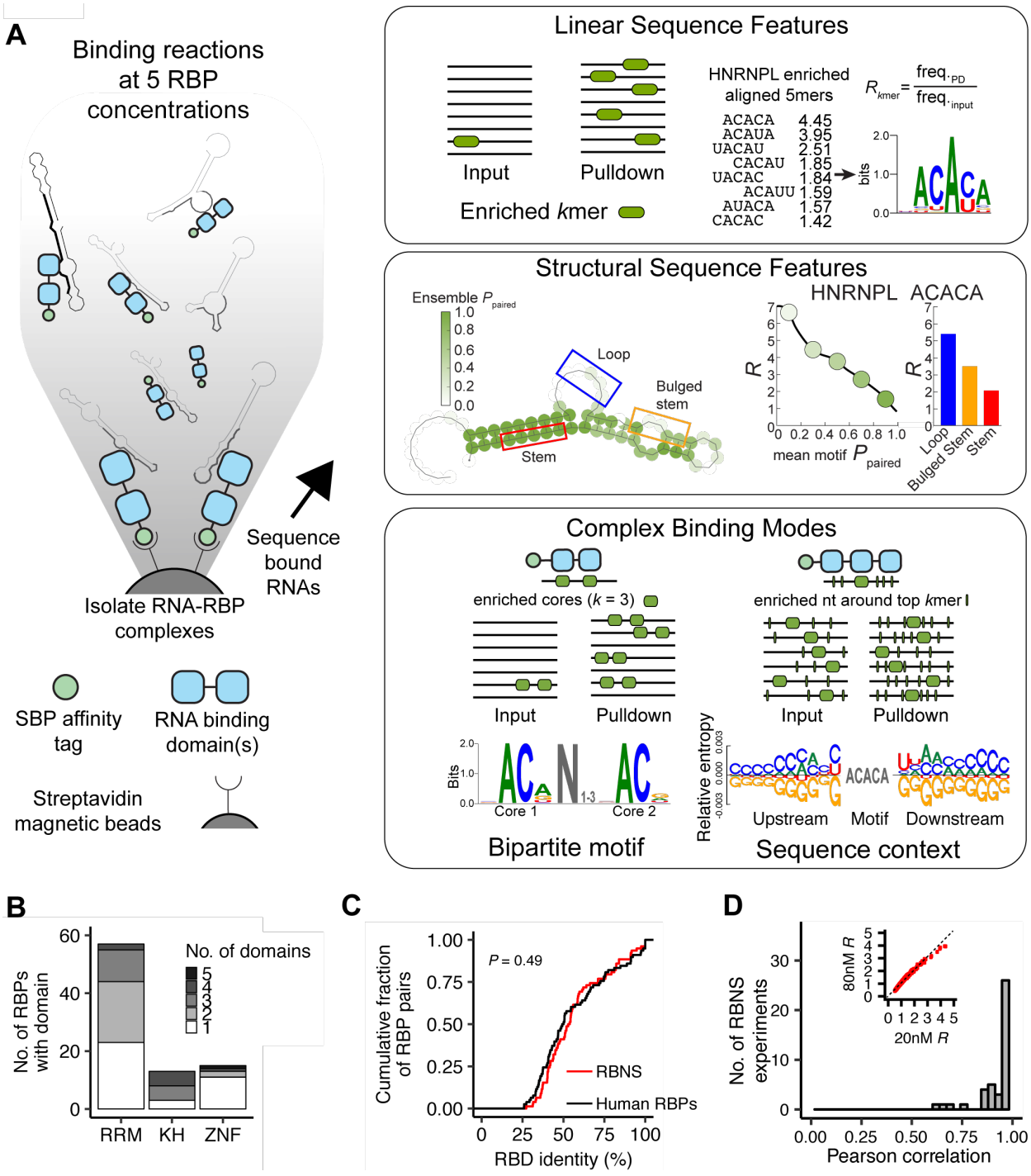

Figure 2-1: **Overview of the high-throughput RNA Bind-n-Seq assay and computational analysis pipeline**

Figure 2-1

(**A**) Schematic of RBNS assay and pipeline. Recombinant RBPs are incubated with a pool of random RNA (black) flanked by adapter sequences (gray). RBP-RNA complexes are isolated with streptavidin magnetic beads, eluted with biotin, and bound RNA is sequenced. Computational analysis of pulldown and input reads reveals linear sequence specificities, secondary structure preferences, and complex binding modes of RBPs.
(**B**) Number of RBPs with one or more of the three most common RBD types assayed.
(**C**) Cumulative distribution of the RBD amino acid identity between each RBP and its most similar analyzed RBP. Distributions are calculated separately for the set of RBPs that has been assayed by RBNS and all other human RBPs (sampled to match the domain distribution analyzed by RBNS). Only domains with 5+ RBPs assayed by RBNS are included (RRM, KH, Zinc finger CCCH-type).
(**D**) Histogram of Pearson correlations between RBNS assays of the same RBP at different protein concentrations. Inset: correlation of 5mer R values of HNRNPL at 20 nM (most enriched concentration) and 80 nM.

2.3 Results

2.3.1 High-throughput RNA Bind-n-Seq Assay

To determine the detailed binding preferences of a large set of human RBPs we developed a high-throughput version of RNA Bind-n-Seq (RBNS), an *in vitro* method capable of determining the sequence, structure, and context preferences of RBPs. In this assay, randomized RNA oligonucleotides (20 or 40 nt) flanked by constant adapter sequences were synthesized and incubated with varying concentrations of an SBP-tagged recombinant protein containing the RBD(s) of an RBP (**Fig.** 2-1**A**, constructs listed in Table S1). RNA-protein complexes were isolated with streptavidin-conjugated affinity resin, washed, and bound RNA was eluted and prepared for deep sequencing. Protein purification, binding assays, and sequencing library preparations were carried out in 96-well format increasing scalability and consistency across experiments (Methods). A typical experiment yielded ∼10-20 million unique reads at each protein concentration, which were compared to the input RNA pool sequenced to similar depth (**Fig.** 2-S1**A**, Table S2). Inclusion of sequencing adapters flanking the randomized RNA region simplified library preparation, preventing ligation biases and amplification of contaminating bacterial RNA carried over from protein purification (Lambert et al. [2014]). Furthermore, as RBPs bind RNA motifs in a wide range of structural contexts *in vivo* (Fukunaga et al. [2014]), RBNS presents RBPs with motifs spanning a broad spectrum of secondary structures, exceeding that of similar reported methods (Cook et al. [2015]) (**Fig.** 2-S1**B**). The high sequence complexity of the interrogated libraries enabled the fine dissection of RNA binding preferences, while use of multiple protein concentrations increased reliability and enabled detection of lower-affinity motifs.

2.3.2 Binding specificities of a diverse set of human RNA binding proteins

RBNS was performed on a fairly diverse set of 78 human RBPs containing a variety of types and numbers of RBDs (**Fig.** 2-1**B**). RBPs were chosen based on a combination of criteria, including: presence of well-established RBDs; evidence of a role in RNA biology (though

this was not required); and secondary criteria related to expression in ENCODE cell lines K562 and HepG2 and availability of validated antibodies for complementary eCLIP analysis (Sundararaman et al. [2016]). Comparing all RBDs in this set, the range of amino acid identity was similar to that of human RBPs overall (**Fig.** 2-1**C**). Together, this set captures a broad swath of proteins that is reasonably representative of human RBPs.

To assess the sequence specificity of each RBP, we developed a computational pipeline that calculates enrichment ("R") values of $k$mers (for k in the range 3-8 nt), where R is defined as the frequency of a $k$mer in protein-bound reads over its frequency in input reads (**Fig.** 2-1**A**, top right). In most cases, R values of top $k$mers exhibited a unimodal profile with increasing protein concentration consistent with increased signal above noise at moderate versus low RBP concentrations, and increased binding of lower-affinity motifs at higher versus moderate concentrations (Lambert et al. [2014]). A mean Pearson correlation across 5mers of 0.96 was observed among experiments performed on the same RBP at different concentrations, indicating high reproducibility (**Fig.** 2-1**D**). A comparison of previously reported binding specificities for 31 factors also assayed by RNAcompete (Ray et al. [2013]) using an independent array-based assay revealed high correlation with our dataset (**Fig.** 2-S1**C-D**, mean Pearson correlation = 0.72), with only four proteins showing correlations below 0.5.

2.3.3 Overlapping specificities of RNA binding proteins

To visualize and compare the primary sequence specificities of the assayed RBPs, we derived sequence motif logos for each RBP by aligning enriched 5mers (Z-score $\geq$ 3, weighted by enrichment above input using an iterative procedure that avoids overlap issues, **Fig.** 2-1**A** top right, Methods). For more than half of the RBPs (41/78), this yielded multiple sequence logos, indicating affinity to multiple distinct motifs that may reflect different binding modes or binding by different RBDs (motif 5mers are listed in Table S3). As expected, the motifs of closely related paralogs (e.g., PCBP1/2/4, RBFOX2/3) clustered closely together (Conway et al. [2016], Smith et al. [2013]) (**Fig.** 2-2**A**). However, unexpectedly, some completely unrelated proteins, often containing distinct classes of RBDs, were also grouped together. Inspection of the dendrogram revealed 15 or so clusters of RBPs with highly similar primary

motifs (nine with three or more members), leaving 18 RBPs with motifs more distinct from other profiled RBPs unclustered (Methods). Notably, eight of the 15 clusters contained two or more proteins with completely different types of RBDs (e.g., cluster 1 contained RRM-, KH- and ZNF-containing proteins as well as factors with multiple RBD types).

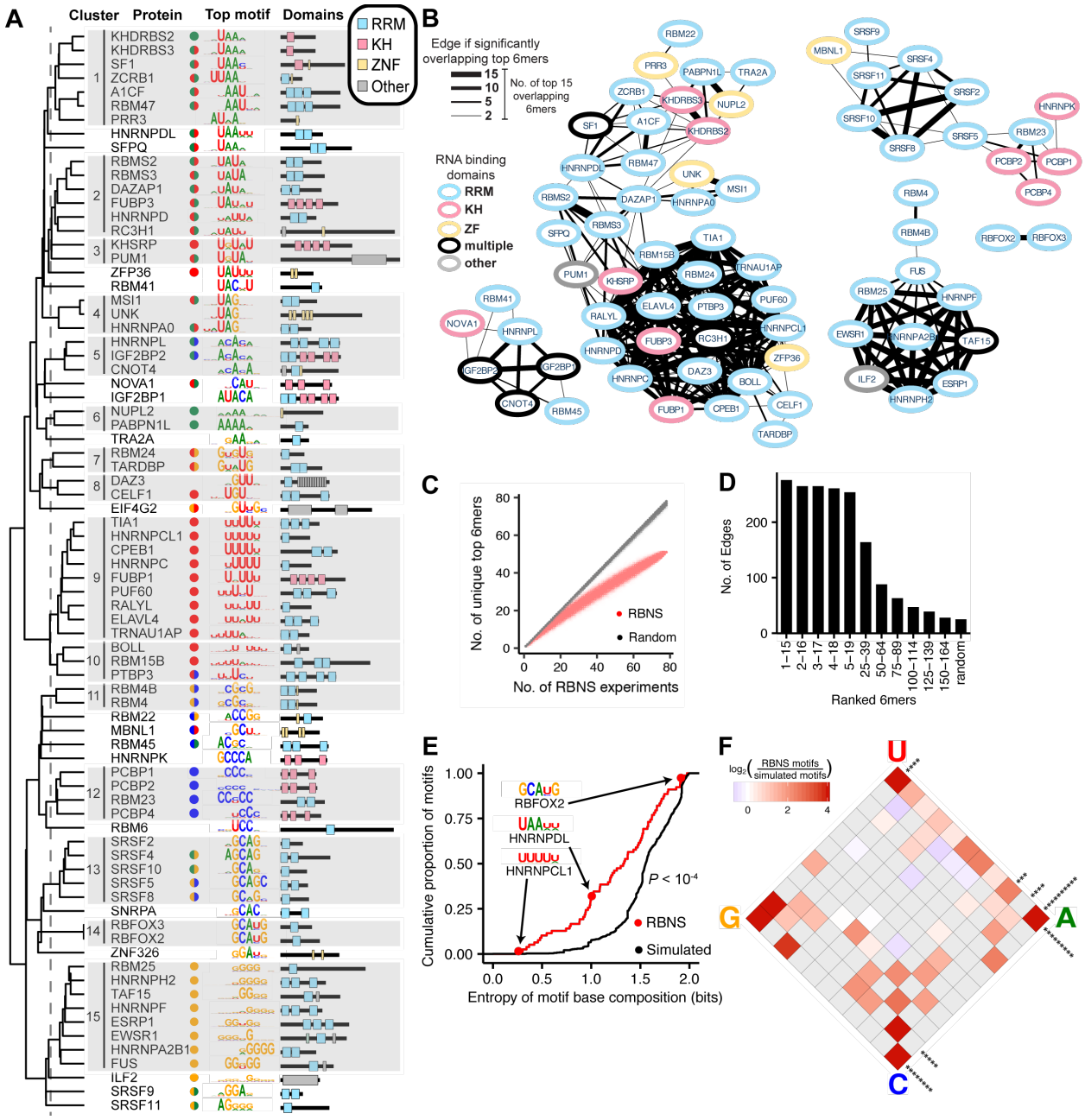

Figure 2-2: **RBPs bind a small subset of the sequence space, characterized by low-entropy motifs**

Figure 2-2

(**A**) From left to right: Dendrogram of hierarchical clustering of RBPs by sequence logo similarity and 15 clusters at indicated branch length cutoff (dashed line); protein name; colored circles representing nucleotide content of RBP motif (one circle if motif is >66% one base, two half-circles if motif is >33% two bases); top motif logo for each protein; protein RBD(s). Each logo represented an average of seven 5mers.
(**B**) Network map connecting RBPs with overlapping specificities (sharing at least two of the top 15 RBNS 6mers). Line thickness increases with number of overlapping 6mers as indicated. Node outline indicates RBD type of each protein.
(**C**) Number of unique top 6mers among subsamplings of the 78 RBNS experiments versus randomly selected 6mers.
(**D**) Edge count between nodes for network maps as shown in **B**, drawn using groups of 15 6mers with decreasing ranks.
(**E**) Entropy of nucleotide composition of RBNS motifs vs. simulated motifs (Methods). *P*-value determined by Wilcoxon rank-sum test.
(**F**) Enrichment of RBNS motifs over simulated motifs among partitions of a 2D simplex of the motif nucleotide composition (Methods). Significance along margins was determined by bootstrap Z-score (number of asterisks = Z-score).

To more rigorously assess the relatedness of RBP binding affinities, we generated a network map with edges connecting RBPs with significantly overlapping sets of top 6mers (at least two of the top fifteen 6mers, $P = 0.001$, hypergeometric test, **Fig.** 2-2**B**). While RBFOX2 and RBFOX3 (which bind GCAUG) were connected only to each other, many proteins were members of larger highly connected groups and the network overall was much more connected than expected ($P < 10^{-5}$ relative to null distribution of RBPs binding all sequences equally, Methods). Indeed, for 27 RBPs the highest ranked 6mer was also the highest ranked 6mer of at least one other RBP. Given that there are 4096 distinct 6mers, this large overlap among highest-affinity 6mers was highly significant compared to the ∼1 overlap expected if RBPs bound all sequences equally (**Fig.** 2-2**C**). A large excess of overlaps also occurred when considering the top fifteen 6mers for each RBP. Furthermore, the excess of overlaps remained when eliminating paralogs and any pairs of proteins containing RBDs sharing at least 40% amino acid identity (**Fig.** 2-S2**A**-**B**). To explore the network's connectivity further, we regenerated network maps with sets of 6mers with progressively decreasing affinities (e.g., 6mers ranked 2-16, 13-17, etc. for each RBP). A monotonic decrease in edges (overlaps of two or more) was observed with decreasing affinity categories, indicating that the connectivity of this RBP map is highest for the 6mers with the highest relative affinity (**Fig.** 2-2**D**). Together, these observations indicate that RBPs recognize a much smaller subset of the available sequence space than expected. The pattern of clustering and overlap of motifs observed in **Fig.** 2-2**A**-**C**, including many clusters of RBPs with distinct RBD types, suggests that unrelated RBPs have evolved to bind similar RNA sequence motifs many times.

2.3.4 RBPs preferentially bind low-complexity motifs

We noted that most motifs were primarily composed of just one or two distinct bases (**Fig.** 2-2**A**). To assess motif composition objectively, we measured the Shannon entropy of the nucleotide composition of each sequence logo, a scale which ranges from 0 bits (homopolymers) to 2 bits (25% of each base). The entropies of actual RBP motifs were substantially lower than control simulated motifs made from sampling motif columns across RBPs ($P < 10^{-4}$, Wilcoxon rank-sum test, Methods), indicating that RBP motifs are biased toward lower compositional complexity (**Fig.** 2-2**E**). This trend applied generally to all compositions with

low complexity, as mapping RBP motif compositions onto a 2-dimensional simplex revealed increased density at all four mononucleotide 'corners', as well as all 6 dinucleotide 'margins' (with A/C, A/U, and C/U most significant, **Fig.** 2-2**F**, **Fig.** 2-S2**C**, all bootstrap $P < 0.05$).

2.3.5 RNA maps from RBNS and knockdown RNA-seq data

Many human RBPs are involved in pre-mRNA splicing. Therefore, it was not unexpected that ∼35% of the 596 'RBNS 6mers' (union of the top fifteen 6mers for all RBPs) matched 6mer splicing elements identified previously in cell-based reporter screens (**Fig.** 2-3**A**, $P = 1.7 \times 10^{-4}$, hypergeometric test) (Ke et al. [2011], Rosenberg et al. [2015], Wang et al. [2012], Wang et al. [2013]) consistent with common roles of the studied RBPs in splicing. The RBNS 6mers conferred stronger regulation than non-RBNS 6mers (**Fig.** 2-3**B** left, $P = 6 \times 10^{-36}$, Wilcoxon rank-sum test), and higher RBNS enrichment (reflecting higher affinity) was associated with increased splicing activity (**Fig.** 2-3**B** right for binned comparisons, overall Spearman $\rho = 0.08$, $P < 10^{-12}$).

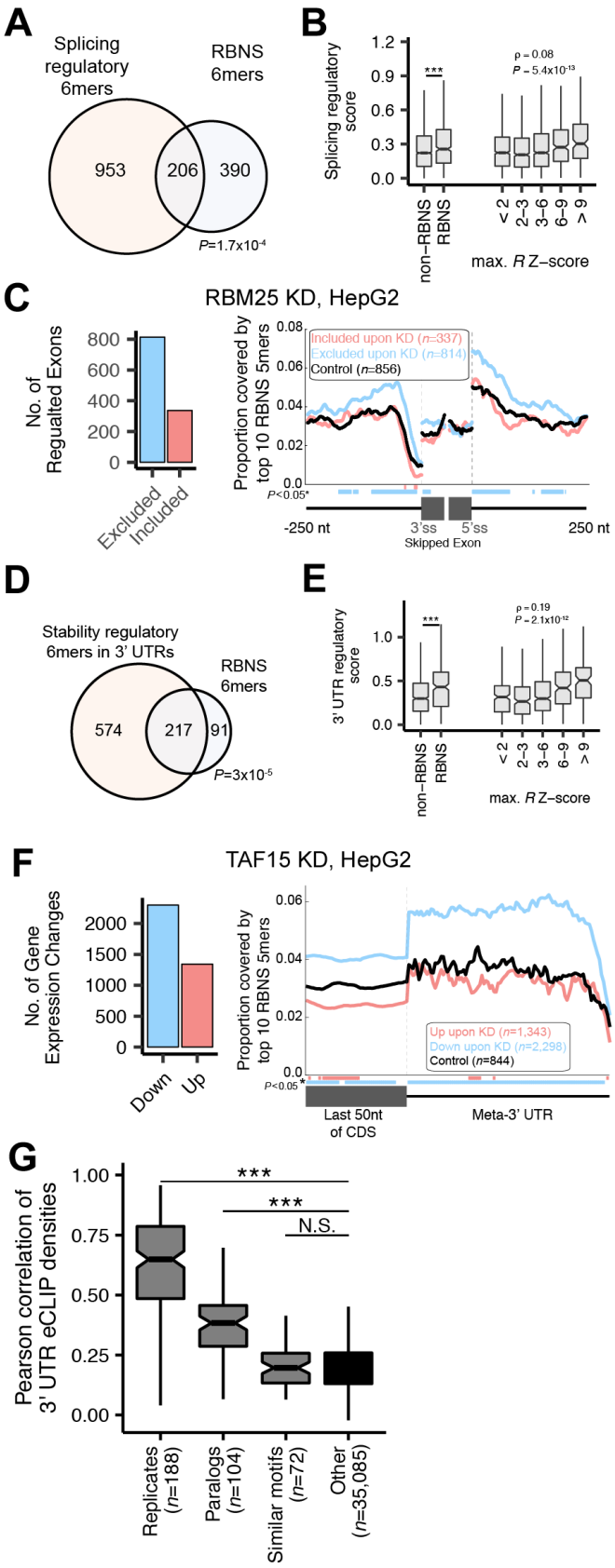

Figure 2-3: **RBNS-derived motifs are associated with regulation of mRNA splicing and stability** *in vivo*

Figure 2-3

(**A**) Overlap of RBNS 6mers and 6mers with splicing regulatory activity (P-value determined by hypergeometric test).

(**B**) Comparison of splicing regulatory scores of, left: RBNS 6mers (top 15 of any RBNS, "RBNS") and all other 6mers ("non-RBNS"); right: all 6mers binned by their maximum R value Z-score across all RBNS experiments (P-values determined by Wilcoxon rank-sum test). Overall Spearman correlation between R Z-score and splicing regulatory score was 0.08 ($P < 10^{-12}$).

(**C**) Left: Number of alternative exons regulated by RBM25 as determined by RNA-seq after RBM25 knockdown in HepG2 cells. Right: Proportion of events covered by RBNS 5mers in exonic and flanking intronic regions near alternative exons excluded upon RBM25 KD (red), included upon RBM25 KD (blue), and a control set of exons (black). Positions of significant difference from control exons upon KD determined by Wilcoxon rank-sum test and marked below the x-axis.

(**D**) Overlap of RBNS 6mers and 6mers with 3' UTR regulatory activity (P-value determined by hypergeometric test among 1303 6mers with sufficient coverage for representation).

(**E**) Comparison of 3' UTR regulatory scores of, left: RBNS 6mers (top 15 of any RBNS, "RBNS") and all other 6mers ("non-RBNS"), P-value determined by Wilcoxon rank-sum test; right: all 6mers binned by their maximum R value Z-score across all RBNS experiments. Overall Spearman correlation between R Z-score and 3' UTR regulatory score was 0.19 ($P < 10^{-11}$).

(**F**) Left: Number of gene expression changes after knockdown of TAF15 in HepG2 cells. Right: Frequency of TAF15 RBNS 5mers along a meta-3' UTR of genes whose expression is decreased (blue), increased (red) or unchanged (black) by TAF15 knockdown. Positions of significant difference from control genes upon KD determined by Wilcoxon rank-sum test and marked below the x-axis.

(**G**) Pearson correlations of eCLIP densities across 100 nt windows of 3' UTRs for all pairs of eCLIP experiments. Pairs of experiments are grouped by category, with all pairs not belonging to 'Replicates', 'Paralogs', or 'Similar motifs' (sharing two of top 5 5mers) placed in 'Other'. P-values determined by Wilcoxon rank-sum test, $^{***}P < 5 \times 10^{-4}$, N.S.= $P > 0.05$.

"RNA maps" for splicing factors have traditionally been built using *in vivo* binding data from CLIP-seq combined with genome-wide assays of splicing changes in response to RBP perturbation (Witten and Ule [2011]). To ask whether *in vitro* data could be used in place of CLIP data to derive such maps of inferred splicing activity, we integrated RBNS data with RNA-seq data from human K562 and HepG2 cells depleted of specific RBPs by shRNA (Van Nostrand et al. [2017]). For example, depletion of RBM25 resulted predominantly in exclusion of cassette exons (**Fig.** 2-3**C**, left). An RNA map for this factor built by assessing the frequency of its top RBNS 5mers near significantly altered exons revealed enrichment of its motif in introns flanking exons that were excluded upon RBM25 knockdown (KD) relative to control introns (**Fig.** 2-3**C**, right). Together, these data support that RBM25 functions as a splicing activator when it binds intronic motifs near alternative exons. This inference is consistent with previous studies of RBM25 regulation of specific alternative exons (Carlson et al. [2017], Gao et al. [2011], Zhou et al. [2008]) and illustrates the potential use of RBNS and RNAi/RNA-seq to inform splicing maps.

By performing this analysis on all 38 RBNS RBPs for which we had KD data in at least one cell type, we observed that 27 of the 38 RBPs showed significant enrichment of their RBNS-derived 5mers in either activated or repressed exons or flanking introns (**Fig.** 2-S3**A**, left). These RNA maps were consistent with previously known patterns of splicing factor activity in many cases, e.g., splicing activation by DAZAP1 (Choudhury et al. [2014]) and PUF60 (Page-McCaw et al. [1999]) and repression by HNRNPC (Choi et al. [1986]) and PTBP1 (Singh et al. [1995]), without use of CLIP. In some cases, RBPs not yet implicated in splicing regulation exhibited splicing maps strongly suggestive of direct function (e.g., ILF2 as a splicing activator). Of note, eight of the nine RBPs with G-rich motifs exhibited splicing activator activity from at least one region, most commonly from introns (FUS being the sole exception). This result matches results of an unbiased screen for intronic splicing enhancers, which identified G-rich *cis*-regulatory sequences and candidate *trans*-acting factors (Wang et al. [2012]). Thus, our approach provides a tool for understanding patterns of splicing regulatory activity by sequence-specific RBPs that augments existing CLIP-based RNA splicing maps (Van Nostrand et al. [2017]).

2.3.6 Protein-bound sequences are associated with *in vivo* regulation of mRNA levels

Besides splicing regulatory activity, we also observed significant overlap between RBNS 6mers and 6mers previously shown to modulate mRNA levels when inserted into reporter 3' UTRs (Oikonomou et al. [2014]) (**Fig.** 2-3**D**, $P = 3 \times 10^{-5}$, hypergeometric test). As observed for splicing regulation, 3' UTR regulatory scores were higher for RBNS motifs (**Fig.** 2-3**E** left, $P = 4 \times 10^{-10}$, Wilcoxon rank-sum test), and regulatory scores increased for 6mers with higher RBNS enrichment (**Fig.** 2-3**E** right for binned comparisons, overall Spearman $\rho = 0.19$, $P < 10^{-11}$).

Again using RNAi/RNA-seq data, we examined an RBP's RBNS motif density in 3' UTRs for genes significantly up- or down-regulated upon KD of that RBP to generate 3' UTR RNA maps. For example, TAF15 knockdown resulted in decreased levels of many mRNAs (**Fig.** 2-3**F** left), and these mRNAs were enriched for TAF15 motifs in their 3' UTRs relative to control genes with unchanged expression upon KD (**Fig.** 2-3**F** right). These data suggest that TAF15 stabilizes mRNAs by binding to G-rich sequences in 3' UTRs, mirroring expression changes observed upon TAF15 depletion in adult mouse brain and human neural progenitor cells (Kapeli et al. [2016]). Just over half of the RBPs with corresponding KD data (20/38) had RNA expression maps that were consistent with a role in regulating mRNA levels (**Fig.** 2-S3**A** right), equally split between stabilizing and destabilizing activity. Interestingly, SRSF5 motifs were highly enriched in 3' UTRs (and the end of the upstream ORF) of genes up-regulated upon KD (**Fig.** 2-S3**B**). 3' UTR binding has been observed in cell types in which SRSF5 undergoes nucleocytoplasmic shuttling (Botti et al. [2017]), and its role in modulating gene expression levels may be related to its role in linking alternative mRNA processing to nuclear export (Müller-McNicoll et al. [2016]).

2.3.7 RBPs with similar motifs often bind distinct transcript locations

As part of a larger analysis of ENCODE RBP data, we compared RBNS motifs to *in vivo* motifs enriched in eCLIP peaks (Van Nostrand et al. [2017]). We observed strong agreement

between eCLIP and RBNS motifs in most cases, with 17 of 26 proteins having significant overlap between RBNS 5mers and 5mers identified *de novo* as enriched in eCLIP peaks (**Fig.** 2-S3**C**, adapted from (Van Nostrand et al. [2017])). Furthermore, RBNS-enriched 5mers were more enriched in reproducible eCLIP peaks identified in multiple eCLIP replicates and in peaks identified in multiple cell types, which likely represent sites of more robust binding (**Fig.** 2-S3**D**). Together, these observations support that RBNS-identified motifs drive the *in vivo* RNA binding specificity of most RBPs.

In cells, RBPs appear to bind only a subset of cognate motifs in expressed transcripts (Taliaferro et al. [2016]). The extent to which RBPs with similar binding motifs bind the same targets *in vivo* is incompletely understood. Comparing eCLIP-derived binding sites for 131 RBPs, we observed moderate correlation of binding locations between pairs of paralogs, which generally bind highly similar motifs *in vitro*, below that for replicate experiments but above that for randomly chosen RBP pairs (**Fig.** 2-3**G**). However, we observed surprisingly little correlation between binding locations of other pairs of RBPs that bound similar motifs *in vitro* (sharing at least two of their top five RBNS 5mers) (**Fig.** 2-3**G**). Their mean Pearson correlation of 0.20 was not different from random pairs of RBPs, even though *in vivo*-enriched motifs generally matched those observed *in vitro* (**Fig.** 2-S3**C**). For example, while TIA1 and HNRNPC both have high affinity for polyU tracts *in vitro* and *in vivo*, they bind distinct sites in many transcripts (example shown in **Fig.** 2-S3**E**).

The low correlation between locations of *in vivo* binding sites of RBPs with similar motifs could result from various factors, including: i) differences in subcellular localization of RBPs resulting in differential access to transcripts; ii) participation of some RBPs in complexes with other factors that alter RNA specificity (e.g., (Damianov et al. [2016])); iii) occlusion of sites by one RBP leading to the exclusion of other RBPs (Zong et al. [2014]); iv) technical differences in efficiency of eCLIP capture of different regions by different RBPs; or v) differences in binding specificities not well captured by conventional motif representations. While all of these factors likely contribute to differential binding to some extent, we focused here on exploring the last possibility, leveraging the depth and sensitivity of the RBNS data to explore determinants of binding beyond canonical short RNA motifs.

2.3.8 Most RBPs analyzed prefer less structured RNAs

RNA secondary structure can impact RBP binding and regulation (Warf and Berglund [2010]), modulate the activity of regulatory RNA sequences (Hiller et al. [2007]), and contribute to improved *in vivo* RBP binding site prediction when considered (Li et al. [2010]). Since potential RBP binding sites in the transcriptome exist in a variety of structural conformations, we determined RNA secondary structure preferences for each RBP by computationally folding 5 million input and compositionally-matched protein-bound reads for all 78 RBNS experiments at all concentrations (Methods). To assess RNA secondary structure preferences, we first computed the base-pairing probabilities (P_{paired}) of occurrences of the top RBNS 6mer and flanking bases in pulldown libraries (**Fig.** 2-S4**A**) and calculated their ratios to the corresponding P_{paired} values in the input library (**Fig.** 2-4**A**). While the great majority of RBPs favored less base-pairing of the motif itself, some like NUPL2 and RBM41 did to a large extent while others did more modestly. Just six proteins favored increased P_{paired} averaged over the 6mer motif, with the strongest preference observed for ZNF326 (23% increase in mean P_{paired} between input and pulldown, **Fig.** 2-S4**A** right). Structural preferences at flanking positions were more variable, including some proteins within the same cluster preferring increased base-pairing five or more bases away from the 6mer (e.g., BOLL, cluster 10 in **Fig.** 2-4**A**) while others preferred decreased base-pairing (e.g., PTBP3).

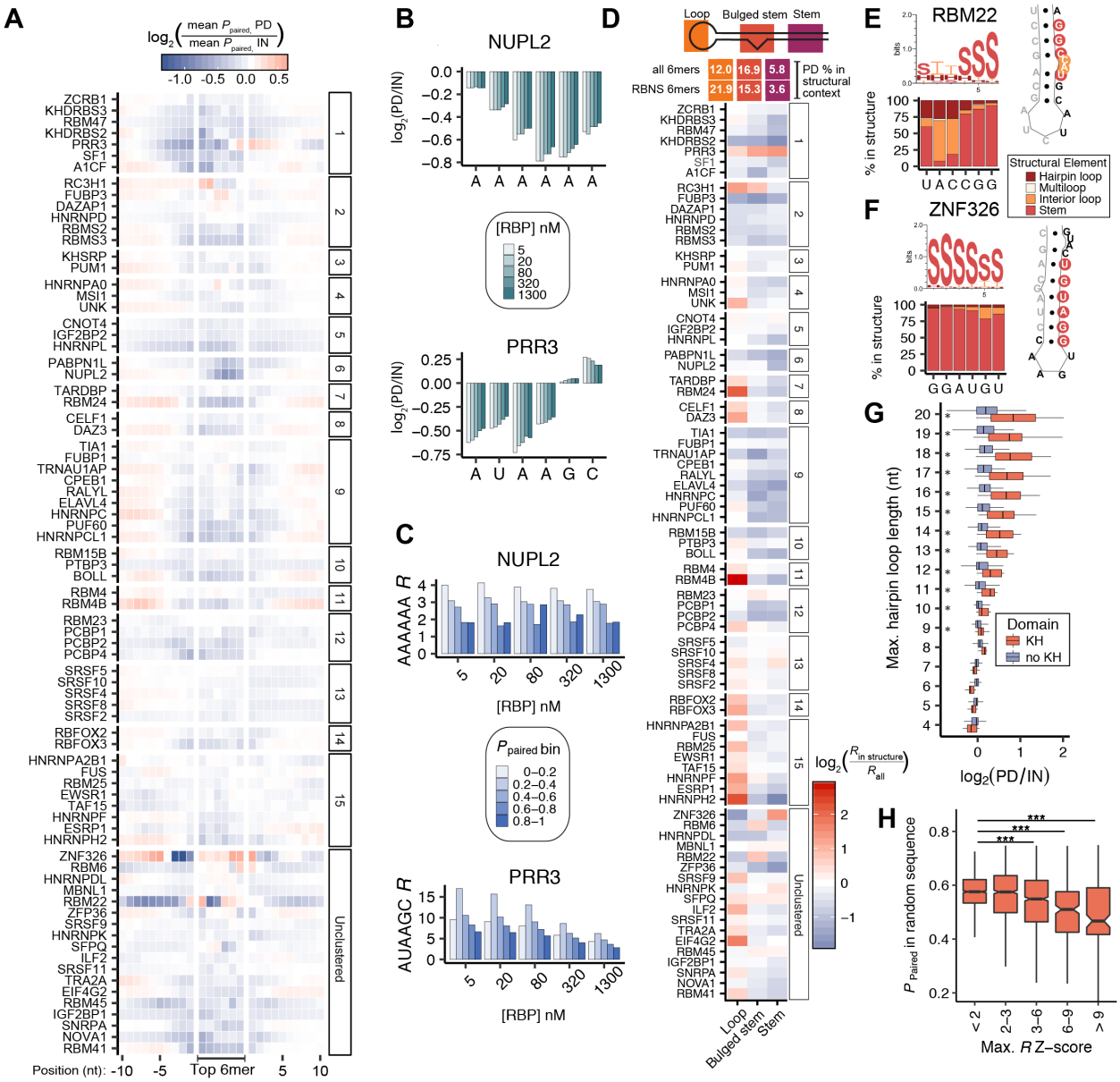

Figure 2-4: **RNA secondary structural preferences of RBPs**

Figure 2-4

(**A**) The $\log_2$(Pulldown P_{paired}/Input P_{paired}) for the most enriched pulldown library over each position of the top 6mer plus 10 flanking positions on each side; RBPs are grouped by motif clusters in **Fig. 2-2A** and from greatest to least mean $\log_2$(Pulldown P_{paired}/Input P_{paired}) over the top 6mer from top to bottom within each cluster.

(**B**) Mean change ($\log_2$) in P_{paired} over each position of the top 6mer at different concentrations of NUPL2 (top) and PRR3 (bottom) relative to the input library.

(**C**) Enrichment of the top 6mer of NUPL2 (top) and PRR3 (bottom) in 5 bins into which all 6mers were assigned based on their average P_{paired}.

(**D**) Top: Three types of structural contexts considered and the percentage of all 6mers and RBNS 6mers (top 6mer for each of 78 RBPs) found in each context in pulldown reads. Bottom: Log-fold change of the top 6mer's recalculated R among 6mers restricted to each structural context relative to the original R.

(**E**, **F**) Left: Percentage of each position of the top 6mer found in the four structural elements for RBM22 (**E**) and ZNF326 (**F**) in pulldown reads. Structure logo for top 6mer is shown above. Right: Representative MFE structures of the top 6mer pairing with the 5' sequencing adapter (gray) for 6mers found at the most enriched positions within the random 20mer (RBM22, position 5; ZNF326, position 6).

(**G**) Enrichment of the percentage of pulldown vs. input reads containing hairpin loops of various lengths, separated by RBPs that contain ($n = 13$) vs. do not contain ($n = 65$) at least one KH domain. Lengths with significant differences determined by Wilcoxon rank-sum test ($P < 0.05$).

(**H**) Average P_{paired} in random sequence for all 6mers binned by maximum R value Z-score across all RBNS experiments ($^{***}P < 0.0005$ by Wilcoxon rank-sum test; overall Spearman correlation $= -0.18$, $P < 10^{-22}$).

Many RBPs showed varying degrees of secondary structure preferences at different positions within the top 6mer. For example, NUPL2, a protein that binds A_6 motifs, strongly disfavored structure at all positions, a pattern observed consistently at all tested protein concentrations (**Fig.** 2-4**B**, top). By contrast, PRR3 disfavored structure at positions 1-4 of its AUAAGC motif but actually favored higher P_{paired} at positions 5 and 6, again consistent across concentrations (**Fig.** 2-4**B**, bottom).

To assess the effect of secondary structure on enrichment, we recomputed R values for all 6mers considering only occurrences of the 6mer in five structure bins ranging from unpaired (average $P_{\text{paired}} < 0.2$ over the 6 positions) to paired (average $P_{\text{paired}} \geq 0.8$) (**Fig.** 2-4**C**, Methods). Consistent with the pattern observed in **Fig.** 2-4**B**, PRR3's top 6mer was most highly enriched in a moderately structured context (P_{paired} 0.2-0.4, **Fig.** 2-4**C**, bottom) while NUPL2's top 6mer was most highly enriched in the least structured context (P_{paired} 0-0.2, **Fig.** 2-4**C**, top). For PRR3 and NUPL2, the R values of the top 6mer were 3- and 4-fold higher, respectively, in the most enriched P_{paired} bin relative to the least enriched bin, underscoring the impact of RNA secondary structure on affinity. Similar magnitudes of enrichment were observed between P_{paired} bins for many other proteins (full listing in Table S4).

2.3.9 RNA structural elements influence binding of some RBPs

Specific RNA structures are known to affect RBP binding in pre-mRNA splicing (Warf and Berglund [2010]), mRNA decay (Goodarzi et al. [2012]), and mRNA localization (Rabani et al. [2008]). To identify preferences for specific structures, we classified each RNA base in the analyzed pulldown and input reads as being part of a stem, hairpin loop, interior loop, or multiloop based on the ensemble of predicted structures (Kerpedjiev et al. [2015]). Averaging over all 78 proteins in pulldown sequences, we found that within hairpin loops the top RBNS 6mers were about two-fold overrepresented compared to all 6mers, while within stems the top RBNS 6mers were about two-fold underrepresented on average (**Fig.** 2-4**D**, top). Correspondingly, the top 6mers of many RBPs were more enriched in a loop context (**Fig.** 2-4**D**, bottom), including RBPs of clusters 7, 8, 11, 14, and almost all members of cluster 15. Fewer RBP motifs were preferentially enriched in bulged stem (9 RBPs) or

stem (8 RBPs) contexts, with generally more modest increases in enrichment than seen in hairpin loops (all enrichments reported in Table S4). Among the strongest bulged stem- and stem-preferring RBPs were the core spliceosomal protein RBM22 (**Fig.** 2-4**E**) - which binds catalytic RNA structural elements in the spliceosome and makes direct contacts with the U6 snRNA Internal Stem Loop and intron lariat (Rasche et al. [2012], Zhang et al. [2017]) - and the zinc-finger protein ZNF326 (**Fig.** 2-4**F**). RBM22 favored a structure with two bases bulged out of a stem while ZNF326 favored a stem within the random region and one base bulged out of its reverse complement paired sequence in the adapter. Unlike most other RBPs, the motifs for these two proteins showed uneven distributions along sequence reads, with increased frequency at the 5' end of the random sequence (**Fig.** 2-S4**B**), and they were commonly predicted to base-pair with the 5' adapter. Examining patterns of enrichment manually, we found evidence that RBM22 prefers AC as the bulged bases, as shown (**Fig.** 2-4**E**), while ZNF326 may simply favor an extended duplex structure independent of sequence.

We also observed a link between RBD type and structural element preference. Large hairpin loops were strongly preferred by 10/13 KH-containing RBPs (all but the FUBP family), while non-KH RBPs showed much more modest preferences for hairpin loops (**Fig.** 2-4**G**). Given that most (7/10) of these RBPs contain more than one KH domain, it is possible that relatively large hairpin loops allow binding of multiple KH domains to the RNA as has been observed in a crystal structure of NOVA1 (Teplova et al. [2011]) and in SELEX analysis of PCBP2 (Thisted et al. [2001]).

For RBPs for which we had corresponding eCLIP data ($n = 26$), we observed a high correlation between RNA secondary structure preferences *in vitro* and *in vivo* (**Fig.** 2-S4**C**). While most RBPs avoided structure in both assays, RBPs that were found to prefer structured or partially structured motifs in RBNS, such as SFPQ, showed this same preference *in vivo*.

A natural question raised by our observation above that human RBPs preferentially bind a small subset of motifs is why RBPs have evolved to bind this particular subset of motifs. While various factors may have contributed to this pattern, we noted a difference in RNA structure. Analyzing RNA folding in random RNAs (**Fig.** 2-4**H**) or in fragments of human introns or exons (see Discussion), we noted that 6mers with higher maximal RBNS

enrichment amongst the 78 experiments were even less structured than 6mers with lower maximal enrichment (overall Spearman correlation = -0.18, $P < 10^{-22}$). Given that most RBPs prefer binding to unstructured motif instances, as observed previously and above (e.g., **Fig.** 2-4**A**), this observation suggests that many RBPs have evolved specificity for motifs that are intrinsically less structured and therefore have greater numbers of accessible occurrences in the transcriptome.

2.3.10 Many RBPs favor pairs of short, spaced motifs

Although structural studies have described a variety of ways that RBDs engage RNA, it is generally thought that a single RBD (e.g., an RRM or KH domain) makes contacts with 3-5 contiguous RNA bases (Auweter et al. [2006]). More than half of the factors in this study contain multiple RBDs (**Fig.** 2-1**B**) and/or multiple types of RBDs, raising the possibility that these RBPs can interact with pairs of short motifs spaced one or more bases apart, hereafter referred to as "bipartite motifs". For example, NMR structures of MSI1's individual RRMs 1 and 2 revealed that they bind GUAG and UAG, respectively, leading to a structural model where both RRMs together bind the sequence UAGN$_{(0-50)}$GUAG (Iwaoka et al. [2017]), and several other examples of RBPs binding to bi- or tripartite motifs have been reported (reviewed in Afroz et al. [2015]), raising the question of how widespread this pattern is.

RBNS is well-suited for the unbiased identification of bipartite motifs as the complexity of the RNA sequence space and the depth of sequencing allow sufficient statistical power to quantify enrichments of longer, spaced $k$mers. We computed enrichments for motifs composed of two 3mer "cores" separated by spacers of 0-10 nt, with spacing 0 representing a traditional contiguous 6mer motif (**Fig.** 2-1**A**, Methods). Since motifs with different spacings have equal information content (6 positions $\times$ 2 bits/position = 12 bits), they can be readily compared. We identified many RBPs that bound bipartite motifs, including PCBP2, ELAVL4, and CELF1, which have been previously reported to bind tandem C-rich, U-rich, and UGU(U/G) sequences, respectively (Teplova et al. [2010], Wang and Tanaka Hall [2001], Teplova et al. [2010]). We found that DAZAP1, which contains two RRMs, preferred AUA followed by another AUA-containing core spaced by 1-3 nucleotides, with no particular

preference for any of the four bases in the spacer (**Fig.** 2-5**A**). RBM45, which contains three RRMs, bound two AC-containing cores separated by a spacer of 1-3 nucleotides, with a slight bias against Gs in the spacer (**Fig.** 2-5**B**). Analysis of all 78 factors revealed that almost one-third of RBPs bound bipartite motifs with similar or greater affinity than linear 6mers, with 18 RBPs showing a significant preference for a bipartite over linear motif at a 5% FDR (**Fig.** 2-5**C**, Table S5, Methods). A complementary analysis testing the preference for specific pairs of spaced cores rather than the aggregate profile over multiple pairs of cores revealed evidence for binding of bipartite motifs for 13 additional RBPs, including several that have been previously reported to bind tandem short sequences (NOVA1 (Teplova et al. [2011]), UNK (Murn et al. [2016]), and the PTB family (Oberstrass et al. [2005])) (**Fig.** 2-S5**A**).

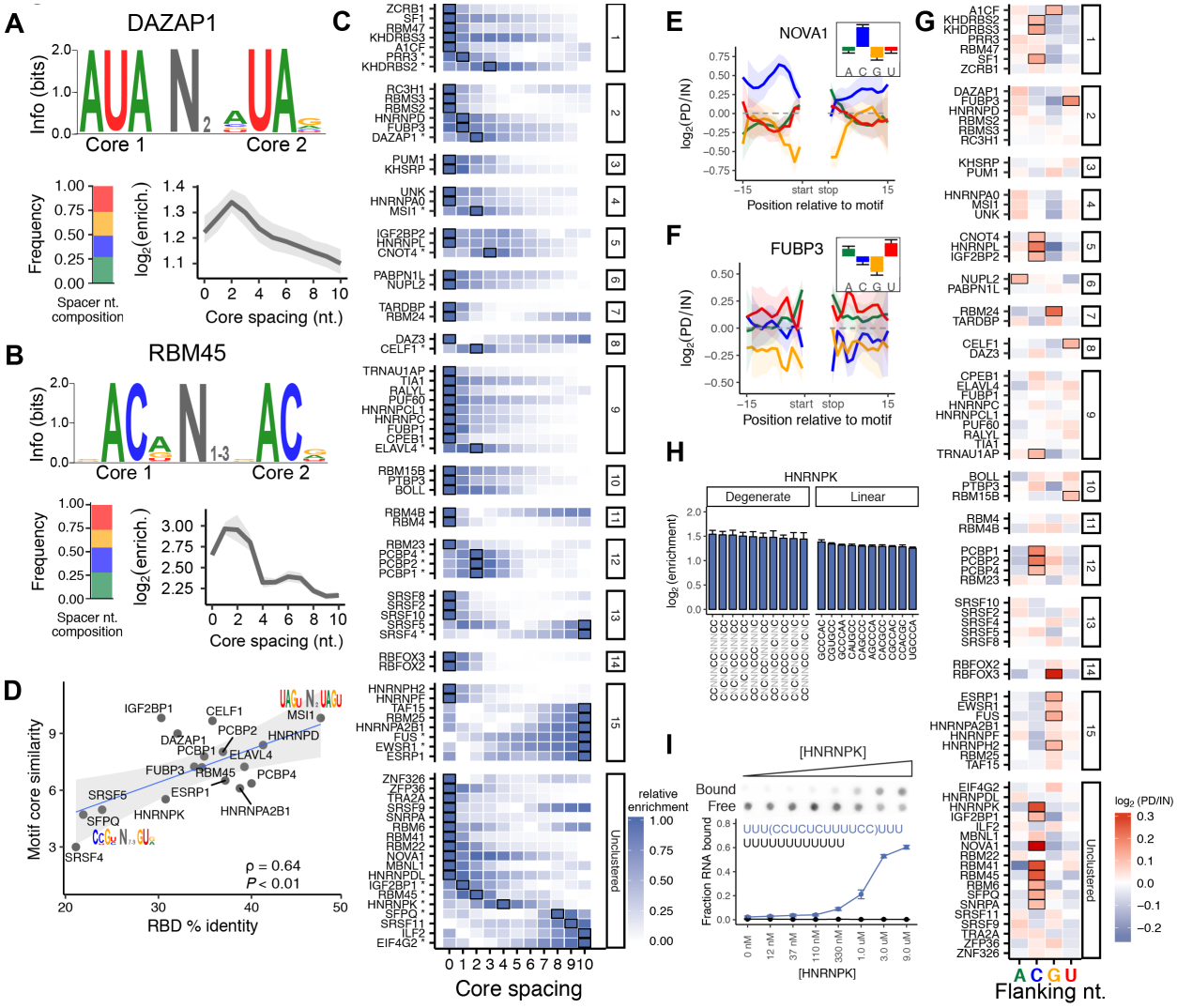

Figure 2-5: **Many RBPs bind bipartite motifs or prefer flanking nucleotide compositions**

Figure 2-5

(**A**, **B**) Top: Sequence logos of bipartite motifs for DAZAP1 (**A**) and RBM45 (**B**). Bottom: Nucleotide composition of the spacer between both motif cores (left) and enrichment as a function of the spacing between cores (right).

(**C**) Core spacing preferences of all RBPs. Each row indicates the relative enrichment as a function of the spacing between cores for a given RBP (i.e., enrichments normalized to the maximum in each row). A box indicates the spacing with maximal enrichment for that RBP, and * to the right of the RBP name signifies the non-zero spacing is significantly preferred over the best linear 6mer. RBPs are grouped by motif clusters in **Fig.** 2-2**A**.

(**D**) Pearson correlation between the maximum identity of RBDs of the same type within an RBP and the similarity between the core motifs of the best bipartite motif. Only RBPs with a significant preference for spacing greater than 0 in **C** and those with at least two RBDs of the same type were used.

(**E**, **F**) Flanking nucleotide compositional preferences surrounding the top five NOVA1 (**E**) and FUBP3 (**F**) 5mers. Reads with no secondary motifs were centered around the top 5mer and the enrichment for each nucleotide in protein-bound reads relative to input reads was calculated (Methods). Inset: mean enrichments across all positions flanking the motif for each of the four nucleotides.

(**G**) Flanking nucleotide compositional preferences of all RBPs. Each row displays the enrichment or depletion for each nucleotide surrounding the RBP's top five 5mers. Boxes indicate significant enrichment ($\log_2(\text{enrichment}) > 0.1$ and $P < 0.001$, Methods).

(**H**) Enrichments of HNRNPK's top 10 linear 6mers (right) and top 10 degenerate sequences of length 12 with 6 Ns (left).

(**I**) Filter assay validation of HNRNPK binding to the oligo UUU(CCUCUCUUUUCC)UUU (blue) and the oligo U_{12} (black) as a negative control (Methods). Dot-blot of filter assay shown on top with fraction of RNA bound quantified below.

As expected, preference for spacing was associated with the presence of more than one RBD (**Fig.** 2-S5**B**, $P = 0.023$, $t-$test), although several exceptions were observed. For example, KHDRBS2 showed the strongest preference for a bipartite motif even though it contains a single KH domain (**Fig.** 2-S5**B**). However, KHDRBS2 also has an N-terminal QUA1 (Quaking-1) domain, a domain which has been shown to promote homodimerization of some STAR (Signal Transduction and Activation of RNA) family proteins (Beuck et al. [2012], Meyer et al. [2010]). Thus, KHDRBS2 may bind bipartite motifs as a homodimer.

We also observed several proteins whose affinity for RNA continuously increased with longer spacers (e.g., TAF15 and ESRP1 in cluster 15 of **Fig.** 2-5**C**), some of which may result from multimerization. For example, FUS, a factor that displayed increasing enrichments as a function of spacing, has a C-terminal RG-rich domain that has been shown to promote cooperative binding to RNA (Schwartz et al. [2013]). Notably, EWSR1, a FUS paralog that is a member of the FET (FUS, EWSR1, TAF15) family and has a similar domain composition, displayed the same preference for increased spacing, suggesting it may also multimerize through low-complexity domains (Schwartz et al. [2015]).

We noted that the two RNA cores bound by MSI1 were nearly identical and that MSI1's two RRMs were highly similar at the amino acid level (∼47% identity). In contrast, SFPQ favored a bipartite motif consisting of two very different RNA cores, and its RBDs were much less similar (∼22% identical). Expanding this observation to all RBPs that preferred spaced motifs and contain at least two RRM, KH, or ZNF domains, we observed a strong positive correlation between the percent identity of sibling RBDs within a protein and the similarity of the bipartite motif RNA cores (**Fig.** 2-5**D**, Pearson correlation = 0.64, $P < 0.01$). These observations support the model that most binding of bipartite motifs in the set of RBPs analyzed involves engagement of RNA by more than one RBD. Furthermore, our observation that many bipartite cores are highly similar to one another is supported by the recent finding that RRMs within the same protein are often the result of recent RRM duplications and therefore highly similar to one another (Tsai et al. [2014]).

2.3.11 RNA sequence context commonly influences RBP binding

It has previously been observed that binding of certain transcription factors may be enhanced by a particular nucleotide composition adjacent to a high-affinity motif (Jolma et al. [2013]) and that such flanking nucleotide biases are also seen around motifs within ChIP-seq peaks (Wei et al. [2010]). We hypothesized that adjacent nucleotide context could play a similar role in modulating RBP specificity by altering local RNA secondary structure or creating additional interactions with the RBP. One such example is Argonaute-2, which preferentially binds miRNA target sites in an AU-rich context, a feature often used to predict miRNA targeting efficacy (Agarwal et al. [2015], Grimson et al. [2007], Nielsen et al. [2007]). For each RBP, we computed the enrichment of each nucleotide at all positions in reads surrounding a high-affinity motif, using only those reads containing one of the top five 5mers and no secondary motifs (Methods).

We found 28 proteins with a significant preference for a particular flanking nucleotide composition (mean $\log_2$(enrichment) > 0.1, $P < 10^{-3}$ by Wilcoxon rank-sum test, Table S5, Methods). For example, NOVA1 preferred to bind its motif in a C-rich context (**Fig.** 2-5**E**) while FUBP3 preferred to bind its motif in a U-rich context (**Fig.** 2-5**F**). We noted an enrichment for RBPs with KH domains within this set ($P < 10^{-3}$, Fisher's exact test), as seen for factors that prefer binding to RNAs within hairpin loops (**Fig.** 2-4**G**). While particular flanking nucleotide compositions may be correlated with the presence of large hairpin loops, we observed a majority of these flanking nucleotide compositional preferences even after controlling for the secondary structure context of the motif, suggesting that nucleotide context effects and secondary structure can each contribute to binding (**Fig.** 2-S5**C**). In most cases, this nucleotide preference was dependent on the presence of a motif in the read, suggesting that flanking sequence promotes or stabilizes RBP binding to a primary motif. However, some RBPs showed these same nucleotide preferences in the absence of a high-affinity motif (e.g., FUS and IGF2BP1, **Fig.** 2-S5**D**), suggesting that these factors have affinity for degenerate sequences with biased nucleotide content.

To formalize these observations and test cases in which biased sequence composition may better describe an RBP's specificity than a linear motif, we calculated enrichments for

degenerate patterns with biased nucleotide composition. For example, HNRNPK, one of the RBPs that showed a preference for C bases in the absence of a high-affinity *k*mer, was far more enriched for the degenerate pattern CNCNCNCNNNCC (enriched 2.9-fold) than the corresponding contiguous 6mer CCCCCC (1.11-fold) with identical information content. In fact, HNRNPK showed greater or equal enrichments for various permutations of the above degenerate pattern (6 Cs with 6 interspersed Ns) relative to the most highly enriched linear 6mers (**Fig.** 2-5**H**). We generalized this observation by calculating enrichments for all degenerate patterns composed of 6 fixed bases and 6 Ns for each RBP (Methods) and found that HNRNPK had higher enrichment for many C-rich degenerate patterns than for its best linear motifs of equal information content (**Fig.** 2-S5**E**). Most other RBPs such as RBFOX2 strongly preferred specific linear sequences over degenerate patterns (**Fig.** 2-S5**E**).

Because binding of RBPs to such degenerate patterns has not been extensively studied, we sought to confirm the preference of HNRNPK for a degenerate sequence, CCNCNC-NNNNCC, containing no motif longer than two bases. Substituting U at the N positions in order to avoid creating RNA secondary structure or other potential motifs, we validated that HNRNPK specifically bound RNAs containing this pattern using a filter binding assay (**Fig.** 2-5**I**). These degenerate patterns were also enriched more than two-fold relative to linear 6mers in HNRNPK eCLIP peaks, supporting *in vivo* binding of such sequences (**Fig.** 2-S5**F**). In all, we identified 17 RBPs whose binding was well-described by degenerate patterns, 14 of which bound spaced motifs (**Fig.** 2-5**C**) and were often enriched for patterns similar to their previously identified bipartite motifs (e.g., CELF1, **Fig.**2-S5**G**). However, at least three RBPs showed enrichment for patterns with no more than 2 contiguous specified bases (e.g., FUBP1, HNRNPK, PUF60; FUBP1 is shown in **Fig.** 2-S5**H**). These patterns may therefore represent degenerate bi- or tripartite motifs, with multiple RBDs each contacting just one or two RNA bases.

2.3.12 Towards a more complete characterization of RBP specificities

Our observation of widespread preferences for secondary structure features, bipartite motifs, and flanking nucleotide composition led us to hypothesize that RBPs favoring similar primary sequence motifs may differ in their preferences for these secondary contextual features in order to select different subsets of targets. For example, PCBP2 and RBM23 (cluster 12) both bind C-rich sequences even though they have distinct RBD composition (3 KH domains versus 2 RRMs). Our analyses indicated that PCBP2 avoids structure within its motif, is capable of binding the bipartite motif CCCNNCCC, and is enriched for flanking C bases. In contrast, RBM23 has no structural preference over its motif and favors a contiguous C-rich motif with no flanking nucleotide compositional preference. Thus, PCBP2 and RBM23 are likely to bind distinct C-rich sites in transcripts.

In order to systematically compare the contributions of these features, we computed "feature-specific" R values for the top 6mer of a cluster. These feature-specific R values measured an RBP's 6mer enrichment as a function of: i) P_{paired} of the 6mer; ii) the average base-pairing probability of the sequence flanking the 6mer (P_{flank}); or iii) the nucleotide frequencies surrounding the 6mer (Methods). Feature-specific R values for bipartite motifs were calculated based on the pattern of preference for spacing between the split motif cores (analogous to **Fig.** 2-5**C**). We then measured a correlation-based distance between feature-specific R value profiles for pairs of RBPs within the same motif cluster (**Fig.** 2-S6**A**). Intra-cluster pairwise distances were significantly higher than distances calculated between replicate RBNS experiments at different RBP concentrations for P_{paired}, flanking nucleotide content, and bipartite motifs. This analysis suggests that RBPs with similar primary motifs are often differentially affected by contextual features (**Fig.** 2-6**A**).

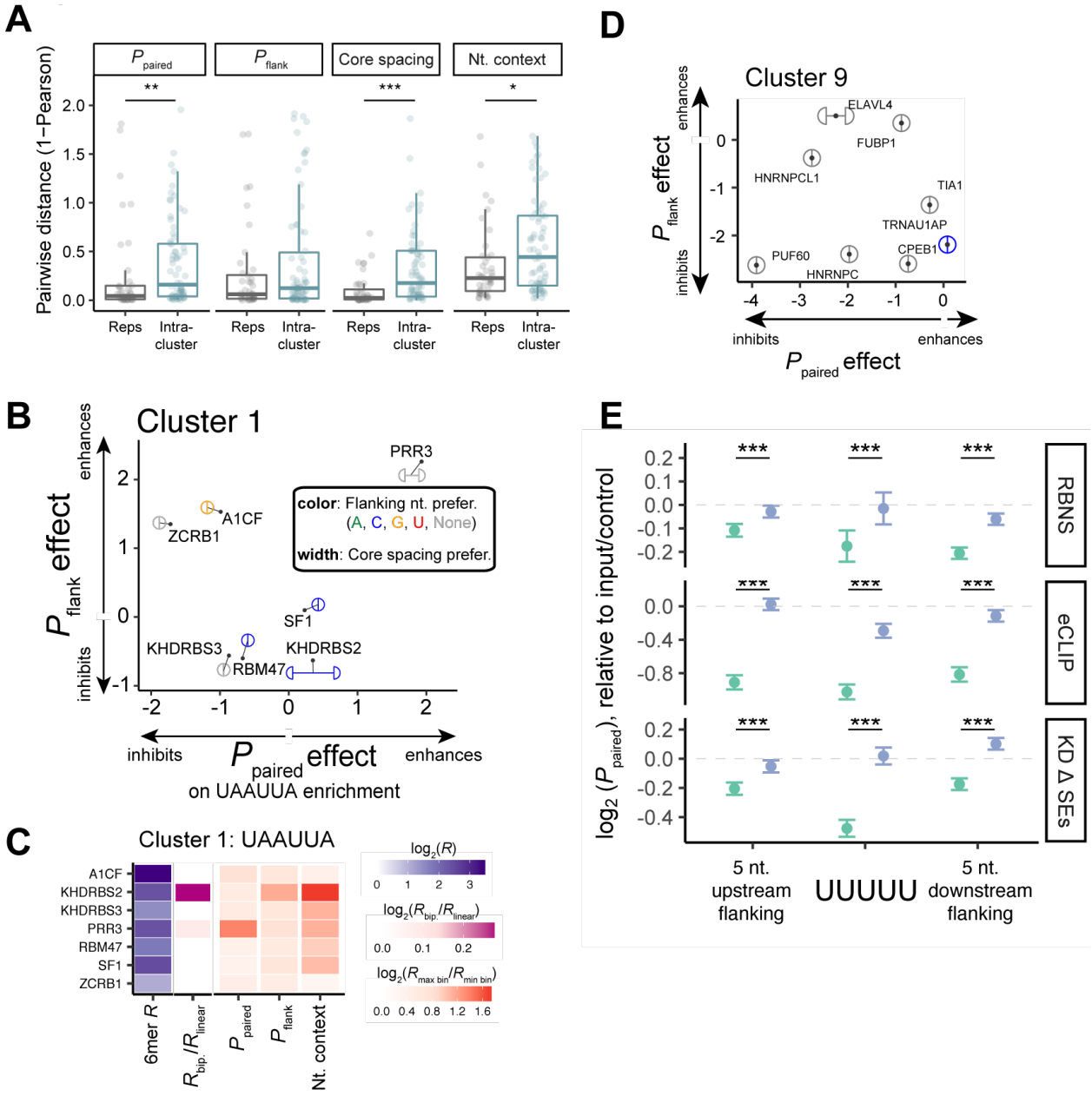

Figure 2-6: **RBPs that bind similar motifs often diverge in sequence context preferences**

Figure 2-6

(**A**) Pairwise distances (1 - Pearson correlation) of feature-specific R values for pairs of RBPs within a motif cluster ("Intra-cluster") compared to distances between controls ("Reps"), where controls are replicate assays of the same RBP performed on different days and at different protein concentrations. Significance determined by Wilcoxon rank-sum test ($^{*}P < 0.05$, $^{**}P < 0.005$, $^{***}P < 0.0005$).

(**B**) Overall dispersal of specificities between cluster 1 RBPs for RNA motif UAAUUA. x- and y- axes represent the degree to which secondary structure over (x) and flanking (y) the motif affect UAAUUA R values. Coloring of circle denotes whether the protein displayed a significant enrichment for flanking nucleotide composition, with gray denoting none. Split semicircles indicate whether the protein had significant preference for a bipartite motif over a linear motif with the distance between the semicircles reflecting the preferred spacing of the cores, and no separation indicating preference for linear motif.

(**C**) From left to right: UAAUUA R value for each RBP in cluster 1; $R_{\text{bipartite}}/R_{\text{linear}}$ 6mer if a bipartite motif was significantly preferred in **Fig.** 2-5**C**; the ratio of maximum R to minimum R for feature-specific R values over 5 context bins for each of the context features P_{paired}, P_{flank}, and flanking nucleotide composition.

(**D**) Top: Same as **B** for cluster 9 RBPs relative to RNA motif UUUUUU.

(**E**) Mean and standard error of the $\log_2$ ratio in P_{paired} over non-overlapping UUUUU occurrences and the 5 nucleotides directly up- and downstream in: Top: RBNS pulldown relative to input; Middle: Intronic eCLIP peaks relative to randomized peak locations; Bottom: 150 nt of intron downstream of exons with increased inclusion upon RBP knockdown relative to same regions downstream of control exons. Significance determined by Wilcoxon rank-sum test ($^{*}P < 0.05$, $^{**}P < 0.005$, $^{***}P < 0.0005$).

To visualize protein preferences for each of the four features among RBPs within the same cluster, we placed RBP markers (consisting of paired semicircles) on a coordinate system according to their structural preferences P_{paired} and P_{flank}, colored each RBP marker based on its flanking nucleotide preferences, and separated the semicircles based on the bipartite spacer preferences (Methods). This visualization of RBPs within the AU-rich cluster 1 revealed that no two RBPs are superimposed in this multidimensional sequence space (**Fig.** 2-6**B**), and divergences were observed within most other clusters as well (**Fig.** 2-S6**B**). Overall, we found that 9/15 clusters diverged significantly in at least one feature, and 5/15 diverged in more than one feature with the most common significant feature being bipartite motif spacing (Table S6). This analysis therefore divides RBP clusters into those likely to bind distinct sites and those likely to compete for binding to the same target sites when coexpressed.

To systematically compare the effects of these contextual features on RBP affinities, we examined the R value differences between the most and least favored motif occurrences for each feature. As shown in **Fig.** 2-S6**C**, the ratio of the most to least favored context for the structure of the motif, the structure flanking the motif, and the nucleotide context for each RBP varied from close to 1.0 for features that did not affect an RBP's motif enrichment to over 10-fold for the P_{paired} context of HNRNPH2, which strongly favored unstructured occurrences of its G_6 motif. Considering the AU-binding RBPs of cluster 1 as an example, KHDRBS2 had an R value for its bipartite motif that was ~20% above that of the top contiguous 6mer. The context features with the greatest effects on binding were the nucleotide context flanking KHDRBS2 motifs and structure over PRR3 motifs with ~3.5- and 2.5-fold increases in R for the most versus least favored contexts, respectively (**Fig.** 2-6**C**, **Fig.** 2-S6**D**). Overall, structure over the motif typically had the greatest impact on R values, with a mean 2.2-fold difference between the most and least favored P_{paired} bins, while the P_{flank} and flanking nucleotide composition had mean 1.7- and 1.9-fold differences, respectively. Together, these context effects appear to impact the overall specificity and affinity of this and other RBPs in the cluster significantly (**Fig.** 2-6**C**, left; **Fig.** 2-S6**C**, bottom). In summary, we find that the specificity of most RNA binding proteins is conferred not only by primary sequence elements but also by a variety of contextual properties of the local RNA sequence

environment.

It is worth emphasizing here that because in our assay binding of an RBP to a specific $k$mer pulls down an entire oligo containing many other $k$mers, the R value greatly underestimates the affinity of the RBP for the $k$mer relative to background. Applying a formalism to estimate quantitative differences in affinity from R values that we introduced previously (Lambert et al. [2014]), we have that an R value of 1.5 for a 6mer within a 20mer oligo corresponds to ∼7-fold increased affinity above background, while an R value of 4 corresponds to ∼45-fold increased affinity. Thus, the ∼1.5- to 3.5-fold increases in R value seen for various context features above generally indicate increases in affinity of severalfold or more.

Finally, we sought to determine if the contextual features observed *in vitro* could help proteins discriminate among potential regulatory targets *in vivo*. We focused on comparison of HNRNPC and TIA1 as they had the following desirable features: (1) both have well-established roles in regulating splicing and KD/RNA-seq yields RNA splicing maps consistent with these roles (**Fig.** 2-S3**A**); (2) both have available *in vivo* binding (eCLIP) data where the eCLIP-derived motif is identical to the RBNS-derived motif (**Fig.** 2-S3**E**); and (3) they have distinct *in vitro* context preferences despite sharing the same top 5mer U5 (**Fig.** 2-6**D**). The two context features most divergent between HNRNPC and TIA1 were structure over the motif (P_{paired}) and flanking it (P_{flank}), with HNRNPC showing a stronger bias against structure in both locations via RBNS (**Fig.** 2-6**E**, top). Examining eCLIP peaks for each factor, we also observed a stronger bias against structure for HNRNPC than for TIA1 on and flanking U_5 motifs in eCLIP peaks (**Fig.** 2-6**E**, middle). Furthermore, we also observed a stronger bias against structure on and flanking U_5 motifs located downstream of HNRNPC-regulated exons than for TIA1-regulated exons (**Fig.** 2-6**E**, bottom). Thus, the contextual features identified by our *in vitro* assay appear to help distinguish *in vivo* binding and regulatory locations of different factors that recognize the same primary motif.

2.4 Discussion

2.4.1 Towards an RNA processing parts list of RNA elements and RBPs

A substantial body of work has aimed to catalog functional RNA elements and their interacting proteins to gain a more complete and mechanistic understanding of RNA processing in cells. For example, analysis of *k*mer frequency and evolutionary conservation in specific regions of the genome has led to the appreciation that many *cis*- sequences are associated with specific types of regulation (Fairbrother et al. [2002], Ke et al. [2011]), even when the *trans*-acting RBPs that bind them are not always known. Here, we characterized 78 human RBPs using a high-throughput version of RBNS, a one-step *in vitro* binding assay that assesses the spectrum of RBP primary sequence binding specificities along with contextual features that influence binding such as secondary structure, bipartite motifs, and flanking nucleotide content.

2.4.2 RBPs recognize a small subset of the available sequence space

Considering a diverse set of 78 proteins, we found that many RBPs bind a relatively small, defined subset of primary RNA sequence space rich in low-complexity motifs primarily composed of just one or two bases. This trend was seen independently for RRM and KH domain proteins that do not share common ancestry, suggesting convergent evolution toward recognition of these types of motifs. These findings are consistent with previous studies implicating AU-, U-, and G-rich sequences as functional elements that regulate stability and splicing (Fu and Ares Jr [2014], Wu and Brewer [2012]). Previous studies have shown certain mono- and dinucleotide-rich sequences occur in clusters to mediate their effects on RNA processing (Barreau et al. [2006], Cereda et al. [2014]). These motifs also have lower propensity to form secondary structures that might block RBP binding, both in random sequences (**Fig.** 2-4**H**) and in sequences from the human transcriptome (**Fig.** 2-S6**E**). Thus, greater average accessibility for binding may have favored evolution of RBP specificity towards a set of motifs. Binding modes involving sliding of RBPs along RNA may also favor low complexity motifs.

For example, HNRNPC was shown to bind runs of uridines with a potential to slide among registers within an RNA (Cieniková et al. [2014]). Such a sliding model would be most feasible for mono- or dinucleotide repeat motifs, while more complex motifs would likely require the RBDs to completely dissociate from the RNA between specific binding interactions.

A similar analysis of DNA binding proteins and transcription factors (not shown) revealed that they do not show the same inherent propensity to target low complexity sequences, perhaps due to differences in the size of the search space and differences in DNA and RNA structure and the biochemical mechanism of binding (Jolma et al. [2013]). While the overlapping specificity across the 78 RBPs is high, we found that 18 of the 78 profiled RBPs have motifs dissimilar from that of any other RBP assayed ('Unclustered' in **Fig.** 2-2**A**), suggesting that a subset of RBPs may have evolved to recognize more distinct sets of RNA targets. These 18 RBPs tended to have broader expression profiles across human tissues than did RBPs within motif clusters, while RBPs with motifs shared by other RBPs were more often expressed tissue-specifically (**Fig.** 2-S6**F**). Thus, some members of the same clusters may serve redundant functions in different cell or tissue types. Even so, coexpression of many members of a cluster appears widespread, and this is likely to provide versatile opportunities for post-transcriptional gene regulatory networks and evolution (Lapointe et al. [2017]).

2.4.3 RBP binding specificities harbor hidden complexity

Closer analysis of the complexity of RBNS data revealed that linear sequence motifs are often insufficient to fully capture RBP binding specificities, with sequence features beyond short motifs contributing to the specificity of most RBPs (**Fig.** 2-S6**C**). We find that nearly all RBPs have strong preferences for or more often against RNA secondary structure, and about half of RBPs were identified as favoring noncontiguous 'bipartite' motifs. While the most common representation of RBP binding sites is a single position weight matrix (PWM), our data suggest that in many cases RBP specificity may be better described by pairs of short motifs with variable spacing, and sometimes by representations describing structural preferences. We hypothesize that these context preferences may be general features that allow RBPs with similar motifs to select distinct targets in cells, a paradigm that has been put forth for pairs of RBPs (Smith et al. [2013]).

In assaying a number of RBPs containing RRM, KH, and zinc finger domains, we noted several commonalities across RBPs with similar RBDs. For example, RRMs and ZNFs were identified that bound to A-, C-, G- and U-rich motifs, while KH domains bound A-, C- and U-rich but not G-rich motifs, in agreement with previous structural studies showing that G-recognition by KH domains is rare (reviewed by Nicastro et al. [2015]). Further, while many motif clusters included RBPs from all three RBD types analyzed (**Fig.** 2-2**A**), 5mer enrichments were more highly correlated among RBPs with the same RBD type relative to RBPs with different RBD types, even after excluding paralogs (**Fig.** 2-S6**G**), suggesting greater residual similarity in binding preferences.

We also noted domain-specific trends among preferences for contextual features. For example, while RBPs overwhelmingly preferred unstructured motifs both *in vitro* (**Fig.** 2-4**A**) and *in vivo* (**Fig.** 2-S4**C**), ZNFs bound motifs encompassing a wide range of secondary structure contexts including those with greater stem content (**Fig.** 2-S6**H**), and RBPs containing ZNFs showed greater variability in their P_{paired} preferences than did other RBPs (**Fig.** 2-S6**I**). These observations are consistent with a recent study finding that more than twenty ZNF-containing proteins selectively bound highly-structured pre-microRNAs (Treiber et al. [2017]). Proteins with KH domains also shared numerous properties, including a preference for binding to large hairpin loops, flanking nucleotide compositional preferences, and common preference for bipartite motifs. Notably, all but one KH-containing RBP assayed by RBNS (SF1) had either more than one KH domain or a known homodimerization domain, suggesting that binding of RNA by pairs of KH domains is very common.

Together, our findings emphasize the complexity of the interactions between proteins and RNA and motivate future studies of the structural basis of the dependence on the various contextual features documented here.

2.5 Supplementary Figures

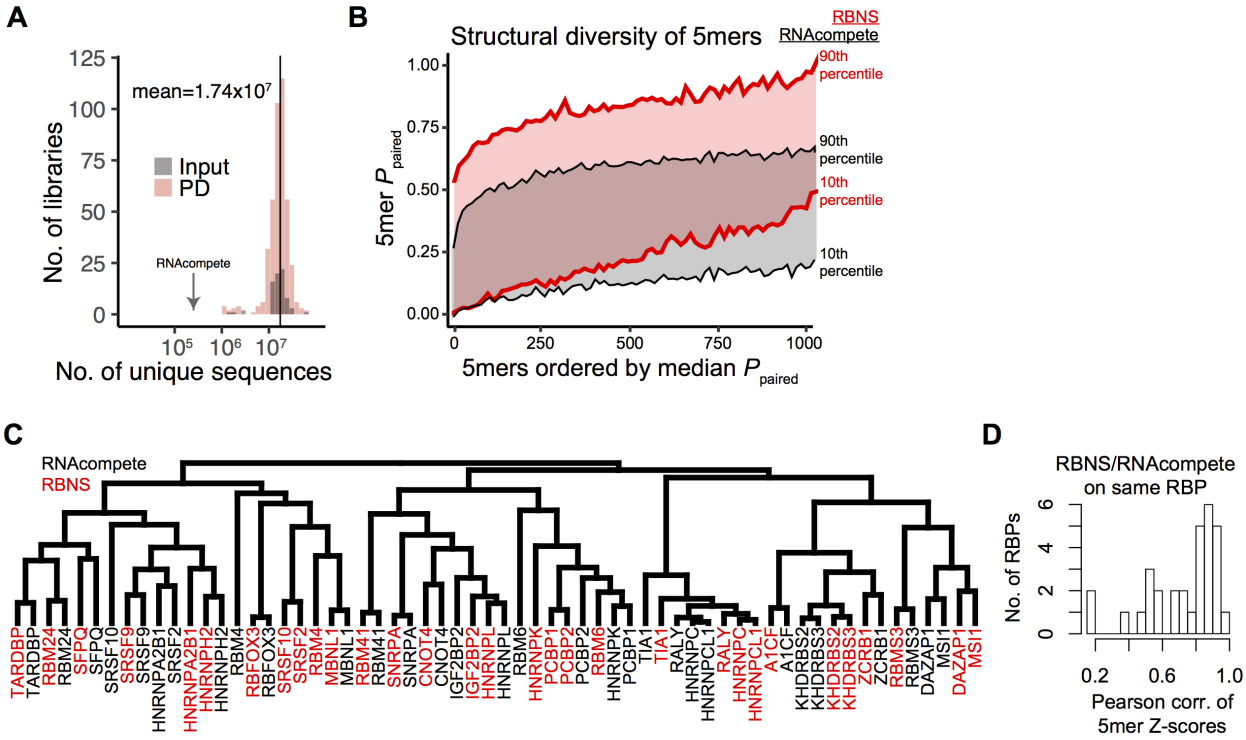

Figure 2-S1: **RBNS assay and comparison to RNAcompete.**
(**A**) Histogram of the number of unique reads sequenced in all RBNS Input and Pulldown libraries, with the ∼250,000 RNAcompete oligo sequences marked for comparison.
(**B**) 10th-90th percentile range of the mean base-pairing probability (P_{paired}) for all 5mer occurrences (averaged over all five positions) in RBNS input libraries and the RNAcompete custom array. The 1,024 5mers are sorted along the x-axis by increasing median RBNS P_{paired}.
(**C**) Dendrogram of clustered RBNS (red) and RNAcompete (black) experiments for the 31 RBPs that were assayed by both methods. Clustering was performed via the hclust function in R (with 'method=average') on the pairwise Pearson correlations of 5mer R Z-scores, for the set of 306 5mers that were enriched ($Z \geq 3$) in any of the 62 experiments.
(**D**) Histogram of Pearson correlations of RBNS and RNAcompete for the 31 RBPs assayed by both methods as described in **C**.

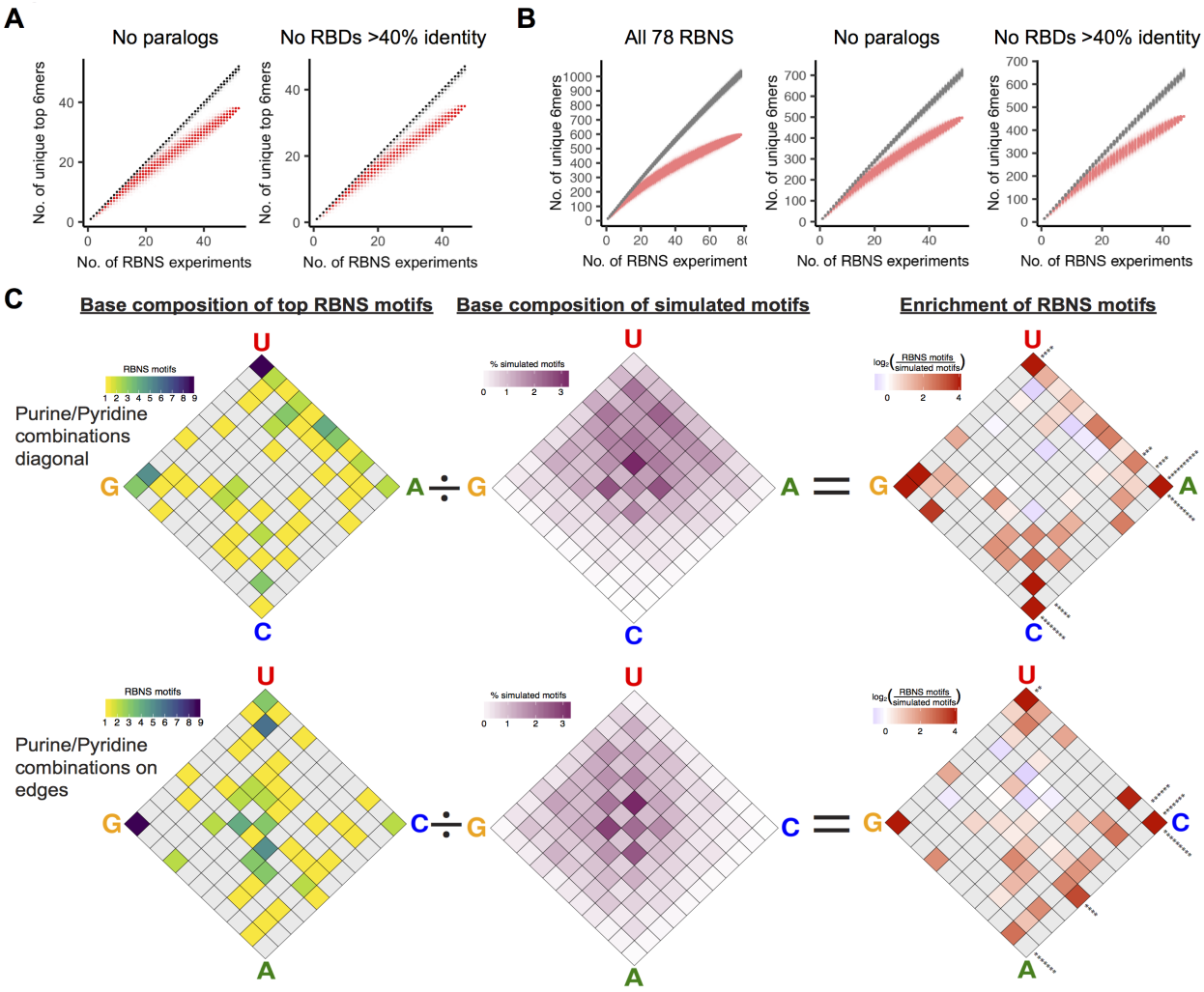

Figure 2-S2: **Overlapping specificities of RBPs**
(**A**) Number of unique top 6mers among random subsamplings of the RBNS experiments versus randomly selected 6mers (similar to **Fig.** 2-2**C**), for the subset of 78 RBNS experiments that excludes any paralogs (left) or any RBPs that share at least 40% identity among any RBDs (right).
(**B**) Similar to **Fig.** 2-2**C** and **Fig.** 2-S2**A**, but for the top 15 6mers of each RBNS experiment instead of just the top 6mer. Black line determined from sets of 15 6mers in which the top 6mer was chosen at random, with the remaining 14 6mers matching the edit distances relative to the top 6mer observed among actual RBNS experiments (Methods).
(**C**) Mapping of the four nucleotide frequencies in motif logos onto a 2D simplex, for both the actual 78 RBNS motifs (left), simulated RBNS motifs (center), and the resulting enrichment of RBNS versus simulated motifs in each simplex partition (right) (Methods). Gray boxes denote that none of the 78 RBNS motif frequencies mapped to that partition. Significance along margins of the enrichment simplex was determined by bootstrap Z-score (number of asterisks = Z-score). The data in the top and bottom rows is the same, with the A and C corners of the simplex switched so that each of the six dinucleotide combinations (AC, AG, AU, CG, CU, GU) is included along an edge in at least one of the two mappings. Upper right simplex is the same as **Fig.** 2-2**F**.

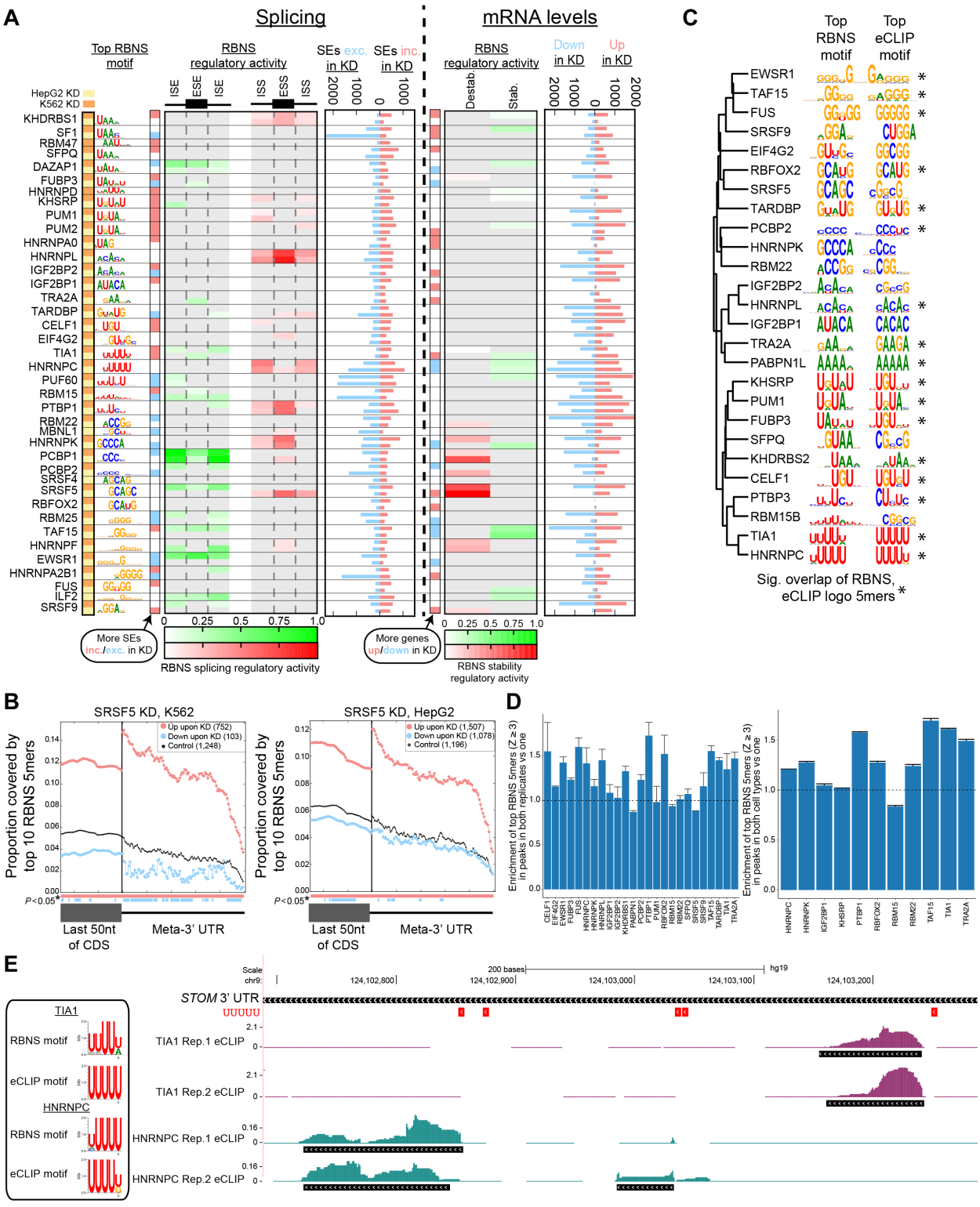

Figure 2-S3: **RBNS-derived splicing and stability RNA maps and RBP binding in the transcriptome**

Figure 2-S3

(**A**) Summary of RBNS-inferred regulatory activity of RBPs in pre-mRNA splicing and mRNA levels.

Left (Splicing): From left to right: All RBPs assayed by both RBNS and KD/RNA-seq, ordered from top to bottom by RBNS motif (same ordering as in **Fig.** 2-2**A**); a bar denoting KD in HepG2/K562; the top RBNS motif logo as in **Fig.** 2-2**A**; a bar denoting whether there were significantly more SEs included upon KD (pink) or excluded upon KD (blue); the inferred RBNS regulatory activity of the RBP over the SE and upstream and downstream 250 nt of flanking intron (Methods), with strength of RBNS splicing regulatory activity of significant regions noted by green (ESE/ISE) or red (ESS/ISS) heat map; total number of SEs changing in each direction upon KD.

Right (mRNA levels): Similar to Left (Splicing), but for inferred activity on gene expression levels based on RBNS motif density in the 3' UTRs of genes significantly up- or down-regulated upon KD (Destab.=increased RBNS density in 3' UTRs of genes up-regulated upon KD; Stab.=increased RBNS density in 3' UTRs of genes down-regulated upon KD).

(**B**) Similar to **Fig.** 2-3**F** right, but for SRSF5 KD in K562 (left) and HepG2 (right) cells.

(**C**) Comparison of RBNS and eCLIP motifs for RBPs assayed by both techniques (reproduced with permission from (Van Nostrand et al. [2017])). The top RBNS and eCLIP motifs are shown for each RBP (Methods), clustered by RBNS motif as in **Fig.** 2-2**A**. 17 of the 26 RBPs with significant overlap in the 5mers comprising the RBNS and eCLIP logos ($P < 0.05$, hypergeometric test) marked with a star to the right of the eCLIP logo.

(**D**) Left: Enrichment of RBNS 5mers (averaged among all RBNS 5mers with Z-score ≥ 3) in eCLIP peaks that occur in both replicates (peaks overlap by at least 1 position) relative to peaks that occur in only one replicate. Right: Enrichment of RBNS 5mers (averaged among all RBNS 5mers with Z-score ≥ 3) in eCLIP peaks that occur in both HepG2 and K562 (peaks overlap by at least 1 position) relative to peaks that occur in only one cell type.

(**E**) Genome browser snapshot of a portion of the 3' UTR of the *STOM* gene and eCLIP entropies ($\log_2$(IP density/input density)) for TIA1 and HNRNPC bound to two different locations in this region. Instances of U_5 motifs shown in top track (red) and eCLIP peaks are marked in black below each entropy track. Motifs derived from RBNS and eCLIP peaks for each RBP shown on the left.

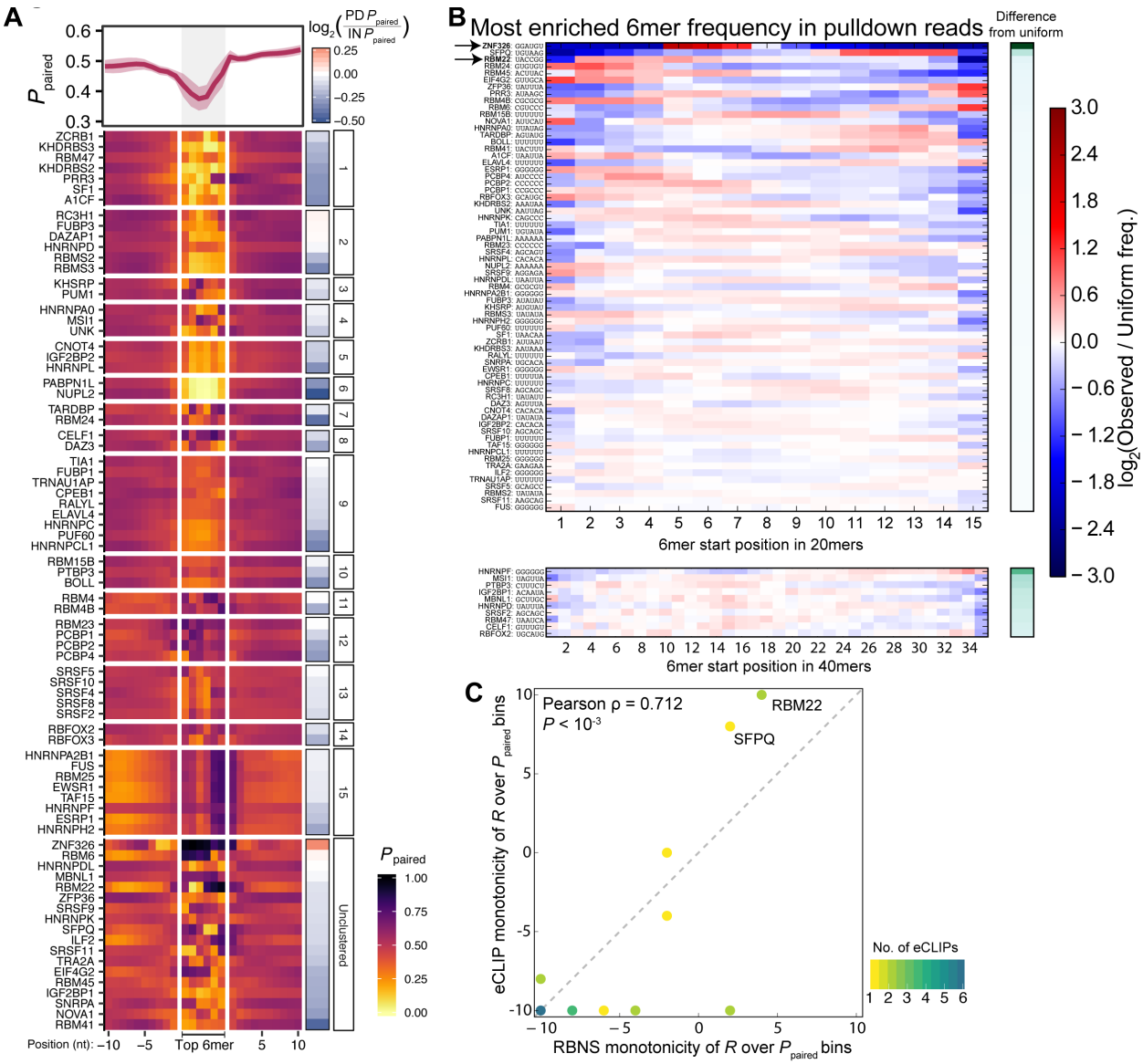

Figure 2-S4: *In vitro* **and** *in vivo* **structural preferences of RBPs and distribution of enrichments across reads**

Figure 2-S4

(**A**) Top: P_{paired} over each position averaged over the 78 RBPs; 95% confidence interval is shadowed. Bottom: Mean P_{paired} in the most enriched pulldown library over the top 6mer plus 10 flanking positions on each side; RBPs are grouped by motif clusters in **Fig.** 2-2**A**. Right: Mean change ($\log_2$) in pulldown vs. input P_{paired} averaged over the top 6mer.
(**B**) The frequency of the top RBNS 6mer at each position of the random region, relative to a uniform distribution at all positions (top: RBPs with random 20mers; bottom: random 40mers). Difference from uniform denoted by the green heat map bar to the right, calculated as the KL-divergence of the observed frequency at each position relative to a uniform distribution; RBPs are sorted by decreasing difference. RBPs and their top 6mers noted on left, with ZNF326 and RBM22 marked as the 1st and 3rd most unequal distributions across 20mers, respectively.
(**C**) Correlation of R value profiles of the top 5mer across five bins of increasing P_{paired} for RBPs assayed by both RBNS (x-axis) and eCLIP (y-axis). For each assay, R was calculated in each of the 5 P_{paired} bins (as in **Fig.** 2-4**C**) and the monotonicity of R over the 5 bins was calculated (-10 monotonicity = 5 bins monotonically decreasing R with increasing P_{paired}; 10 monotonicity = 5 bins monotonically increasing R with increasing P_{paired}).

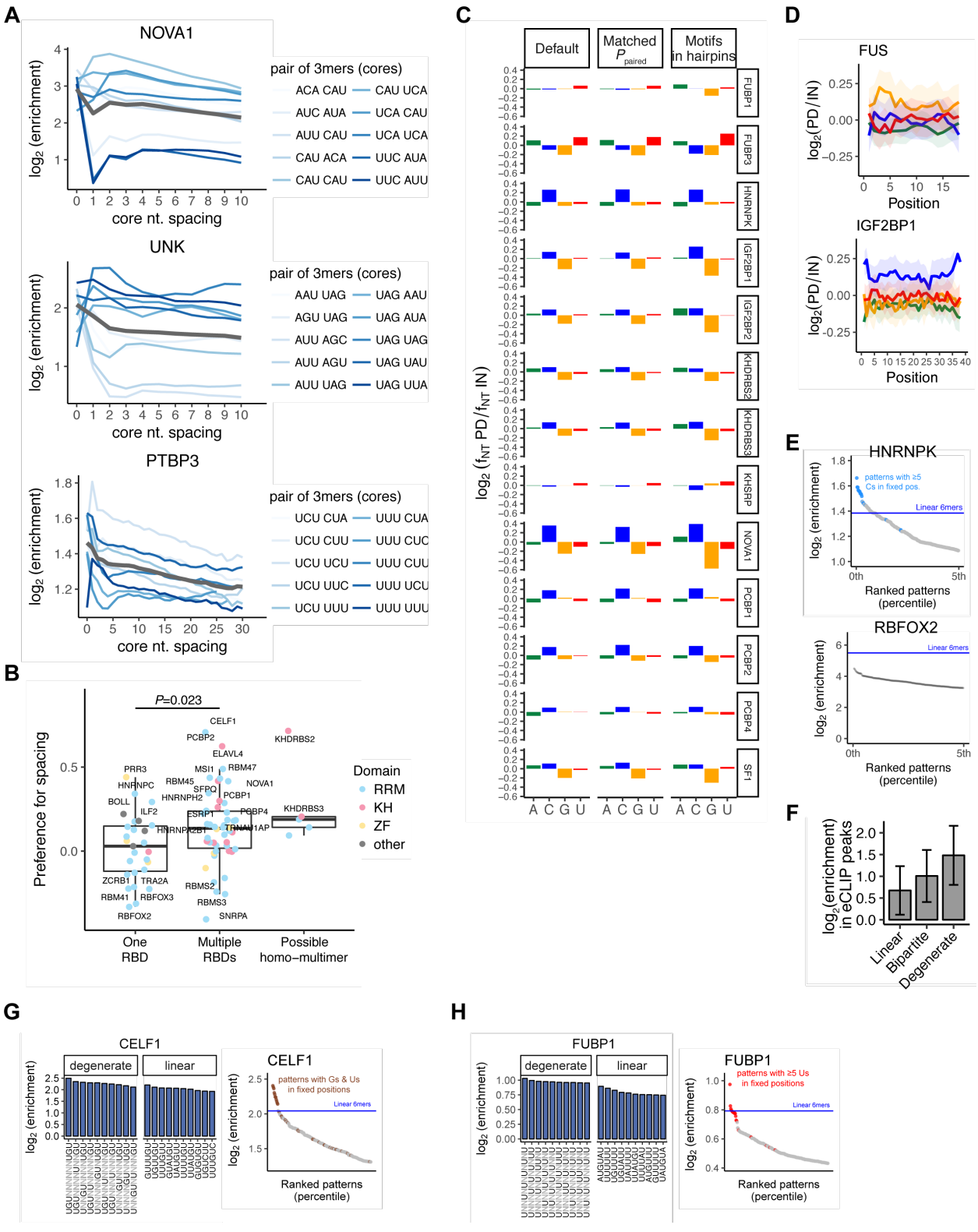

Figure 2-S5: **Bipartite core spacing, flanking nucleotide composition, and degenerate pattern binding preferences**

Figure 2-S5

(**A**) Enrichment ($\log_2$) as a function of the spacing between pairs of cores (3mers) for NOVA1 (top), UNK (middle), and PTBP3 (bottom). Individual 3mer pairs are shown in blue, and the average across all pairs is shown in gray (analogous to **Fig.** 2-5**A-B**).

(**B**) Relative preference for spacing (Methods) grouped by whether RBPs have a single RBD, multiple RBDs, or have been shown to multimerize in the literature with the domain responsible for multimerization included in the RBNS construct. P-value determined by t-test.

(**C**) Flanking nucleotide compositional preferences for RBPs with KH domains, with and without controls for RNA secondary structure. "Matched P_{paired}" contains motif occurrences that are sampled to have the same mean base-pairing probability in the input and pulldown libraries. "Motifs in hairpin" contains motif occurrences that have the 5-letter code "HHHHH" in the minimum free energy structure in both the input and pulldown libraries.

(**D**) Enrichment for particular nucleotides in reads with no high-affinity motifs for FUS (top) and IGF2BP1 (bottom).

(**E**) Enrichments for degenerate patterns of length 12 with 6 fixed bases shown for HNRNPK (top) and RBFOX2 (bottom). Only the top 5% of degenerate patterns are shown and the red lines indicate the enrichment of the linear 6mers. For HNRNPK, degenerate patterns where 5 out of the 6 fixed positions are C are in blue.

(**F**) Enrichment of HNRNPK motifs in eCLIP data. Linear: top 10 6mers; Bipartite: top 10 spaced 6mers; Degenerate: top 10 degenerate 12mers with 6 Cs (those shown in **Fig.** 2-5**H** left).

(**G, H**) Enrichments for top 10 degenerate 12mers with 6 Ns for CELF1 (**G**) and FUBP1 (**H**).

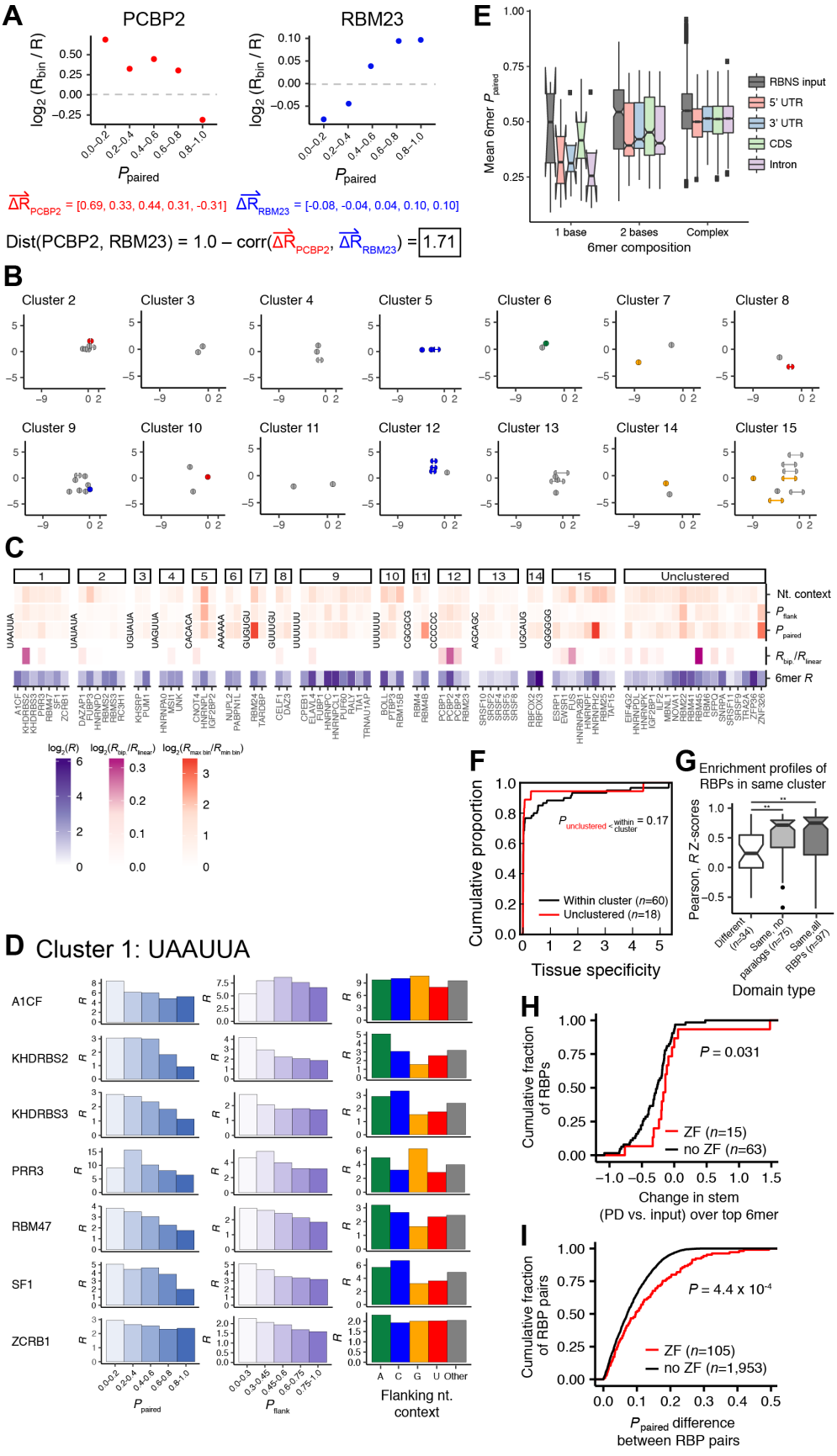

Figure 2-S6: **Sequence context effects on RBP binding**

Figure 2-S6

(**A**) Example of the distance between feature-specific R value profiles for PCBP2 and RBM23 for P_{paired}.

(**B**) Dispersal of specificities as in **Fig.** 2-6**B** for all other clusters.

(**C**) From bottom to top: R value for the 'exemplar' 6mer for each RBP (6mer denoted to the left of each cluster), $R_{\text{bipartite}}/R_{\text{linear}}$ 6mer if a bipartite motif was significantly preferred in **Fig.** 2-5**C**; the ratio of maximum R to minimum R for feature-specific R values over 5 context bins for each of the context features (P_{paired}, P_{flank}, and flanking nucleotide composition).

(**D**) For cluster 1 RBPs bound to UAAUUA, from left to right: R by P_{paired} bin (as in **Fig.** 2-4**C**); R by P_{flank} bin; and R by nucleotide context bin.

(**E**) Mean average P_{paired} over 6mers in RBNS input 20mers and human transcript regions (5' UTR, 3' UTR, CDS, Intron), separated by whether the 6mer is a homopolymer ('1 base'), contains two distinct bases ('2 bases'), or contains three or four distinct bases ('Complex').

(**F**) Tissue specificity of gene expression profile in 40 human tissues from the GTEx consortium for RBPs within a motif cluster versus those unclustered. P-value determined by Wilcoxon rank-sum test.

(**G**) Distribution of Pearson correlations of the R Z-scores of top 5mers for RBPs within the same motif cluster, separated by RBP pairs that have different RBD types (left) or the same RBD type (right: all RBPs; center: no paralogs). The R Z-scores of the top 18 5mers (i.e., the median number of enriched 5mers over all RBPs) for each RBP pair were used, and RBPs with only 1 RBD type were included. $^{**}P < 0.05$ by Wilcoxon rank-sum test.

(**H**) Change in stem content ($= f_{\text{stem,PD}} \times \log_2(f_{\text{stem,PD}}/f_{\text{stem,IN}})$, where $f_{\text{stem,PD}}$ is shown for RBM22 and ZNF326 in **Fig.** 2-4**E-F**) averaged over the top 6mer of each RBP, separated by RBPs that do vs. do not contain a ZNF. P-value determined by Wilcoxon rank-sum test.

(**I**) Comparison of differences in P_{paired} preferences among pairs of RBPs ($|P_{\text{paired}}$ $\text{ratio}_{\text{RBP1}}$ - P_{paired} $\text{ratio}_{\text{RBP2}}|$, where P_{paired} ratio is $P_{\text{paired,PD}}/P_{\text{paired,IN}}$ averaged over the top 6mer as shown in the right bar of **Fig.** 2-S4**A**), separated by RBPs that do vs. do not contain a ZNF. P-value determined by Wilcoxon rank-sum test.

2.6 Methods

2.6.1 Cloning of RNA binding protein domains

In most cases, RBPs were selected from a curated set of high-confidence annotations consisting of factors with well-defined RNA binding domains or with previous experimental evidence of RNA binding (Van Nostrand et al. [2017]). Regions of each protein containing all RBDs plus ∼50 amino acids flanking the RBD were cloned into the pGEX6 bacterial expression construct (GE Healthcare). A list of all constructs generated and primer sequences used is given in Table S1.

2.6.2 Bacterial expression and protein purification

Transformed Rosetta Cells (Novagen) were cultured in SuperBroth until optical density reached 0.6, cultures were transferred to 4°C and allowed to cool. Protein expression was induced for 14-20 hrs with IPTG at 15°C. Cells were pelleted, lysed (Qproteome Bacterial Protein Prep Kit, Qiagen) for 30 min in the presence of protease inhibitor cocktail (Roche), sonicated and clarified by centrifuging at >8,000 rpm, passed through a .45 μM filter (GE), and purified using GST-sepharose in either column format (GST-trap FF, GE) or 96-well format (GSTrap 96-well Protein Purification Kit, GE). Generally, 250 mL bacterial cultures used for column purifications and 50 mL for 96-well plate purifications (note: 8 wells of a 96-well plate were used per protein so that up to 12 proteins were purified per plate at a time). Eluted proteins were concentrated by centrifugation (Amicon Ultra-4 Centrifugal Filter Units) and subjected to buffer exchange (Zeba Spin Desalting Columns, 7K MWCO, Life Technologies) into Final Buffer (20 mM Tris pH 7, 300 mM KCl, 1 M DTT, 5 mM EDTA, 10% glycerol). Proteins were quantified using Bradford Reagent (Life Technologies), and purity and quality of protein was assessed by PAGE followed by Coomassie staining (all gels are available at https://www.encodeproject.org/search/?type=Experiment&assay_title=RNA+Bind-N-Seq&assay_title=RNA+Bind-n-Seq).

2.6.3 Production of random RNAs by *in vitro* transcription

Single-stranded DNA oligonucleotide and random template were synthesized (Integrated DNA Technologies) and gel-purified as previously described (Lambert et al. [2014]). Synthesis of random region of the template DNA oligo was hand-mixed to achieve balanced base composition. An oligo matching the T7 promoter sequence was annealed to the random template oligo by mixing in equal parts, bringing it to 70°C for 2 min, and allowing it to cool by placing at room temperature.

T7 Template:
CCTTGACACCCGAGAATTCCA(N)20GATCGTCGGACTGTAGAACTCCCTATAGTGAGTCGTA

T7 oligo: TAATACGACTCACTATAGGG

RNA was synthesized by transcribing 6 μL of 25 μM annealed template and T7 oligo in a 100 μL reaction (Hi-Scribe T7 transcription kit (NEB) according to manufacturer's protocol) or with a custom protocol using T7 polymerase (NEB) for larger-scale preps. RNAs were then DNase-treated with RQ1 (Promega) and subjected to phenol-chloroform extraction. RNA was suspended in nuclease free water and resolved on a 6% TBE-Urea gel (Life Technologies). RNA was excised and gel-extracted as previously reported (Lambert et al. [2014]). RNA was aliquoted and stored at -80°C.

Final transcribed RNA with sequencing adapters:
GGGGAGUUCUACAGUCCGACGAUC$(N)_{20}$UGGAAUUCUCGGGUGUCAAGG

2.6.4 RNA Bind-n-Seq Assay

All steps of the following binding assay were carried out at 4°C. Dynabeads MyOne Streptavidin T1 (Thermo) were washed 3X in binding buffer (25 mM tris pH 7.5, 150 mM KCl, 3 mM $MgCl_2$, 0.01% tween, 500 μg/mL BSA, 1 mM DTT). 60 μL of beads per individual protein reaction were used. 60 μL RBP diluted (see below for protein concentrations used) in binding buffer were allowed to equilibrate for 30 minutes at 4°C in the presence of 60 μL of washed Dynabeads MyOne Streptavidin T1. After 30 min of incubation, 60 μL of random RNA diluted in binding buffer was added, bringing the total reaction volume to 180 μL. The

final concentration per reaction of each of the components was 1 μM RNA; 5, 20, 80, 320 or 1300 nM of RBP; and 60 μL of Dynabeads MyOne Streptavidin T1 stock slurry washed and prepared in binding buffer. Each reaction was carried out in a single well of a 96-well plate. After 1 hr, RBP-RNA complexes were isolated by placing 96-well plate on a magnetic stand for 2 min. Unbound RNA was removed from each well and the bound RNA complexes were washed with 100 μL of wash buffer (25 mM tris pH 7.5, 150 mM KCl, 0.5 mM EDTA, 0.01% tween). Immediately after adding wash buffer the plate was placed on the magnet and wash was removed after $\sim$1 minute. This procedure was repeated 3 times. RBP-RNA complexes were eluted from Dynabeads MyOne Streptavidin T1 by incubating reaction at room temperature for 15 minutes in 25 μL of elution buffer (4 mM biotin, 1x PBS), the eluate was collected, the elution step was repeated, and eluates were pooled. RNA was purified from elution mixture by adding 40 μL AMPure Beads RNAClean XP (Agencourt) beads and 90 μL of isopropanol and incubating for 5 minutes. 96-well plate was placed on a magnetic stand and supernatant was discarded. Beads were washed twice with 80% ethanol, dried, and RNA was eluted in 15 μL of nuclease-free water. The extracted RNA was reverse transcribed into cDNA with Superscript III (Invitrogen) according to manufacturer's instructions using the RBNS RT primer. To prepare the input random library for sequencing, 0.5 pmol of the RBNS input RNA pool was also reverse transcribed. To make Illumina sequencing libraries, primers with Illumina adapters and sequencing barcodes were used to amplify the cDNA by PCR using Phusion DNA Polymerase (NEB) with 10-14 PCR cycles. PCR primers always included RNA PCR 1 (RP1) and one of the indexed primers as previously reported (Lambert et al. [2014]). PCR products were then gel-purified from 3% agarose gels and quantified and assessed for quality on the Bioanalyzer (Agilent). Sequencing libraries for all concentrations of the RBP as well as the input library were pooled in a single lane and sequenced on an Illumina HiSeq 2000 instrument.

2.6.5 RNA Bind-n-Seq data processing and motif logo generation

RBNS $k$mer enrichments (R values) were calculated as the frequency of each $k$mer in the pulldown library reads divided by its frequency in the input library; enrichments from the pulldown library with the greatest enrichment were used for all analyses of each respective

RBP. Mean and standard deviation of R values were calculated across all 4^k $k$mers for a given k to calculate the RBNS Z-score for each $k$mer.

RBNS motif logos were made from following iterative procedure on the most enriched pulldown library for $k = 5$: the most enriched $k$mer was given a weight equal to its enrichment over the input library (=R-1), and all occurrences of that $k$mer were masked in both the pulldown and input libraries so that stepwise enrichments of subsequent $k$mers could be used to eliminate subsequent double counting of lower-affinity 'shadow' $k$mers (e.g., only GGGGA occurrences not overlapping a higher-affinity GGGGG would count towards its stepwise enrichment). All enrichments were then recalculated on the masked read sets to obtain the resulting most enriched kmer and its corresponding weight (=stepwise R-1), with this process continuing until the enrichment Z-score (calculated from the original R values) was less than 3. All $k$mers determined from this procedure were aligned to minimize mismatches to the most enriched $k$mer, with a new motif started if the $k$mer could not be aligned to the most enriched $k$mer in one of the following 4 ways: one offset w/ 0 mismatches (among the 4 overlapping positions); 1 offset w/ 1 mismatch; no offset w/ 1 mismatch; 2 offsets w/ 0 mismatches. The frequencies of each nucleotide in the position weight matrix, as well as the overall percentage of each motif, were determined from the weights of the individual aligned $k$mers that went into that motif; empty unaligned positions before or after each aligned $k$mer were given pseudocounts of 25% each nucleotide, and outermost positions of the motif logo were trimmed if they had had unaligned total weight >75%. To improve the robustness of the motif logos, the pulldown and input reads were each divided in half and the above procedure was performed independently on each half; only $k$mers identified in corresponding motif logos from both halves were included in the alignments to make the final motif logo (weight of each $k$mer averaged between the two halves). In **Fig.** 2-2**A**, only the top RBNS motif logo is shown if there were multiple (all motifs displayed on the ENCODE portal within the "Documents" box of each experiment, with the proportion of each motif logo determined by computing the relative proportion of each motif's composite $k$mer weights). Motif logos were made from the resulting PWMs with Weblogo 2.0 (Crooks et al. [2004]). In addition to those displayed for 5mers with a Z-score=3 cutoff, for comparison motif logos were also made using: 5mers with Z-score=2 cutoff, 6mers with Z-score=2 cutoff, and 6mers with Z-score=3

cutoff; additionally, different rules of when to start a new logo vs. add to an existing one were tried. Logos for 5mers with Z-score=3 cutoff and the rules for starting a new motif described above appeared to strike the best balance of capturing a sufficient number of *k*mers to accurately represent the full spectrum of the RBP's binding specificity but did not create a number of secondary, largely similar motifs, and thus these parameters were used across all 78 RBPs.

The RBNS pipeline is available at: https://bitbucket.org/pfreese/rbns_pipeline. More specialized software is being prepared for release to coincide with publication.

2.6.6 Clustering of RBNS motifs

A Jensen-Shannon divergence (JSD)-based similarity score between each pair of top RBNS motif logos was computed by summing the score of the j overlapping positions between RBP A and RBP B:

$$\sum_{\text{aligned pos. } i=1,\ldots,j} \text{info}_{\text{A,i}} \times \text{info}_{\text{B,i}} \times \left(1 - \sqrt{\text{JSD}[\overrightarrow{ACGU}_{\text{A,i}} || \overrightarrow{ACGU}_{\text{B,i}}]}\right)$$

where $\text{info}_{\text{A,i}}$ is the information content in bits of motif A at position i and $\overrightarrow{ACGU}_{\text{A,i}}$ is the vector of motif A frequencies at position i (vectors sum to 1).

This score rewards positions with higher information content (scaled from positions positions with 100% one nucleotide given maximum weight to degenerate positions with 25% each nucleotide given zero weight) and more aligned positions (more positions j contributing to the summed score).

This similarity score was computed for each possible overlap of the two logos (subject to at least four positions overlapping, i.e., $j \geq 4$), and the top score with its corresponding alignment offset was used. The matrix of these scores were normalized to the maximum score over all RBP pairs and clustered using the linkage function with centroid method in scipy.cluster.hierarchy to obtain the dendrogram shown in **Fig.** 2-2**A**, with the 15 RBP groupings derived from a manually-set branch length cutoff.

This branch length cutoff was chosen to balance the competing interests of maximizing

the number of paralogous proteins within the same cluster (more stringent cutoffs eliminated PCBP4 from the cluster containing PCBP1 and PCBP2; it also did not include RBM4 and RBM4B to be in a cluster) and minimizing differences between primary motifs within the same cluster (less stringent cutoffs included the distinct UAG- containing MSI1/UNK/HNRNPA0 motifs within the same cluster as the AU-rich RBPs, for example).

2.6.7 Comparison with RNAcompete

5mer scores were derived from publicly available 7mer Z-scores by computing the mean across all 7mers containing a given 5mer (http://hugheslab.ccbr.utoronto.ca/supplementary-data/RNAcompete_eukarya/). Correlations between RBNS and RNA-compete experiments were computed by taking the Pearson correlation of Z-scores for all 5mers which had a Z-score ≥ 3 for at least one of the 31 RBPs in common between both assays.

2.6.8 Overlap of RBNS 6mers with splicing and stability regulatory elements

Splicing regulatory elements were taken from: ESS and ESE: Ke et al. [2011] and Rosenberg et al. [2015]; ISE: Wang et al. [2012]; ISS: Wang et al. [2013]. 3' UTR regulatory 6mers were derived from Oikonomou et al. [2014]. Only 6mers with $\geq$100 occurrences across all designed sequences were used (totaling 1303 6mers) in order to derive a mean 6mer score with sufficient coverage in different contexts. 6mer repressor and activator scores were obtained by averaging scores ($\log_2$(frequency) as described in the original manuscript) across all oligos containing that 6mer in the low (L10) and high (H10) Dual-reporter Intensity Ratio bins, respectively. Activator and repressor scores were averaged across both replicates (Libraries A and B). 6mers with an overall score ≥ 0.25 were used, where regulatory score=|$\log_2$(repressor score)-$\log_2$(activator score)|.

2.6.9 Analysis of eCLIP for motif discovery, regulation and overlapping targets

eCLIP datasets were produced by the Yeo Lab through the ENCODE RBP Project and are available at:
https://www.encodeproject.org/search/?type=Experiment&assay_title=eCLIP.

For all analyses, only eCLIP peaks with an enrichment over input ≥ 2 were used. Peaks were also extended 50 nucleotides in the 5' direction as the 5' start of the peak is predicted to correspond to the site of crosslink between the RBP and the RNA.

To produce eCLIP logos in a similar manner for comparison with RBNS logos, an analogous procedure to creating the RBNS motif logos was carried out on the eCLIP peak sequences: the two halves of the RBNS pulldown reads were replaced with the two eCLIP replicate peak sequences, and the input RBNS sequences were replaced by random regions within the same gene for each peak that preserved peak length and transcript region (5' and 3' UTR peaks were chosen randomly within that region; intronic and CDS peaks were shuffled to a position within the same gene that preserved the peak start's distance to the closest intron/exon boundary to match sequence biases resulting from CDS and splice site constraints). The enrichment Z-score threshold for 5mers included in eCLIP logos was 2.8, as this threshold produced eCLIP logos containing the most similar number of 5mers to that of the Z=3 5mer RBNS logos. Each eCLIP motif logo was filtered to include only 5mers that occurred in both corresponding eCLIP replicate logos. eCLIP motif logos were made separately for all eCLIP peaks, only 3' UTR peaks, only CDS peaks, and only intronic peaks, with the eCLIP logo of those 4 (or 8 if CLIP was performed in both cell types) with highest similarity score to the RBNS logo shown in **Fig.** 2-S3**C**, where the similarity score was the same as previously described to cluster RBNS logos. To determine overlap significance of RBNS and eCLIP, a hypergeometric test was performed with the 5mers in all (not just the top) logos for: RBNS logo 5mers, eCLIP logo 5mers (for peaks in the region with highest similiarity score to the RBNS logo), and 5mers in their intersection among the background of all 1,024 5mers; overlap was deemed significant if $P<0.05$.

All eCLIP/RBNS comparisons were for the same RBP with the following exceptions in

which the eCLIP RBP was compared to its paralogous RBNS protein: KHDRBS2 (KHDRBS1 RBNS); PABPN1 (PABPN1L RBNS); PTBP1 (PTBP3 RBNS); PUM2 (PUM1 RBNS); and RBM15 (RBM15B RBNS).

For **Fig.** 2-3**G**, the Pearson correlation between eCLIP experiments was assessed by computing the mean eCLIP coverage across 3' UTRs of all genes. 3' UTRs were split into windows of ∼100 nucleotides and the mean base-wise coverage (eCLIP coverage divided by input coverage) was calculated in each window. Pairs of RBPs were assigned as paralogs according to their classification in Ensembl. Pairs of RBPs were assigned as having overlapping motifs if at least 2 of their 5 top 5mers overlapped; RBPs with specificities determined from RBNS and RNAcompete (Ray et al. [2013]) were pooled.

2.6.10 Analysis of RNA-seq datasets for regulation and RBNS Expression & Splicing Maps

RNA-seq after shRNA knockdowns of individual RBPs in HepG2 and K562 cells (two KD and two control RNA-seq samples per RBP) were produced by the Graveley Lab as part of the ENCODE RBP Project and are available at:
https://www.encodeproject.org/search/?type=Experiment&assay_title=shRNA+RNA-seq.

Splicing changes upon KD were quantified with MATS (Shen et al. [2012]), considering only skipped exons (SEs) with at least 10 inclusion + exclusion junction-spanning reads and a Ψ between 0.05 and 0.95 in the averaged control and/or KD samples. SEs that shared a 5' or 3' splice site with another SE (i.e., those that are part of an annotated A3'SS, A5'SS, or Retained Intron) were eliminated. If multiple pairs of upstream & downstream flanking exons were quantified for an SE, only the event with the greatest number of junction-spanning reads was used. SEs significantly excluded or included upon KD were defined as those with a P-value < 0.05 and $|\Delta\Psi| \geq 0.05$. Control SEs upon KD were those with a P-value=1 and $|\Delta\Psi| \leq 0.02$.

Differentially expressed genes upon KD were called from DEseq2 (Love et al. [2014]), considering genes that had a 'baseMean' coverage of at least 1.0 and an adjusted P-value $<$ 0.05 and $|\log_2(\text{FC})| \geq 0.58$ (1.5-fold up or down upon KD). Candidate control genes upon

KD were taken from those with a P-value > 0.5 and $|\log_2(\text{FC})| \leq 0.15$; from this set of genes, a subset matched to the deciles of native (i.e., before KD) gene expression levels of the differentially expressed genes was used. The last 50 nt of each gene's ORF and 3' UTR sequence were taken from the Gencode version 19 transcript with the highest expression in the relevant cell type (HepG2 or K562).

'RBNS splicing maps' were made by taking the three sets of SEs included, excluded, or control upon KD and extracting their exonic and upstream/downstream flanking 250 nt sequences. At each position of each event, it was determined whether the position overlapped with one of the top 10 RBNS 5mers for that RBP in any of the five registers overlapping the position. Then to determine if the RBNS density was significantly higher or lower for included/excluded SEs at a position relative to control SEs at that position, the number of positions in a 20 bp window on each side (total 41 positions) covered by RBNS motifs was determined for each of the events, with significance determined by P-value<0.05 in a Wilcoxon rank-sum test on the control vs. changed events in the desired direction upon KD. Exonic regions were deemed to have ESE or ESS RBNS regulatory activity if 20 of the 100 exonic positions among SEs excluded or included upon KD, respectively, had significantly higher RBNS motif coverage than control SEs. The upstream and downstream intronic regions were each individually deemed as ISE or ISS regions if 50 of the 250 intronic positions had significantly higher RBNS motif coverage. For each significant region for each RBP, the ratios of $\log_2$(RBNS density over changing SEs/RBNS density over control SEs) of all significant positions in that region were summed, and the maximum value was normalized to 1 over all RBPs.

'RBNS stability maps' were made in an analogous manner, but for genes up- or down-regulated compared to control genes upon KD. The 3' UTR sequence was divided into 100 segments of roughly equal length and the proportion of positions covered by RBNS motifs in each segment were used for each bin of the meta-3' UTR. An RBP was deemed to have significant RBNS regulatory activity if 10 of the 100 positions of the meta-3' UTR for up- or down-regulated genes had increased RBNS density relative to control genes.

2.6.11 Generation of random sets of ranked 6mer lists with edit distances to top 6mer matching RBNS

Because the ranked lists of top enriched $k$mers (e.g., the top 15 6mers) are highly constrained depending on what the most enriched $k$mer is (e.g., 6mers 2-15 are typically Hamming distance of 1 and/or shifted by 1 from the top 6mer), as background sets for comparison to actual RBNS 6mer lists we sought to create groups of 6mers that matched the observed RBNS patterns of Hamming distances and shifts from the top 6mer for any given randomly selected $k$mer. To do this, for each of the 78 RBNS experiments we first calculated the edit distance from 6mer_i to 6mer_1, where 6mer_1 is the most enriched 6mer and i=2, ..., 15 is the ith enriched 6mer (e.g., 6mer_8 might have a mismatch at position two compared to 6mer_1 and then be shifted to the right by 1 position). Then, for all 4,096 starting 6mers, we created 78 ranked lists of 15 6mers, each of which matched the observed edit distances to the top 15 list of an actual RBNS experiment. The expected number of network edges in **Fig.** 2-2**B**, and the 'random' number of edges in **Fig.** 2-2**D** were performed by selecting random lists from these 4,096×78 possibilities.

2.6.12 RBNS RBP groups without paralogs or RBPs with any RBD pair sharing 40% identity

- No Paralogs (n =52):

  A1CF, BOLL, CELF1, CNOT4, CPEB1, DAZ3, EIF4G2, ELAVL4, ESRP1, EWSR1, FUBP1, HNRNPA2B1, HNRNPC, HNRNPK, HNRNPL, IGF2BP1, ILF2, MBNL1, NUPL2, PABPN1L, PRR3, PTBP3, PUM1, RBFOX2, RBM15B, RBM22, RBM23, RBM24, RBM25, RBM4, RBM41, RBM45, RBM47, RBM6, RBMS2, RC3H1, SF1, SFPQ, SNRPA, SRSF10, SRSF11, SRSF2, SRSF4, SRSF8, TARDBP, TIA1, TRA2A, TRNAU1AP, UNK, ZCRB1, ZFP36, ZNF326

- No RBPs sharing >40% identity among any RBDs (n =47):

  A1CF, BOLL, CELF1, CNOT4, CPEB1, EIF4G2, ELAVL4, EWSR1, FUBP3, HNRNPA0, HNRNPCL1, HNRNPDL, HNRNPH2, HNRNPL, IGF2BP1, ILF2, KHDRBS3,

MBNL1, NOVA1, NUPL2, PABPN1L, PCBP2, PRR3, PTBP3, PUF60, PUM1, RBFOX3, RBM15B, RBM22, RBM24, RBM25, RBM41, RBM45, RBM4B, RBM6, RBMS2, SFPQ, SNRPA, SRSF11, SRSF8, SRSF9, TARDBP, TIA1, TRA2A, TRNAU1AP, ZFP36, ZNF326

Pairwise RBD alignments were performed using ClustalW2 (Larkin et al. [2007]) and percent identities (as shown in **Fig.** 2-1**C** and **Fig.** 2-5**D**) were calculated as the percentage of identical positions relative to the number of ungapped positions in the alignments.

2.6.13 Network map of overlapping affinities

The lists of top 15 6mers for each RBP were intersected to get the number in common - those with 2 or more were deemed significant and connected by an edge ($P < 0.05$ by hypergeometric test, as well as by simulations based on the empirical distribution from random sets of ranked 6mer lists with edit distances to top 6mer matching RBNS as described above). The resulting network was visualized with Cytoscape (Shannon et al. [2003]).

2.6.14 Motif entropy analysis

To construct a set of 'simulated' motifs that matches the overall nucleotide composition of the 78 RBNS motifs but removes any positional correlations within a motif, individual columns of each RBNS motif (including all motifs for an RBP if there was more than one) were pooled to be sampled from. A 'simulated' motif was constructed by randomly sampling 5 or 6 columns (with probability 2/3 and 1/3, respectively, to roughly match the lengths of RBNS motifs) from this pool and concatenating them, repeated to construct 100,000 shuffled motifs.

The frequency of the four bases in each logo was calculated by averaging over all positions in the motif. This frequency vector (=[f_A, f_C, f_G, f_U], $f_A+f_C+f_G+f_U=1$) was mapped onto a square containing corners at [+/-1, +/-1] using two different orderings of the 4 corners, which together contain all 6 dinucleotide combinations (AC, AG, AU, CG, CU, GU) as edges:

1. Purine/Pyrimidine diagonals:

A U

C G

$$\vec{A} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$\vec{C} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

$$\vec{G} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$\vec{U} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

2. Purine/Pyrimidine edges:

C U

A G

$$\vec{A} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

$$\vec{C} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$\vec{G} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$\vec{U} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

To map the frequency vector to its coordinates (u, v) within the unit circle, the frequency vector was normalized to the largest component:

$$\vec{F} = [F_A, F_C, F_G, F_U] = [f_A/f_{\max}, f_C/f_{\max}, f_G/f_{\max}, f_U/f_{\max}],$$

where $f_{\max} = \max(f_A,\ f_C,\ f_G,\ f_U)$, and (u, v) was computed as:

$$\begin{aligned} |\vec{F}| &= \sqrt{F_A^2 + F_C^2 + F_G^2 + F_U^2} \\ \begin{bmatrix} u \\ v \end{bmatrix} &= \frac{F_A\vec{A} + F_C\vec{C} + F_G\vec{G} + F_U\vec{U}}{\sqrt{2} \times |\vec{F}|} \end{aligned}$$

The elliptical grid mapping was used to convert the (u, v) coordinates within the unit circle to the corresponding position (x, y) within a square containing corners at [+/-1, +/-1]:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\left(\sqrt{2 + u^2 - v^2 + 2\sqrt{2}u} - \sqrt{2 + u^2 - v^2 - 2\sqrt{2}u}\right) \\ \frac{1}{2}\left(\sqrt{2 - u^2 + v^2 + 2\sqrt{2}v} - \sqrt{2 - u^2 + v^2 - 2\sqrt{2}v}\right) \end{bmatrix}$$

The simplex grid shown was divided into 11 equal parts along both dimensions, and the density in each of the 121 squares was computed for the 78 RBNS motifs and 100,000 shuffled motifs to get enrichments.

To determine significance via bootstrapping, 1,000 different shuffled motif distributions over the grid were computed. In each of the 1,000 bootstraps, the 100,000 shuffled motifs were drawn from a different starting pool of motif columns: rather than all 78 RBPs' motifs contributing once to the pool, a random sampling (with replacement) of the 78 RBPs was performed, and those motifs' columns served as the starting pool for the 100,000 shuffled motifs. The mean and standard deviation of these 1,000 bootstraps were computed for each margin, and margins for which the density of the 78 RBNS motif logos had a Z-score greater than 2 were marked significant (number of asterisks = Z-score, rounded down).

2.6.15 RNA secondary structure analysis

The RNA base-pairing probability was extracted from the partition function of RNAfold: "RNAfold -p –temp=X", where X was 4°C or 21°C depending on what temperature the binding reaction was conducted at (See Table S3) (Lorenz et al. [2011]). For each pulldown library, reads were randomly selected to match the distribution of C+G content among input reads; all enrichments were recalculated for these C+G-matched pulldown reads for **Fig.** 2-4, **Fig.** 2-S4, **Fig.** 2-6, and **Fig.** 2-S6. Reads were folded with the 5' and 3' adapters (24 and

21 nt, respectively), resulting in folded sequences of length 65 and 85 for 20mer and 40mer RBNS experiments, respectively.

Secondary structural element analyses were performed by using the forgi software package (Kerpedjiev et al. [2015]). For each read, to mirror the partition function rather than relying solely on the Minimum Free Energy structure, 20 random suboptimal structures with probabilities equal to their Boltzmann weights were sampled and averaged over ("RNAsubopt –temp=X –stochBT=20"). In **Fig.** 2-4**D**, 6mers counting toward: 'loop' were: H_6, M_6, I_6; 'stem' was S_6; 'bulged stem' were 6mers matching the pattern SXXXXS, where XXXX contained 1-3 S.

For **Fig.** 2-6**A**, **Fig.** 2-6**C**, **Fig.** 2-S6**C**, **Fig.** 2-S6**D**, bin limits for the motif structure analyses (P_{paired}) were: 0-0.2 (bin 1); 0.2-0.4 (bin 2); 0.4-0.6 (bin 3); 0.6-0.8 (bin 4); and 0.8-1.0 (bin 5). Bin limits for flanking structure analyses (P_{flank}) were: 0-0.3 (bin 1); 0.3-0.45 (bin 2); 0.45-0.6 (bin 3); 0.6-0.75 (bin 4); 0.75-1.0 (bin 5). P_{paired} was calculated as the average over the six positions of the 6mer; P_{flank} was calculated as the average over all other positions in the read. The continuous measures of preference for motif and flanking preference for the x- and y-axes in **Fig.** 2-6**B**, **Fig.** 2-6**D**, **Fig.** 2-S6**B** were computed as:

$$\begin{bmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} log_2(R_{\text{bin 1}}/R_{\text{original}}) \\ log_2(R_{\text{bin 2}}/R_{\text{original}}) \\ log_2(R_{\text{bin 3}}/R_{\text{original}}) \\ log_2(R_{\text{bin 4}}/R_{\text{original}}) \\ log_2(R_{\text{bin 5}}/R_{\text{original}}) \end{bmatrix}$$

RBNS structure profiles were compared to eCLIP structure profiles in the region with the greatest number of eCLIP peaks. Bound RBNS motifs were selected from the transcriptome region that showed the highest enrichment for the number of peaks (5' UTR/3' UTR/introns/CDS). Motifs that were not bound were selected from the same gene regions as bound motifs and matched for the same genes. Motifs were folded with 50 nucleotides of flanking sequence on both sides using RNAfold. Motifs (both bound and not bound) were then binned by their mean base-pairing probability (same bins as RBNS), and the fraction of bound motifs in each bin was computed. The monotonicity of R over P_{paired} bins for RBNS

and eCLIP was computed by taking all 10 comparisons over the 5 bins, adding 1 if R was greater in the higher P_{paired} bin or subtracting 1 if it was lower in the higher P_{paired} bin.

The structure of 6mers in random sequences (**Fig.** 2-4**H**) was calculated by creating random 65mers, folding them at 4°C, and computing the mean P_{paired} over each of the 15 6mers within the region corresponding to the random RBNS 20mer positions (i.e., positions 25-44 of the 65mer, inclusive). The structure of 6mers in human Gencode version 19 transcript regions (**Fig.** 2-S6**E**) was calculated by taking all consecutive 65 nt sequences (the length of the RBNS 20mer + adapter sequences) fully within one of the four respective transcript regions. Each 65 nt sequence was folded at 37°C, and the mean P_{paired} over each of the 15 6mers within the region corresponding to the random RBNS 20mer positions was calculated. The mean P_{paired} for each 6mer was then computed over all occurrences of that 6mer within the given transcript region, and the 6mer was classified as being composed of "1 base" (e.g., GGGGGG), "2 bases" (e.g., GAGGAA), or "Complex" (3+ unique bases, e.g., UGGAGU).

2.6.16 Determination of bipartite motifs

Enrichments were computed for all pairs of the top 10 enriched 3mers, with a spacer of length $i = 0 - 10$ (in total: $10 \times 10 \times (i + 1)$ combinations), where the enrichment was defined as the fraction of pulldown reads with a motif relative to the fraction of input reads with a motif. The enrichment for each spacing was computed as the mean enrichment of the 10 most enriched combinations of that particular spacing (**Fig.** 2-5**A-B**). Nucleotide composition of the spacer (as shown in **Fig.** 2-5**A-B**) was the mean nucleotide frequency across positions between both motif cores, relative to the corresponding nucleotide frequency between the same motif cores the input libraries. Preference for spacing (**Fig.** 2-S5**B**) was computed as the change in the mean enrichment for the top 10 spaced combinations ($i > 0$) relative to the mean enrichment of the top 10 non-spaced combinations ($i = 0$, i.e., top 10 6mers): $\log_2(\text{enrichment}_{\text{spaced}}/\text{enrichment}_{\text{linear}})$. Significance was determined by setting a False Discovery Rate (FDR) using 0 nM control libraries as follows: samples of 10 3mer cores were repeatedly drawn and the observed relative enrichments were used to set an FDR at each spacing s. Motif cores were sampled such that the relationships between sampled 3mers were the same as the relationship observed for that particular protein's enriched cores.

2.6.17 Assessment of flanking nucleotide compositional preferences

For a given RBP, we only considered (protein-bound and input) reads that: a) contained one of the top 5 enriched 5mers; b) contained no additional secondary motifs, where secondary motifs were the top 50 enriched 5mers or all 5mers with $R \geq 2$, whichever set was larger. The remaining protein-bound and input reads were then subsampled to match the distribution of motifs and the positions of those motifs along a read. These reads were further subsampled to match the distribution of mean base-pairing probabilities over the motif (bins used were [0- 0.1),[0.1-0.2),...,[0.9,1.0]). For the analysis in **Fig.** 2-S5**C**, protein-bound and input reads were instead subsetted only to reads where the motif was in a hairpin configuration (H_5 in the MFE). The flanking nucleotide enrichment was then determined by centering these reads on the motif and computing the relative enrichment ($= \log_2(f_{\text{PD,NT}}/f_{\text{IN,NT}})$) for each nucleotide at each position relative to the motif. We excluded the two nucleotides immediately adjacent to the motif on either side (to avoid capturing the extension of a core motif) as well as the first and last position of the random region in order to avoid certain nucleotide biases that can occur due to the presence of adaptor sequences. The overall enrichment (**Fig.** 2-5**G**) is the mean enrichment across all assessed positions, with significance assessed by a Wilcoxon rank-sum test.

Binding to mono- or dinucleotide rich sequence (e.g., **Fig.** 2-S5**E**) in absence of a motif was done analogously, except only using reads that did not contain any of the top 50 5mers or any 5mer with $R \geq 2$. Enrichments for degenerate patterns were calculated as the mean of the 10 best degenerate $k$mers matching that pattern (e.g., mean of top 10/4096 12mers matching CCNNNCCNNNCC in the example in **Fig.** 2-5**H**, 2-S5**H**). We first calculated enrichments for patterns where the fixed positions (e.g., CCCCCC in the previous example) contained only one or two nucleotides to assess which RBPs were biased towards binding to degenerate nucleotide-rich sequences, but later performed exhaustive searches where the fixed $k$mer was allowed to cover the entire sequence space (i.e., 4096 possible sequences in fixed positions $\times$ 210=(10 choose 4) patterns with 6 fixed positions and 6 internal Ns).

2.6.18 Filter binding assay

Filter binding assay was performed with the oligo UUU(CCUCUCUUUUCC)UUU, i.e., the pattern CCNCNCNNNNCC flanked by Us and with Ns substituted with Us to avoid creating high-affinity motifs and ensure the oligo was void of secondary structure. The negative control oligo used was U_{12}. Custom RNA oligonucleotides were synthesized by IDT (Integrated DNA Technologies) and RBPs were purified as described earlier (see Cloning of RNA binding protein domains). RNA was end-labeled with ^{32}P by incubating with Polynucleotide Kinase (NEB) according to manufacturer protocol. The assay was done following the protocol described in (Rio [2012]) for use with a 96-well dot-blot apparatus (Biorad). RBP and radio-labelled RNA were incubated in 50 μL binding buffer (500 μL 2M KCl, 10 μL 1M DTT, 400 μL 40% glycerol, 200 μL 1M Tris in 10 mL) for 1 hour at room temperature. Final concentration of RNA was 1 nM and protein concentration ranged from 10 nM-10 μM (three-fold serial dilutions spanning this range).

2.6.19 Calculation of feature-specific R values and relative entropy of context features

Feature-specific R values were calculated by assigning all 6mers into their respective bin for the feature under consideration for both the pulldown and input libraries, converting the counts into frequencies within each bin for both libraries, and computing the R value for the 6mer under consideration using the pulldown and input bin frequencies.

For **Fig.** 2-6**A**, bins used to compute feature-specific R values for each feature were the following:

- P_{paired}: bin 1=0-0.2; bin 2=0.2-0.4; bin 3=0.4-0.6; bin 4=0.6-0.8; bin 5=0.8-1.0
- P_{flank}: bin 1=0-0.3; bin 2=0.3-0.45; bin 3=0.45-0.6; bin 4=0.6-0.75; bin 5=0.75-1.0
- Core spacing: bin 1 = 0 nt spacing; bin 2 = 1 nt spacing; ... ; bin 11 = 10 nt spacing, where the spacing corresponds to the spacing between the two cores of a bipartite motif.

- Nucleotide context: 16 bins, where the first four bins are quartiles of the percentage of A content flanking a 6mer based on the composition of input reads (bins 5-8, 9-12, and 13-16 are analogous for C, G, and U content, respectively). Each 6mer occurrence was therefore counted 4 times, into the corresponding bin for each of the four nucleotides.

Feature-specific R values within each bin were compared to the overall R value of the 6mer without binning (i.e., $\log_2(R_{\text{bin}}/R_{\text{original}})$) to create the feature-specific enrichment profile for a particular context feature (example for P_{paired} for two RBPs in **Fig.** 2-S6**A**).

For **Fig.** 2-6**C** and **Fig.** 2-S6**C-D**, feature-specific R values were computed for the 'exemplar' 6mer for each group (i.e., the 6mer with the lowest summed ranks among all cluster members). The enrichments of the exemplar 6mer over five context bins were calculated for P_{paired} and P_{flank} as described above; the ratio of the maximum to minimum R values over these 5 bins for each feature was then computed. Nucleotide context bins were created using the empirical distribution of nucleotide flanking contents for reads with the same 6mer in the input according to: bin 1) 'high A' (flanking A content in the 75th percentile of input reads); bins 2), 3) and 4) 'high C', 'high G', and 'high U' (analogous to high A for the respective nucleotides); bin 5) 'other' (all other reads). As was done in **Fig.** 2-5**E-G**, only reads with one exemplar 6mer and no additional high affinity 6mers (top 100 6mers) were used.

For **Fig.** 2-6**E**, RNA folding was done as described above. To determine the $\log_2$ ratio of base-pairing probabilities, control U_5 occurrences were determined as previously described for each datatype (RBNS reads, eCLIP peaks, control exons in knockdown data). $\text{Log}_2(P_{\text{paired}}$ ratio) values were then bootstrapped from the P_{paired} distributions in pulldown relative to input for each datatype, with significance assessed by a Wilcoxon rank-sum test.

2.6.20 Tissue specificity of RBP gene expression

Tissue specificity was measured as the information content deviation from a uniform distribution among all tissues as in (Gerstberger et al. [2014]). For each RBP, the $\log_2$(TPM+1) was calculated for each of the 40 GTEx tissues (GTEx Consortium [2015]), and the tissue specificity was computed as the difference between the logarithm of the total number of

samples ($N = 40$) and the Shannon entropy of the expression values for an RBP:

$$S = H_{\max} - H_{\text{obs}} = \log_2(N) - \sum_{i=1,\ldots,N} \left[p_i \times \log_2(p_i) \right],$$

Where $p_i = x_i \; / \; \left(\sum_{i=1,\ldots,N} x_i \right)$ for $x_i = \log_2(\text{TPM}_i + 1)$ in sample i.

The data used for the analyses were obtained from dbGaP accession number phs000424.v2.p1 in Jan. 2015. TPMs were measured using kallisto (Bray et al. [2016]) on the following samples: Adipose-Subcutaneous: SRR1081567; AdrenalGland: SRR1120913; Artery-Tibial: SRR817094; Bladder: SRR1086236; Brain-Amygdala: SRR1085015; Brain- AnteriorCingulateCortex: SRR814989; Brain-CaudateBasalGanglia: SRR657731; Brain- CerebellarHemisphere: SRR1098519; Brain-Cerebellum: SRR627299; Brain-Cortex: SRR816770; Brain-FrontalCortex: SRR657777; Brain-Hippocampus: SRR614814; Brain- Hypothalamus: SRR661179; Brain-NucleusAccumben: SRR602808; Brain-SpinalCord: SRR613807; Brain-SubstantiaNigra: SRR662138; Breast-MammaryTissue: SRR1084674; Cervix: SRR1096057; Colon: SRR1091524; Esophagus: SRR1085211; FallopianTube: SRR1082520; Heart-LeftVentricle: SRR815517; Kidney-Cortex: SRR809943; Liver: SRR1090556; Lung: SRR1081283; MinorSalivaryGland: SRR1081589; Muscle-Skeletal: SRR820907; Nerve-Tibial: SRR612911; Ovary: SRR1102005; Pancreas: SRR1081259; Pituitary: SRR1077968; Prostate: SRR1099402; Skin: SRR807775; SmallIntestine: SRR1093314; Spleen: SRR1085087; Stomach: SRR814268; Testis: SRR1081449; Thyroid: SRR808886; Uterus: SRR820026; Vagina: SRR1095599.

Chapter 3

A Large-Scale Binding and Functional Map of Human RNA Binding Proteins

Under review, posted to bioRxiv on 8/23/17:

EL Van Nostrand[§], P Freese[§], GA Pratt[§], X Wang[§], X Wei[§], R Xiao[§], SM Blue, D Dominguez, NAL Cody, S Olson, B Sundararaman, L Zhan, C Bazile, LPB Bouvrette, J Chen, MO Duff, KE Garcia, C Gelboin-Burkhart, M Hochman, NJ Lambert, H Li, TB Nguyen, T Palden, I Rabano, S Sathe, R Stanton, J Bergalet, B Zhou, A Su, R Wang, BA Yee, AL Louie, S Aigner, X Fu, E Lecuyer, CB Burge, BR Graveley, GW Yeo. "A Large-Scale Binding and Functional Map of Human RNA Binding Proteins". http://doi.org/10.1101/179648

My contributions:

Analysis of comparison of *in vitro* and *in vivo* binding specificities and relationship between sequence-specific binding and regulation (**Fig.** 3-3, **Fig.** 3-S3, and **Fig.** 3-S4); association between TIA1 binding and RNA expression upon KD (**Fig.** 3-S5**e-g**); methods development of RNA splicing maps pipeline in conjunction with ELVN, GAP, and BAY (**Fig.** 3-5, **Fig.** 3-S6, and **Fig.** 3-S8); integrative analysis of RBP data types in cryptic exon suppression (**Fig.** 3-S1); tissue specificity of RBP expression (**Fig.** 3-8**e**, **Fig.** 3-S11);

writing of text section "*In vivo* binding is largely determined by *in vitro* binding specificity" with CBB as well as editing remainder of text.

3.1 Abstract

Genomes encompass all the information necessary to specify the development and function of an organism. In addition to genes, genomes also contain a myriad of functional elements that control various steps in gene expression. A major class of these elements function only when transcribed into RNA as they serve as the binding sites for RNA binding proteins (RBPs) which act to control post-transcriptional processes including splicing, cleavage and polyadenylation, RNA editing, RNA localization, translation, and RNA stability. Despite the importance of these functional RNA elements encoded in the genome, they have been much less studied than genes and DNA elements. Here, we describe the mapping and characterization of RNA elements recognized by a large collection of human RBPs in K562 and HepG2 cells. These data expand the catalog of functional elements encoded in the human genome by addition of a large set of elements that function at the RNA level through interaction with RBPs.

3.2 Introduction

RBPs have emerged as critical players in regulating gene expression, controlling when, where, and at what rate RNAs are processed, trafficked, translated, and degraded within the cell. They represent a diverse class of proteins involved in co- and post-transcriptional gene regulation (Gerstberger et al. [2014], Glisovic et al. [2008]). RBPs interact with RNA to form ribonucleoprotein complexes (RNPs), governing the maturation and fate of their target RNA substrates. Indeed, they regulate numerous aspects of gene expression including pre-mRNA splicing, cleavage and polyadenylation, RNA stability, RNA localization, RNA editing, and translation. In fact, many RBPs participate in more than one of these processes. For example, studies on the mammalian RBP Nova using a combination of crosslinking and immunopreciptation (CLIP)-seq and functional studies revealed that Nova not only regulates alternative splicing, but also modulates poly(A) site usage (Licatalosi et al. [2008]). Moreover, in contrast to regulation at the transcriptional level, post-transcriptional regulatory steps are often carried out in different subcellular compartments of the nucleus (e.g. nucleoli, nuclear speckles, paraspeckles, coiled bodies, etc.) and/or cytoplasm (e.g. P-bodies, endoplasmic reticulum, etc.) by RBPs that are enriched within these compartments. These regulatory roles are essential for normal human physiology, as defects in RBP function are associated with diverse genetic and somatic disorders, such as neurodegeneration, auto-immune defects, and cancer (Kao et al. [2010], King et al. [2012], Lagier-Tourenne et al. [2010], Nussbacher et al. [2015], Paronetto et al. [2007], Lukong et al. [2008], Martini et al. [2002]).

Traditionally, RBPs were identified by affinity purification of single proteins (Sonenberg et al. [1979a], Sonenberg et al. [1979b]). However, several groups have recently used mass spectrometry-based methods to identify hundreds of proteins bound to RNA in human and mouse cells (Baltz et al. [2012], Castello et al. [2012], Kwon et al. [2013], Brannan et al. [2016]). Recent censuses conducted by us and others indicate that the human genome may contain between 1,072 (Sundararaman et al. [2016]) and 1,542 (Castello et al. [2012]) RBP-encoding genes. This large repertoire of RBPs likely underlies the high complexity of post-transcriptional regulation, motivating concerted efforts to systematically dissect the binding properties, RNA targets, and functional roles of these proteins.

The dissection of RBP-RNA regulatory networks therefore requires the integration of multiple data types, each viewing the RBP through a different lens. *In vivo* binding assays such as CLIP-seq provide a set of candidate functional elements directly bound by each RBP. Assessments of *in vitro* binding affinity help understand the mechanism driving these interactions, and (as we show) improve identification of functional associations. Functional assays that identify targets whose expression or alternative splicing is responsive to RBP perturbation can then fortify evidence of function. For example, observation of protein binding by CLIP-seq within introns flanking exons whose splicing is sensitive to RBP levels in RNA-seq provides support for the RBP as a splicing factor and for the binding sites as splicing regulatory elements. *In vivo* interactions of RBPs with chromatin can also be assayed to provide insight into roles of some RBPs as transcription regulators and can provide evidence for co-transcriptional deposition of RBPs on target RNA substrates. The regulatory roles of RBPs are also informed by the subcellular localization properties of RBPs and of their RNA substrates. Furthermore, these data resources comprised of multiple RBPs profiled using the same methodology may be integrated to enable the identification of factor-specific regulatory modules, as well as a factor's integration into broader cellular regulatory networks, through novel integrated analyses.

3.3 Results

3.3.1 Overview of data and processing

To develop a comprehensive understanding of the binding and function of the human RBP repertoire, we used five assays to produce 1,076 replicated datasets for 352 RBPs (**Fig.** 3-1, Supplementary Table 1). The RBPs characterized by these assays have a wide diversity of sequence and structural characteristics and participate in diverse aspects of RNA biology (**Fig.** 3-1). Functionally, these RBPs are most commonly known to play roles in the regulation of RNA splicing (96 RBPs, 27%), RNA stability and decay (70, 20%), and translation (69, 20%), with 161 RBPs (46%) having more than one function reported in the literature. However, 82 (23%) of the characterized RBPs have no known function in RNA biology other than being annotated as involved in RNA binding (**Fig.** 3-1). Although 57% of the RBPs surveyed contain well-characterized RNA binding domains (RNA recognition motif (RRM), hnRNP K homology (KH), zinc finger, RNA helicase, ribonuclease, double-stranded RNA binding (dsRBD), or pumilio/FBF domain (PUM-HD)), the remainder possess either less well studied domains or lack known RNA binding domains altogether (**Fig.** 3-1). Each of the five assays used focused on a distinct aspect of RBP activity:

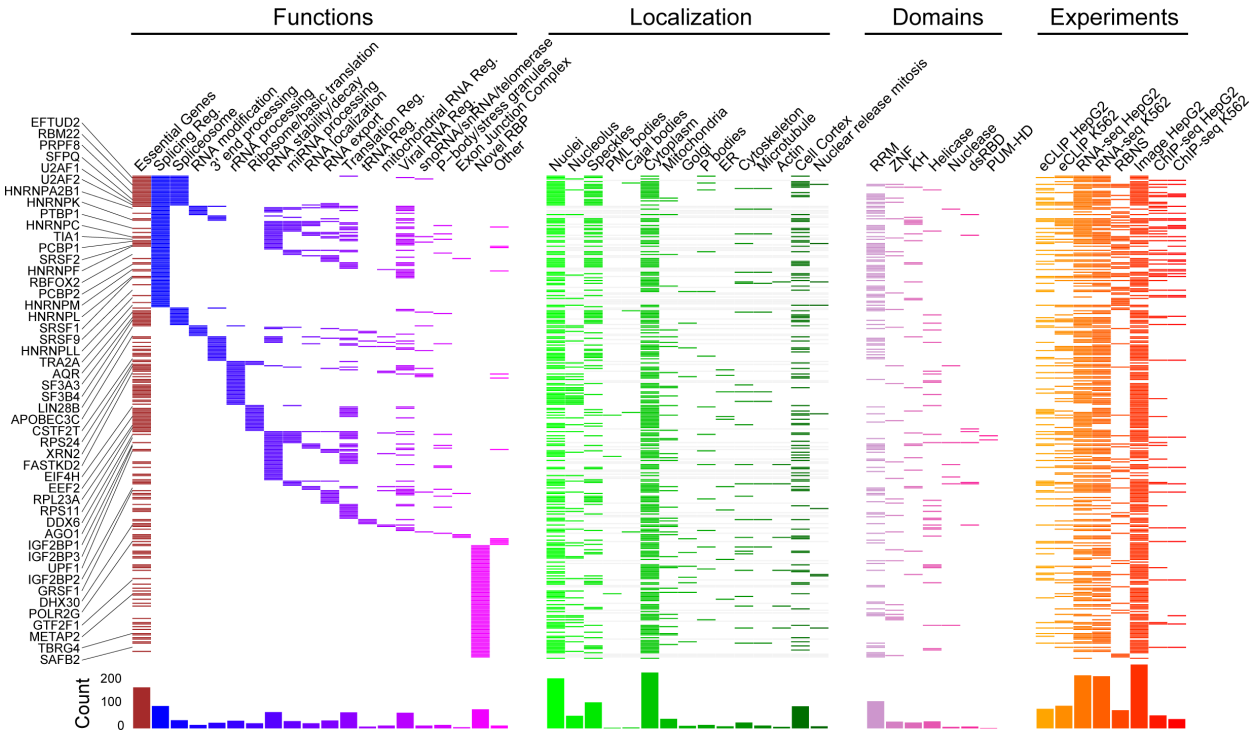

Figure 3-1: **Overview of experiments and data types**
The 352 RNA binding proteins (RBPs) profiled by at least one ENCODE experiment (orange/red) are shown, with localization by immunofluorescence (green), essential genes from CRISPR screening (maroon), manually annotated RBP functions (blue/purple), and annotated protein domains (pink). Histograms for each category are shown on bottom, and select RBPs highlighted in this study are indicated on left.

- Transcriptome-wide RNA binding sites of RBPs: We identified and validated hundreds of immunoprecipitation-grade antibodies that recognize human RBPs (Sundararaman et al. [2016]) and developed enhanced CLIP (eCLIP) followed by sequencing to efficiently identify RNA targets bound by these RBPs at scale (Van Nostrand et al. [2016]). We performed eCLIP to profile 97 RBPs in K562 cells and 84 RBPs in HepG2 cells, for a total of 126 RBPs (of which 55 were characterized in both cell types). This effort identified 717,765 significantly enriched peaks (relative to size-matched input controls for each RBP) that cover 16.5% of the annotated mRNA transcriptome and 2.4% of the pre-mRNA transcriptome (**Fig.** 3-2**d**, **Fig.** 3-S2**d**).

- RBP-responsive genes and alternative splicing events: To obtain insight into the functions of the RBP binding sites identified by eCLIP, we used shRNA- or CRISPR-mediated depletion followed by RNA-seq of 224 RBPs in K562 and 228 RBPs in HepG2 cells, for a total of 251 RBPs (of which 200 were characterized in both cell types). This effort identified 223,118 instances of RBP-mediated differential gene expression involving 14,281 genes affected upon knockdown of at least one RBP as well as 205,919 cases of RBP-mediated alternative splicing events involving 37,829 alternatively spliced events impacted upon knockdown of at least one RBP.

- *In vitro* RBP binding motifs: To identify the *in vitro* RNA sequence and structural binding preferences of RBPs, we developed a high-throughput version of RNA Bind-N-Seq (RBNS, Lambert et al. [2014]) that assays binding of purified RBPs to pools of random RNA oligonucleotides. This effort identified the binding specificities of 78 RBPs. Short oligonucleotides of length k=5 ($k$mers) with high RBNS affinity clustered into a single motif for about half of the RBPs assayed (37/78). The remaining RBPs had more complex patterns of binding, best described by two motifs (32/78), or even three or more motifs (9 RBPs). These data also indicate that many RBPs are sensitive to the sequence and RNA structural context in which motifs are embedded.

- RBP subcellular localization: Post-transcriptional gene regulation occurs in different intracellular compartments. For instance, rRNA maturation and pre-mRNA splicing primarily occur in sub-regions of the nucleus, whereas mRNA translation and default

mRNA decay pathways operate in the cytoplasm. To illuminate functional properties of RBPs in intracellular space, we took advantage of our validated antibody resource (Sundararaman et al. [2016]) to conduct systematic immunofluorescence (IF) imaging of 274 RBPs in HepG2 cells and 268 RBPs in HeLa cells, in conjunction with a dozen markers for specific organelles and subcellular structures. These data, encompassing 230,000 images and controlled vocabulary localization descriptors, have been organized within the RBP Image Database (http://rnabiology.ircm.qc.ca/RBPImage).

- RBP association with chromatin: Recent work has suggested that RBP association with chromatin may play roles in transcription and co-transcriptional splicing (Naftelberg et al. [2015], Ji et al. [2013]). To generate a large-scale resource of chromatin association properties for RBPs, we identified the DNA elements associated with 58 RBPs in HepG2 cells and 45 RBPs in K562 cells by ChIP-seq, for a total of 63 RBPs (of which 34 RBPs were characterized in both cell types). These experiments identified 622,443 binding locations covering 3.1% of the genome.

To facilitate integrated analyses, all data for each data type were processed by the same data processing pipeline, and consistent, stringent quality control metrics and data standards were uniformly applied to all experiments. Although only 8 RBPs were investigated using all five assays, 249 of the 352 RBPs (71%) were studied using at least two different assays and 129 (37%) were subjected to at least three different assays, providing opportunities for integrated analysis using multiple datasets. As an example of how these complementary datasets provide distinct insights into RNA processing regulation, we considered the Vascular Endothelial Growth Factor A gene (VEGFA) (**Fig.** 3-2**a**). Although the RBPs IGF2BP1 and IGF2BP3 both showed significant binding to the VEGFA 3' UTR, knockdown of IGF2BP1 increased VEGFA mRNA levels, while knockdown of IGF2BP2 decreased them, pointing to opposing regulatory effects. Knockdown of HNRNPK also yielded decreased VEGF mRNA, as well as a significant decrease in inclusion of exon 6. This splicing event is likely directly regulated by HNRNPK, as the flanking introns contain multiple HNRNPK eCLIP peaks, some of which contain the preferred binding motif for HNRNPK (GCCCA) identified in RBNS experiments. Similar integrated analysis can provide insight into mechanisms of cryptic exon

repression, illustrated by HNRNPL binding to a region downstream of a GTPBP2 cryptic exon that contains repeats of HNRNPL's top RBNS motif, contributing to production of GTPBP2 mRNA with a full-length open reading frame (**Fig.** 3-S1).

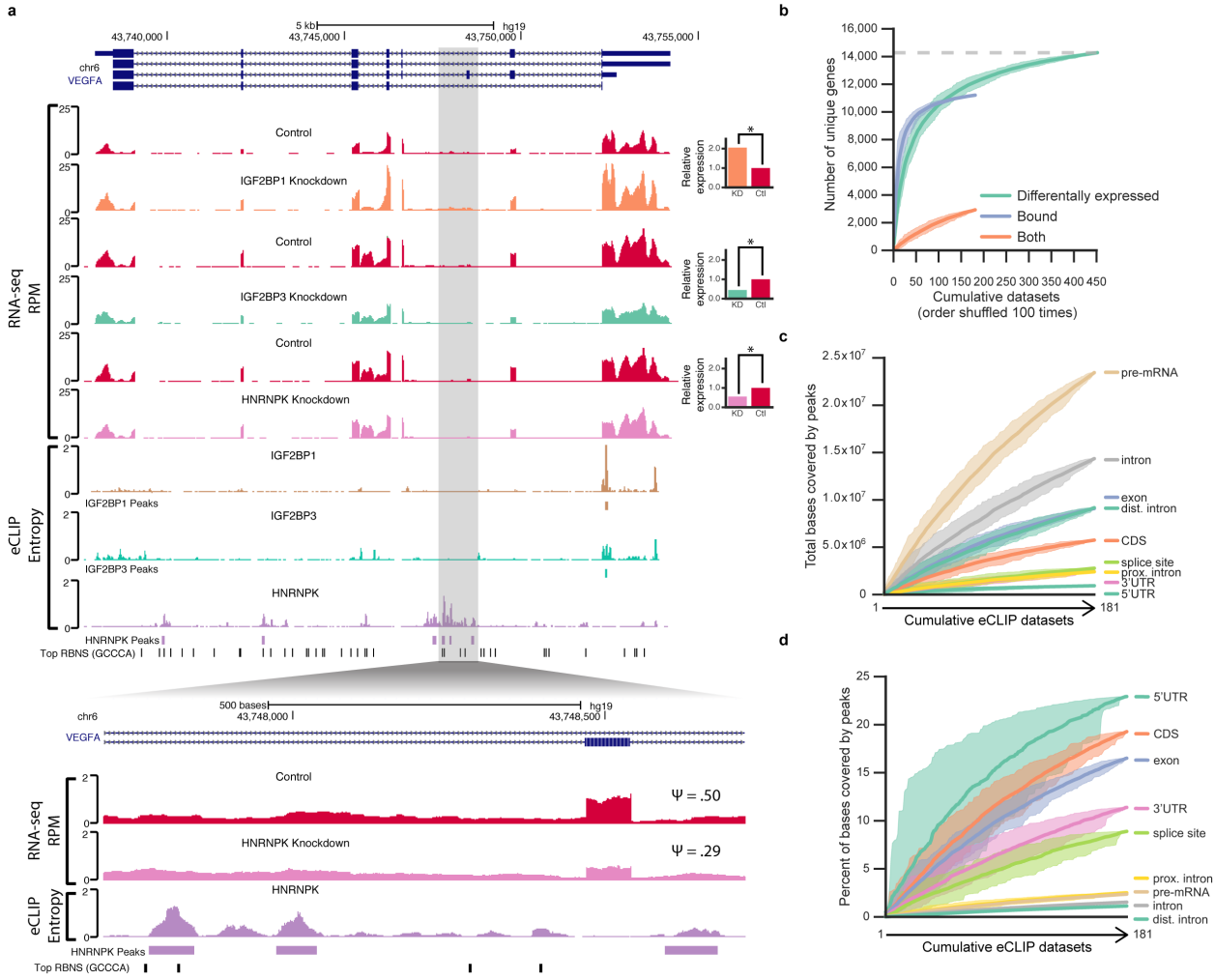

Figure 3-2: **Integrative analysis of RBP binding and function**
(**a**) Combinatorial expression and splicing regulation of VEGFA. Tracks indicate RNA-seq read density (as reads per million, RPM) and eCLIP relative information (in IP versus input). Asterisk indicates significant expression changes ($|\log_2(\text{fold-change})| > 1$ and p-value < 0.05), and percent spliced in (Ψ) is indicated.
(**b**) Lines indicate the number of genes that are (green) differentially expressed upon RBP knockdown in RNA-seq (FDR < 0.05, p-value < 0.05, and $|\log_2(\text{fold-change})| > 1$), (blue) bound in eCLIP, or (orange) both bound and differentially expressed in the same cell type. The ordering of each data type was randomly shuffled 100 times, with the shaded region indicating the tenth and ninetieth percentiles. Grey line indicates the number of genes expressed with TPM > 1 in either K562 or HepG2.
(**c-d**) Lines indicate the mean of (**c**) the cumulative total number of bases covered by peaks, or (**d**) the cumulative fraction of bases covered by peaks for 100 random orderings of the 181 eCLIP datasets, separated by transcript regions as indicated. Shaded region indicates tenth through ninetieth percentiles.

The scale of data available enabled us to query the degree to which we have saturated the discovery of RBP binding sites on RNA and RBP-associated RNA processing events. In total, 14,281 genes were differentially expressed in at least one knockdown experiment (**Fig.** 3-2**b**), including 72.8% of genes expressed in both cell types and 71.2% of those expressed in at least one of the two (**Fig.** 3-S2**a-b**). Similarly, 11,211 genes were bound in at least one eCLIP dataset, representing 86.1% of genes expressed in both cell types and 75.6% of those expressed in at least one (**Fig.** 3-2**b**, **Fig.** 3-S2**a-b**). However, only 2,827 genes were both bound by and responsive to knockdown of the same RBP, suggesting that a large fraction of knockdown-responsive expression changes result from indirect effects and that many binding events do not directly affect gene expression levels in the conditions assayed here (**Fig.** 3-2**b**, **Fig.** 3-S2**a-b**). Similar analysis of alternative splicing changes revealed that differentially spliced events were saturated to a lesser degree than differentially expressed genes. (Greater variability in the cumulative number of unique events resulted from the large number of splicing changes in one knockdown dataset, the RNA helicase and spliceosomal protein AQR (Zhang et al. [2017]) (**Fig.** 3-S2**c**).)

Considering RBP binding, we observed a total of 23.4 Mb of annotated pre-mRNA transcripts covered by at least one reproducible RBP binding site, representing 9.1 Mb of exonic and 13.4 Mb of intronic sequence (**Fig.** 3-2**c**). This total represents only 1.5% of annotated intronic sequence (1.1% of distal intronic, 2.5% of proximal intronic, and 8.9% of splice site), but 16.5% of annotated exonic sequences (22.9% of 5' UTR, 19.3% of CDS, and 11.4% of 3' UTR, respectively) were covered by at least one peak (**Fig.** 3-2**d**). Restricting our analysis to only genes expressed (TPM$\geq$1) in both cell types resulted in substantial increases in these percentages (**Fig.** 3-S2**d**). We found that profiling a new RBP often resulted in greater increases in covered bases of the transcriptome than did re-profiling the same RBP in another cell line, with the marginal 181st dataset averaging a 0.38% (for newly profiled RBPs) or 0.32% (for RBPs profiled in a second cell type) increase (**Fig.** 3-S2**e-g**). We additionally observed an average 1.4% increase in covered bases when we added eCLIP datasets from H1 and H9 embryonic stem cells for RBFOX2 and IGF2BP1-3 (all of which were also profiled in either K562 or HepG2), suggesting that substantial numbers of additional RBP binding sites remain to be detected in cell types distinct from K562 and HepG2 (**Fig.** 3-S2**g**).

Although these results correspond well with previous work suggesting that RNAs are often densely coated by RBPs (Silverman et al. [2014]), it remains to be seen what fraction of these peaks mark alternative regulatory interactions versus constitutive RNA processing. Indeed, many may mark relative association of proteins which coat or broadly interact with RNAs as part of their basic function (such as association of RNA Polymerase II component POLR2G with pre-mRNAs, or spliceosomal component association with splice sites).

3.3.2 *In vivo* binding is largely determined by *in vitro* binding specificity

Binding of an RBP to RNA *in vivo* is determined by the combination of the protein's intrinsic RNA binding specificity and other influences such as RNA structure and protein cofactors. To compare the binding specificities of RBPs *in vitro* and *in vivo*, we calculated the enrichment or R value of each 5mer in RBNS-bound sequences relative to input sequences and compared it to the corresponding enrichment of the 5mer in eCLIP peaks relative to randomized locations in the same genes (R_{eCLIP}). Significantly enriched 5mers *in vitro* and *in vivo* were mostly in agreement, with 17 of the 26 RBPs having significant overlap in the 5mers that comprise their motif logos (**Fig.** 3-3**a**, left). The top RBNS 5mer for an RBP was almost always enriched in eCLIP peaks (**Fig.** 3-3**a**, center). In most cases, similar degrees of enrichment and similar motif logos were observed in eCLIP peaks located in coding, intronic or UTR regions, suggesting that RBPs have similar binding determinants in each of these transcript regions (**Fig.** 3-3**a**, center; **Fig.** 3-S3**a**). Strikingly, the single most enriched RBNS 5mer occurred in 30% or more of the peaks for several RBPs including SRSF9, TRA2A, RBFOX2, PTBP3, TIA1, and HNRNPC. For most RBPs, at least half of eCLIP peaks contained one of the top five RBNS 5mers. Therefore, instances of these 5mers provide candidate nucleotide-resolution binding locations for the RBP (**Fig.** 3-3**a**, right). Such precise binding locations have a number of applications, e.g., they can be intersected with databases of genetic variants to identify those likely to alter function at the RNA level. When two or more distinct motifs were enriched in both RBNS and eCLIP, the most enriched motif *in vitro* was usually also the most enriched *in vivo* (5 out of 7 cases). These

observations are consistent with the idea that intrinsic binding specificity observed *in vitro* explains a substantial portion of *in vivo* binding preferences for most RBPs studied to date.

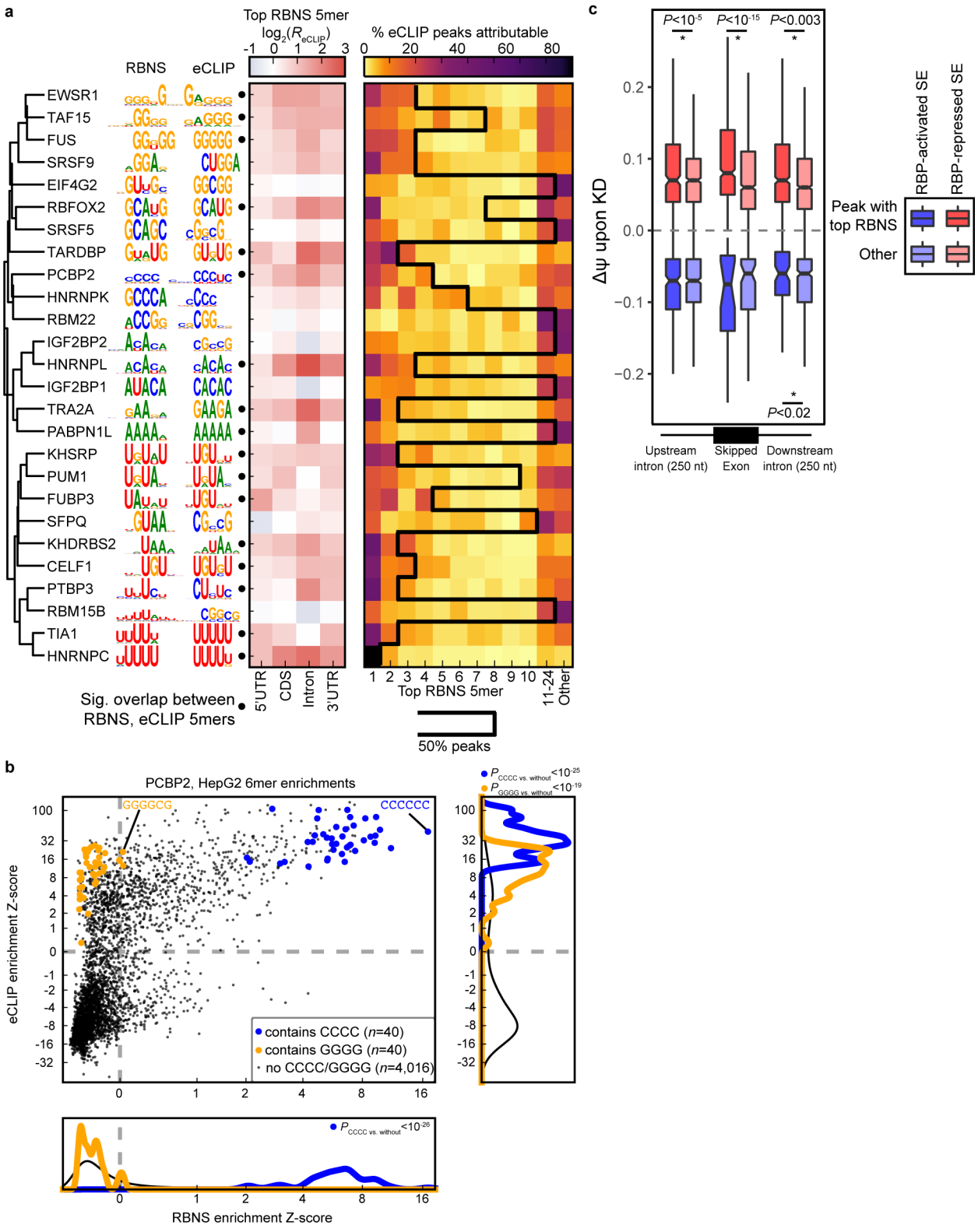

Figure 3-3: **Sequence-specific binding** *in vivo* **is determined predominantly by intrinsic RNA affinity of RBPs**

Figure 3-3

(**a**) Left: Top sequence motif of RBNS versus eCLIP-derived enriched 5mers clustered by similarity of RBNS motifs. Filled circles to the right of the eCLIP logo indicate if the groups of 5mers comprising the RBNS and eCLIP motifs overlap significantly (hypergeometric $p < 0.05$). Center: Enrichment of the top RBNS 5mer in eCLIP peaks (R_{eCLIP}) within different genomic regions. Right: The proportion of eCLIP peaks attributed to each of the 10 highest affinity RBNS 5mers, as well as the #11-24 RBNS 5mers combined. The black line indicates the number of top RBNS 5mers required to explain >50% of eCLIP peaks for each RBP (maximum, 24 5mers).

(**b**) Comparison of PCBP2 *in vivo* vs. *in vitro* 6mer enrichments, with 6mers containing CCCC and GGGG highlighted. Significance was determined by Wilcoxon rank-sum test and indicated if $p < 0.05$. x- and y-axes are plotted on an arcsinh scale.

(**c**) Comparison of the magnitude of splicing change upon RBP knockdown for SEs containing eCLIP peaks with vs. without the top RBNS 5mer, separated by the direction of SE change upon KD and location of the eCLIP peak relative to the SE. Significance was determined by Wilcoxon rank-sum test and indicated if $p < 0.05$.

For a minority of RBPs (12/26), the top five RBNS 5mers explained less than half of eCLIP peaks. Some of these factors appear to have affinities to RNA structural features or to more extended RNA sequence elements not well represented by 5mers (Dominguez et al. [2017]), while the sequence-specific binding of others may be driven to a large extent by cofactors. In some cases, RBNS revealed affinity to a subset of the motifs that were enriched in eCLIP peaks. For example, C-rich 6mers were most enriched in RBNS data for PCBP2 and also in PCBP2 eCLIP peaks (**Fig.** 3-3**b**). In this example, and in several others, a subset of similar eCLIP-enriched *k*mers were not enriched at all by RBNS (e.g., the G-rich 6mers in **Fig.** 3-3**b**). Such 'eCLIP-only' motifs, which were often G-, GC-, or GU-rich (**Fig.** 3-S3**b**), may represent RNA binding of cofactors in complex with the targeted RBP (e.g., G-rich motifs enriched near RBFOX2 peaks may represent sites bound by HNRNPF, HNRNPH and HNRNPM in complex with RBFOX2, Mauger et al. [2008], Damianov et al. [2016]) or could represent biases in crosslinking or in genomic sequences near eCLIP peaks (Sugimoto et al. [2012], Hauer et al. [2015]).

The extent to which strength and mode of binding correlates with eCLIP read density and regulatory activity is not well understood. We focused on regulation of splicing because a large proportion of the available cell type/RBP combinations that included KD, eCLIP, and RBNS data involved RBPs with known roles in splicing, and splicing changes could be readily detected in the KD/RNA-seq data. For most datasets involving KD of known splicing RBPs (18/28), eCLIP binding to one or more specific regions near alternative exons was associated with increased splicing changes upon KD of the RBP. In contrast, this association was observed for only one of the seven datasets involving RBPs that lacked known splicing functions (Fisher exact $p<0.05$, **Fig.** 3-S4**a**). To explore the relationship between sequence-specific binding and regulation, we classified eCLIP peaks as RBNS+ or RBNS- depending on whether they contained the highest-affinity RBNS motif (Methods). We then asked whether these classes of peaks differed in their association with splicing regulation. Examining exon-proximal regions commonly associated with splicing regulation, we found that RBNS+ eCLIP peaks conferred stronger regulation of exon skipping - with an average ~25% increase in $|\Delta\Psi|$ than did RBNS- peaks (**Fig.** 3-3**c**). Thus, sequence-specific binding appears to confer stronger regulation than non-sequence-specific binding, and RBNS motifs

can be used to distinguish a subset of eCLIP peaks that has greater regulatory activity. The *in vitro* data were needed to make this distinction, because a similar analysis of eCLIP peaks classified by presence/absence of the top eCLIP-only 5mer exhibited minimal differences in splicing regulatory activity (**Fig.** 3-S4**b**). In general, RNA binding directed by intrinsic RNA affinity may involve longer-duration interactions that more consistently impact recruitment of splicing machinery.

3.3.3 Functional Characterization of the RBP Map

Analysis of the knockdown/RNA-seq data enables inference of the function of some RNA elements identified by eCLIP and RBNS. Regulation of RNA stability, which alters steady-state mRNA levels, can be observed by an increase or decrease in mRNA expression upon knockdown of an RBP. We compared differentially expressed genes upon RBP knockdown with eCLIP binding to three regions of mRNAs: 5' UTR, CDS, and 3' UTR. We observed that binding of 17 RBPs correlated with increased expression upon knockdown, including RBPs with previously identified roles in induction of RNA decay (such as UPF1, XRN2, and DDX6) (**Fig.** 3-4**a**, **Fig.** 3-S5**a**) as well as previously unknown candidates including METAP2, a methionyl aminopeptidase which has been co-purified with poly(A)-selected RNA but has no previously known RNA processing roles (Castello et al. [2012]). Although METAP2 was highly bound throughout the transcriptome, CDS regions were on average 3.4-fold enriched in METAP2 eCLIP, well above the 2.4-fold average enrichment of 3' UTR and 1.4-fold depletion of intronic regions (**Fig.** 3-S5**b-c**). We further observed a trend in which increasing METAP2 eCLIP fold-enrichment correlated with progressively stronger increases in RNA expression upon knockdown, supporting an RNA regulatory role (**Fig.** 3-4**b**).

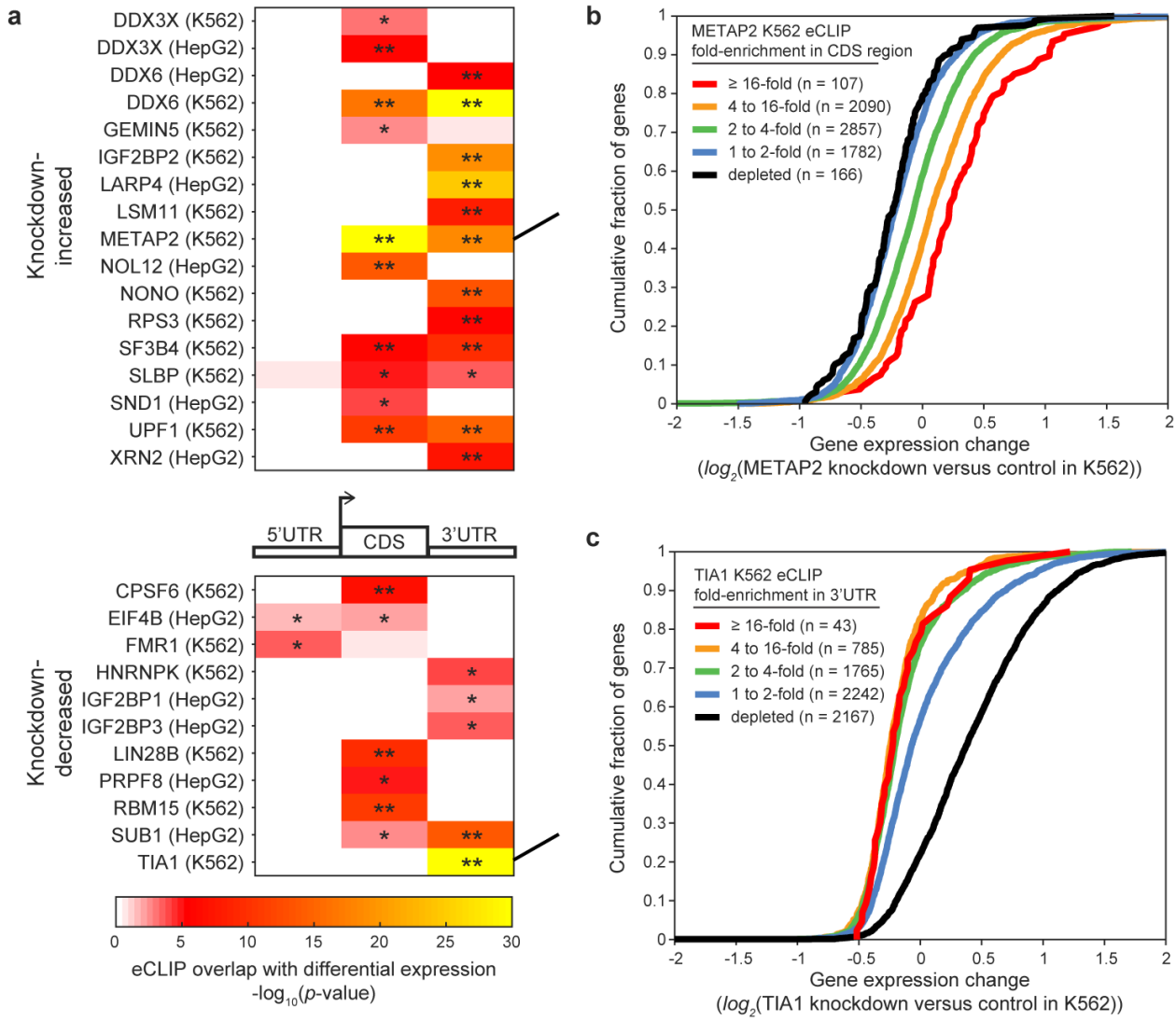

Figure 3-4: **Association between RBP binding and RNA expression upon knockdown**

(**a**) Heatmap indicates significance of overlap between genes significantly bound ($p < 10^{-5}$ and $>$ 4-fold enriched in eCLIP versus input) and genes significantly (top) increased or (bottom) decreased ($p < 0.05$ and FDR < 0.05) in RBP knockdown RNA-seq experiments. Significance was determined by Fisher's Exact test or Yates' Chi-Square approximation where appropriate; * indicates $p < 0.05$ and ** indicates $p < 10^{-5}$ after Bonferroni correction. Shown are all overlaps meeting a $p < 0.05$ threshold.

(**b-c**) Lines indicate cumulative distribution plots of gene expression fold-change (knockdown versus control) for indicated categories of eCLIP binding of (**b**) METAP2 in K562 and (**c**) TIA1 in K562.

Additionally, we observed 11 RBPs for which binding correlated with decreased RNA levels following knockdown (**Fig.** 3-4**a**), including the stress granule component TIA1. Surprisingly, although our transcriptome-wide analysis indicated that transcripts with 3' UTR binding of TIA1 decreased upon knockdown in K562 cells (suggesting a globally stabilizing role for TIA1) (**Fig.** 3-4**c**), little to no stabilization activity was observed for 3' UTR-bound mRNAs in HepG2 cells (**Fig.** 3-S5**d**). Using TIA1 RBNS motif content in 3' UTRs rather than eCLIP binding sites, we additionally observed cell type-specific enrichment of TIA1 motifs in destabilized transcripts upon KD in K562, with no significant effect (though a slight motif enrichment in stabilized genes upon KD) in HepG2 (**Fig.** 3-S5**e-g**). This distinction is reminiscent of previous studies, which indicate that TIA1 can either induce RNA decay when tethered to a 3' UTR (Yamasaki et al. [2007]), or stabilize target mRNA levels through competition with other RBPs including HuR (Wigington et al. [2015]). Thus, our results provide further evidence that TIA1 can regulate mRNA stability through dynamic cell type-specific interactions.

3.3.4 RBP association with splicing regulation

Next, we considered how localized RBP binding was associated with splicing regulation. To do this, we generated a 'splicing map' for each RBP (Ule et al. [2006]), which depicts the average RBP binding at relative positions in the regulated exon and flanking introns of exons that show RBP-responsive splicing changes, relative to binding at non-responsive alternative exons (**Fig.** 3-5**a**, **Fig.** 3-S6). Considering 57 RBPs with eCLIP and knockdown/RNA-seq performed in the same cell line, we observed a wide variety of RNA maps (**Fig.** 3-5**b**). Binding of SR proteins typically correlated with decreased inclusion upon knockdown and hnRNP protein binding correlated with increased inclusion upon knockdown, consistent with classical models of antagonistic effects of SR and hnRNP proteins on splicing (Erkelenz et al. [2013]) (**Fig.** 3-S7**a-b**).

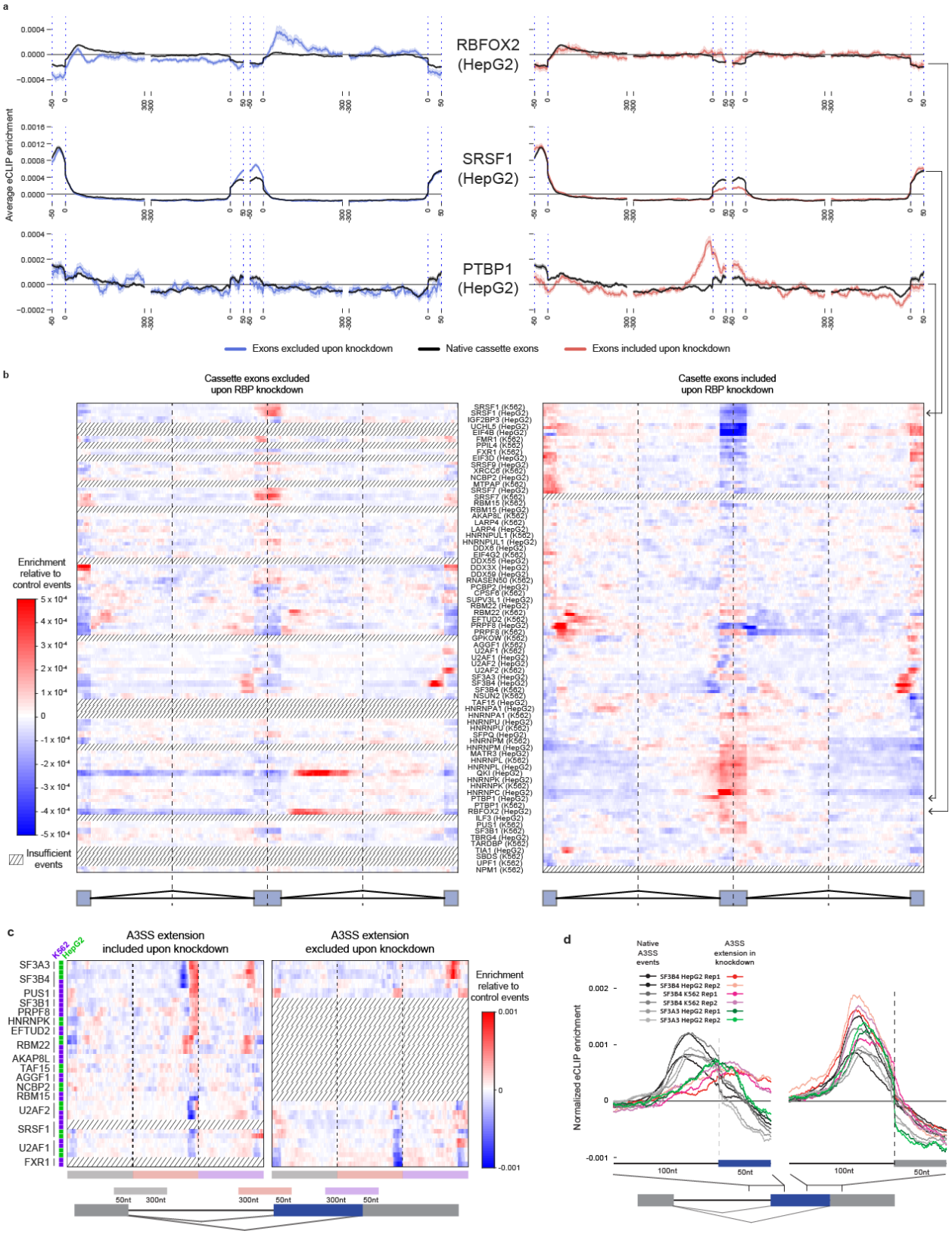

Figure 3-5: **Integration of eCLIP and RNA-seq identifies splicing regulatory patterns**

Figure 3-5

(**a**) Normalized splicing maps of RBFOX2, SRSF1, and PTBP1 for cassette exons excluded (left) and included (right) upon knockdown, relative to a set of 'native' cassette exons with $0.05 < \Psi < 0.95$ in controls. Lines indicate average eCLIP read density in IP versus input for indicated exon categories, with standard error shown by the shaded area. The displayed region shown extends 50 nt into exons and 300 nt into introns.
(**b**) Heatmap indicates the difference between normalized eCLIP read density at cassette exons excluded (left) or included (right) upon RBP knockdown versus native cassette exons. Shown are all RBPs with eCLIP and knockdown RNA-seq data, with dashed lines indicating datasets with fewer than 100 significantly altered events.
(**c**) As in (**b**), shown for alternative 3' splice site events. Dashed lines indicate datasets with fewer than 50 significantly altered events. The displayed regions include the upstream common 5' splice site (grey box), the extended alternative 3' splice site (orange box), and the distal alternative 3' splice site (purple box).
(**d**) Lines indicate mean normalized eCLIP enrichment in IP versus input for SF3B4 and SF3A3 at (red/purple/green) alternative 3' splice site extensions in RBP knockdown or (black) alternative 3' splice site events in control HepG2 or K562 cells.

Surprisingly, in addition to canonical alternative splicing regulators such as RBFOX2 and PTBP1, we observed significant differential association of spliceosomal components near knockdown-sensitive alternative splice sites (**Fig.** 3-5**b**, **Fig.** 3-S7**c-d**). For example, we observed that comparing cassette exons to constitutive exons revealed strikingly different binding patterns, with cassette exons characterized by increased association of 5' splice site machinery such as PRPF8, EFTUD2, and RBM22 at the upstream 5' splice site but depletion at the cassette exon 5' splice site (**Fig.** 3-S7**c**). Branch point recognition factors such as SF3B4 and SF3A3 showed similar depletion for the cassette exon branch point and enrichment at the downstream branch point, leading to a distinctive pattern of spliceosomal association at the cassette exon (**Fig.** 3-S7**c**). Moreover, when considering non-spliceosomal RBPs we similarly observed that while RBP association was higher at cassette-bordering proximal intron regions relative to constitutive exons, the upstream 5' splice site and proximal intron showed an even greater enrichment (**Fig.** 3-S7**d**). Together, these observations for non-spliceosomal RBPs suggest that the upstream 5' splice site of alternative exons represent a greater source of regulatory RBP binding than previously believed.

Splicing maps were also constructed for alternative 5' (A5SS) and alternative 3' splice site (A3SS) events (**Fig.** 3-5**c**). Again, we noticed differential association of spliceosomal components (**Fig.** 3-5**c**, **Fig.** 3-S8**a**). Focusing on A3SS events, we noted a particularly prominent pattern of association for branch point factors SF3B4 and SF3A3, which typically bind to the branch point region ∼50 nt upstream of the 3' splice site. For both, we observed that enrichment was shifted downstream towards the 3' splice site for the set of A3SS events where the distal (downstream) 3' splice site is preferred upon knockdown (**Fig.** 3-5**c-d**, **Fig.** 3-S8**b**). These genome-level results generalize previous mini-gene studies showing that 3' splice site scanning and recognition originates from the branch point and can be blocked if the branch point is moved close to the 3' splice site AG (Bradley et al. [2012]), and suggest that regulated branch point recognition plays a key role in A3SS regulation by restricting recognition by the 3' splice site machinery (Smith et al. [1993]) (**Fig.** 3-S8**c**).

In summary, the RBPs we have surveyed that participate in alternative splicing display a wide diversity of regulatory modes. Moreover, although the target splicing events differ, the splicing map of a given RBP is highly consistent between cell types (**Fig.** 3-5**b-c**, **Fig.** 3-

S7**a-b**, **Fig.** 3-S8**a**). Thus, performing eCLIP and KD/RNA-seq in a single cell type may be sufficient to elucidate the splicing map for an RBP, but multiple cell types must be surveyed to identify the full repertoire of direct regulatory events.

3.3.5 RBP Association with Chromatin

Increasing evidence suggests that regulatory RNAs, both coding and non-coding, are broadly involved in gene expression through their interaction with chromatin (Engreitz et al. [2016], Skalska et al. [2017]). Based on the expectation that these regulatory events must enlist specific RBPs participating in co-transcriptional RNA processing, we surveyed 58 RBPs in HepG2 and 45 RBPs in K562 cells for their association with chromatin by ChIP-seq (Supplementary Table 1). We selected RBPs for analysis based on their complete or partial localization in the nucleus and on the availability of antibodies that could efficiently immunoprecipitate each factor. These RBPs belong to a wide range of functional categories, including SR and hnRNP proteins, spliceosomal components, and RBPs that have been generally considered to function as transcription factors, such as POLR2G and GTF2F1. Interestingly, 30 of 58 RBPs (52%) in HepG2 cells and 29 of 45 RBPs (64%) in K562 cells showed specific interactions with chromatin, with several hundred to more than ten thousand specific binding peaks identified for each RBP.

We next characterized the RBP-chromatin interactions with respect to established chromatin activities, including DNase I hypersensitive sites and various histone marks (**Fig.** 3-6**a**). This analysis revealed a general preference of RBPs for euchromatic relative to heterochromatic regions, especially gene promoters, although individual RBPs showed distinct preferences. Collectively, even this moderately sized set of RBPs occupied ∼30% of all DNase hypersensitive or open chromatin regions and ∼70% of annotated gene promoters, which was similar between the two cell types, suggesting broad involvement of RBPs in chromatin activities in the human genome. These data provide a foundation for future investigation of direct roles of specific RBPs in transcriptional control, exemplified by recent studies on SR proteins (Ji et al. [2013]) and RBFOX2 (Wei et al. [2016]).

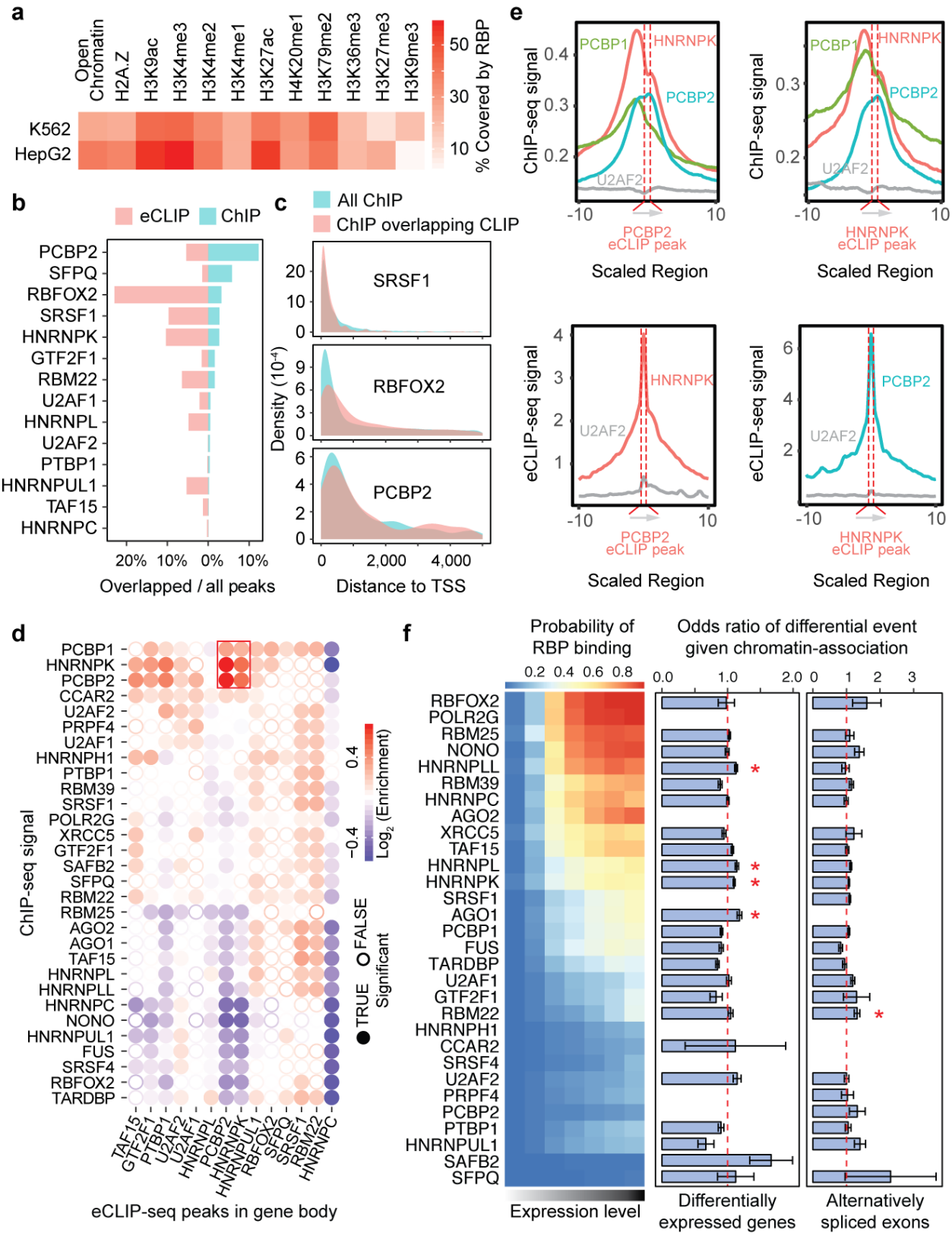

Figure 3-6: **Chromatin-association of RBPs and overlap with RNA binding**

Figure 3-6

(**a**) Collective RBP binding activities on specific chromatin regions, including DNase I hypersensitive sites and various histone marks in HepG2 and K562 cells.
(**b**) Percentage of RBP eCLIP peaks overlapped with corresponding ChIP-seq peaks (pink) or percentage of RBP ChIP-seq peaks overlapped with corresponding eCLIP peaks (green). RBPs are sorted by decreasing level of overlapped ChIP-seq peaks.
(**c**) Distributions of overall chromatin binding activities (green) and overlapped chromatin and RNA binding activities (pink) of representative RBPs.
(**d**) Clustering of overlapped chromatin and RNA binding activities of different RBPs in non-promoter regions. Color key indicates the degree of ChIP enrichment at eCLIP peaks relative to surrounding regions with significance as True ($p < 0.001$) or False ($p > 0.001$) indicating the significance of the enrichment.
(**e**) Cross-RBP comparison of chromatin and RNA binding activities. Top: ChIP-seq density of indicated RBPs around PCBP2 or HNRNPK eCLIP peaks. Bottom: eCLIP density of indicated RBPs around PCBP2 or HNRNPK eCLIP peaks.

It is well established that a variety of RNA processing events are co-transcriptionally coupled (Naftelberg et al. [2015]). It is therefore possible that some RBP-chromatin association events are coupled with their direct RNA binding activities in cells. To explore this possibility, we intersected ChIP-seq peaks with eCLIP peaks for individual RBPs in HepG2 cells, revealing specific RBPs with relatively high degrees of overlap in these two types of interactions (**Fig.** 3-6**b**). Interestingly, the distribution of the overlapped regions varies among individual RBPs. For example, the chromatin and RNA binding activities of SR proteins, as exemplified by SRSF1, are predominantly coincident near gene promoters, while most other RBPs show such overlapping activities further into the gene body (**Fig.** 3-6**c**). These data revealed coordinated actions of RBPs at the chromatin and RNA levels, suggesting that a fraction of their RNA binding events are chromatin-associated. We next sought to quantify the similarity of each RBP pair's chromatin and RNA binding activities by computing the overlap in ChIP-seq and eCLIP-seq signal in non-promoter regions (**Fig.** 3-6**d**). Clustering of the data indicated that many RBPs might function in conjunction with other RBP(s) to coordinate their chromatin and RNA binding activities (red signals in **Fig.** 3-6**d**). Particularly interesting is the high correlation between poly(rC) binding proteins HNRNPK and PCBP1/2, which share a common evolutionary history and domain composition yet perform diverse functions (Makeyev and Liebhaber [2002]) (red box in **Fig.** 3-6**d**). To illustrate the relationship between RNA and chromatin interactions, we plotted the ChIP-seq density of these three RBPs relative to PCBP2 and HNRNPK eCLIP peaks in non-promoter regions. We found that chromatin binding signals were typically centered around eCLIP peaks, although HNRNPK and PCBP1 (to a lesser degree) were slightly biased for chromatin binding upstream of RNA binding, indicative of a specific topological arrangement of these potential RBP complexes on chromatin in a manner dependent on the direction of transcription (**Fig.** 3-6**e**, top panels). We next asked whether the signal overlapping eCLIP peaks was specific to the DNA binding activity of these RBPs or if it also held for related RNA binding activities. For this purpose, we plotted the eCLIP density of multiple RBPs relative to another RBP's eCLIP peaks. This analysis revealed a significant degree of overlap among the RNA binding activities of HNRNPK and PCBP2, which contrasts to a randomly selected RBP for comparison (**Fig.** 3-6**e**, bottom panels). Combined, these data strongly reinforce coordinated

chromatin and RNA binding activities of specific RBPs.

To investigate whether there is correlation between gene expression and RBP association, we clustered the probability of each RBP's association with genes binned by increasing expression levels, revealing a pervasive positive correlation between gene expression and RBP association for the majority of RBPs, with the exception of SAFB2 and SFPQ which have only a few binding sites (**Fig.** 3-6**f**, left panels). These data implicated regulatory roles of RBPs in gene expression through chromatin association. To pursue this point, we compared the frequency that a gene is differentially expressed upon RBP knockdown depending on whether or not the RBP was chromatin-associated at that gene, controlling for the different expression levels of these two groups. We found that chromatin association of HNRNPLL, HNRNPL, HNRNPK, and AGO1 correlated with a significantly increased chance of gene expression change upon knockdown (**Fig.** 3-6**f**, right panels). To explore the connection between alternative splicing regulation and RBP-chromatin association, we calculated similar ratios for each RBP and found that RBM22 chromatin association was associated with a significant increase in alternative splicing changes upon knockdown (**Fig.** 3-6**f**, right panels). Together, these data suggest that chromatin-association of RBPs affects RNA processing and gene expression.

3.3.6 RBP regulatory features in subcellular space

The systematic imaging screen revealed that RBPs display a broad diversity of localization patterns (**Fig.** 3-7**a**), with most factors exhibiting targeting to multiple structures in the nucleus and cytoplasm (**Fig.** 3-7**b**). Next, we integrated RBP localization data with other datasets generated here. First, we considered the nuclear relative to cytoplasmic ratios for each RBP. As expected, we observed a significant shift towards increased binding to unspliced transcripts for nuclear RBPs, whereas cytoplasmic RBPs were enriched for binding to spliced transcripts (**Fig.** 3-S9**a-b**). Next, we identified a collection of 80 RBPs that exhibit robust localization to SC35 (SRSF2)-labeled nuclear speckles, a class of subnuclear structures enriched for proteins involved in pre-mRNA splicing (Spector and Lamond [2011]). Consistent with the established function of speckles in splicing, analysis of splicing changes associated with RBP depletion revealed that speckle-localized RBPs impact larger numbers

of splicing events compared to non-speckle proteins (**Fig.** 3-7**c**). Notably, the top 12 (and 19 of the top 20) RBP knockdowns with the strongest impact on splicing (in terms of number of altered splicing events) corresponded to speckle-localized RBPs, including spliceosomal components U2AF1, U2AF2, SF3B4, and SF3A1 as well as splicing regulators HNRNPK and SRSF1 (**Fig.** 3-7**c**). By contrast, nucleolar RBPs had significantly less impact on pre-mRNA splicing than non-nucleolar RBPs, as expected (**Fig.** 3-7**c**). Furthermore, when we queried the enrichment for RNA binding modalities in eCLIP data, we observed similarly striking enrichment for speckle-localized RBPs binding to proximal introns as well as snRNAs (notably RNU2 and RNU6) (**Fig.** 3-S9**c**), again consistent with nuclear speckles playing a key role in pre-mRNA splicing.

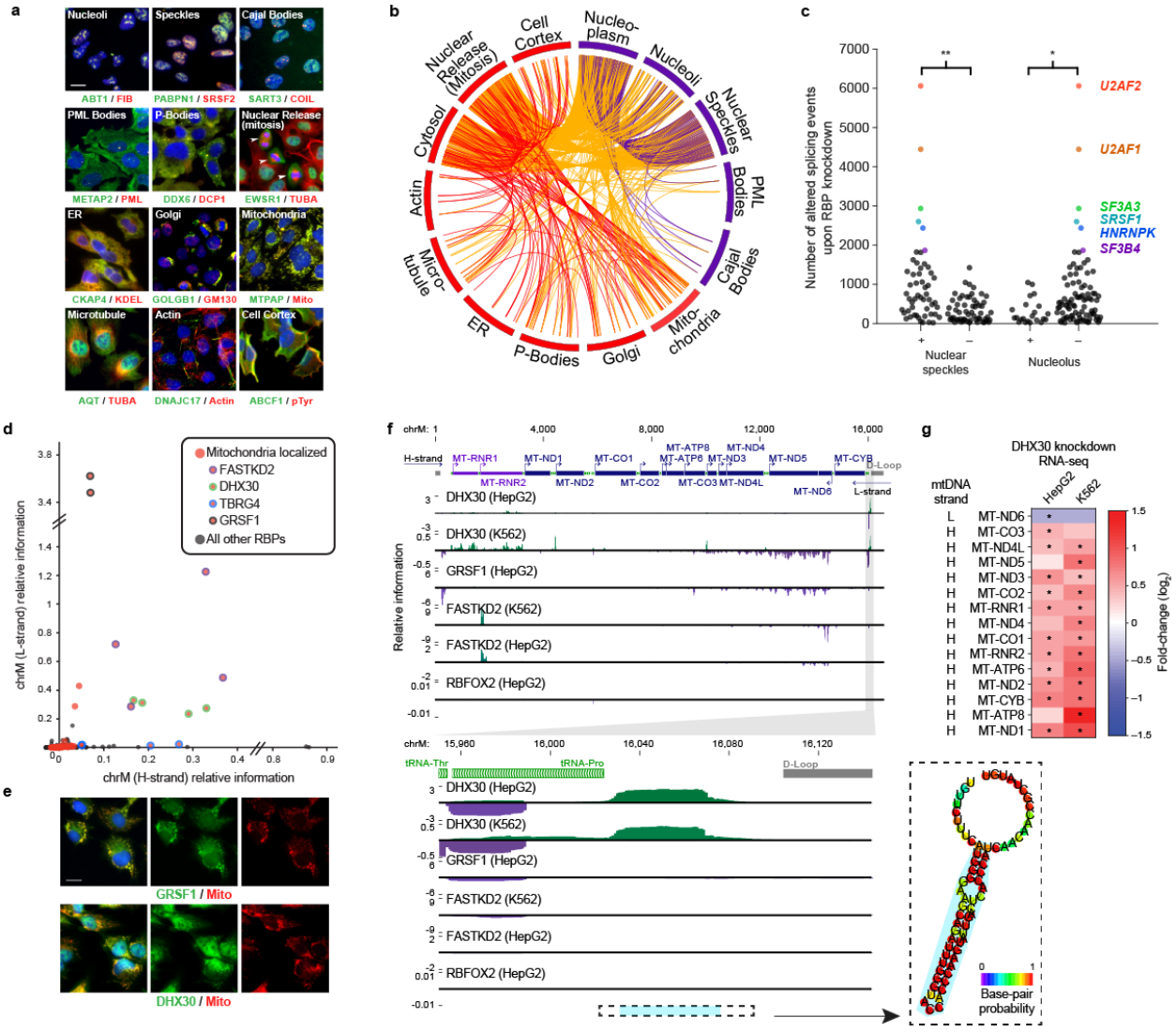

Figure 3-7: **RBP subcellular localization, binding, and regulation**

Figure 3-7

(**a**) Example RBPs (green) co-localized with twelve interrogated markers (red).
(**b**) Circos plot with lines indicating co-observed localization patterns (red: within cytoplasm; purple: within nucleus; orange: between cytoplasm/nucleus).
(**c**) Points indicate the number of differential splicing events observed upon knockdown of each RBP, separated by the presence or absence of localization in nuclear speckles (center) or the nucleolus (right). * indicates $p < 0.05$ and ** indicates $p < 10^{-4}$ by Wilcoxon rank-sum test.
(**d**) Points indicate eCLIP relative information content (IP versus input) for mitochondria H-strand (x-axis) or L-strand (y-axis) for RBPs with mitochondrial localization by IF (red, with specific RBPs noted) and all other RBPs (black).
(**e**) IF images of mitochondrial localization of GRSF and DHX30.
(**f**) Genome browser tracks indicate binding (as eCLIP relative information content) along (top) the mitochondrial genome or (bottom) a ∼300 nt region for indicated RBPs. (right) Inset shows RNA secondary structure prediction (RNAfold) for the indicated region in blue.
(**g**) Heatmap indicates gene expression change upon DHX30 knockdown for all mitochondrial protein-coding and rRNA transcripts. * indicates significant expression changes ($p < 0.05$ and FDR< 0.05 from DEseq2 analysis).

Focusing on localization to specific cytoplasmic organelles, we noted that 42 RBPs exhibited localization to mitochondria, an organelle with unique transcriptional and RNA processing regulation (Rackham et al. [2002]). We observed numerous RBPs with significant binding to the mitochondrial genome in either Heavy (H)-strand (TBRG4), Light (L)-strand (GRSF1), or both (FASTKD2, DHX30) (**Fig.** 3-7**d-e**). Validating these observations, mitochondrial localization by immunofluorescence was associated with significantly increased eCLIP binding to mitochondrial RNA (**Fig.** 3-S9**d**). DHX30 is essential for proper mitochondrial ribosome assembly and oxidative phosphorylation (Antonicka and Shoubridge [2015]). Intriguingly, in addition to widespread association with many mitochondrial genes that was consistent with previous RIP-seq findings (Antonicka and Shoubridge [2015]) (**Fig.** 3-S9**e**), we observed dramatic enrichment at an unannotated region on the mitochondrial H-strand downstream of all annotated genes and just upstream of the replication D loop, which has strong potential to form a stem-loop structure (**Fig.** 3-7**f**). We further observed that DHX30 knockdown resulted in increased expression of nearly all H-strand transcripts, but decreased expression of L-strand transcript ND6 (**Fig.** 3-7**g**). Identification of the termination signal for mitochondrial H-strand transcription has remained elusive; it is tempting to speculate this site of DHX30 association could mark such signal. These examples illustrate how intracellular localization of RBPs can be used synergistically with binding and loss-of-function studies to infer aspects of post-transcriptional regulation that occur in different cellular compartments and organelles.

3.3.7 Preservation of RBP regulation across cell types

Next we evaluated whether RBP regulation is preserved across cell types. Analyzing the 55 RBPs profiled by eCLIP in both K562 and HepG2 cell lines, we found that only a small portion (averaging 11.0%) of peaks were found in genes with cell type-specific expression (TPM $\geq$ 1 in one cell type and TPM $<$ 0.1 in the other) (**Fig.** 3-8**a**, **Fig.** 3-S10**a**). An average of 70.6% of peaks were located within genes expressed in both cell types (TPM $\geq$ 1 in both), although only 10.3% were present in genes that were unchanged between K562 and HepG2 (defined as fold-difference $\leq$ 1.2), and 31.6% of peaks were within genes that changed by at least two-fold (**Fig.** 3-8**a**, **Fig.** 3-S10**a**). To illustrate, we observed that

RBFOX2 eCLIP peaks at least 8-fold enriched in HepG2 were typically also enriched in K562 (average enrichment of 6.1-fold) if the bound gene was expressed within five-fold of the level in HepG2 cells (Covering the 'Unchanged', 'Weakly differential', and 'Moderately differential' classes in **Fig.** 3-8**a**). In contrast, 89.8% of HepG2 peaks in genes with cell type-specific expression were not enriched in K562 (**Fig.** 3-8**a**). Indeed, 49.7% of RBFOX2 HepG2 peaks that were not enriched in K562 occurred in genes with cell type-specific expression whereas only 5.5% occurred in genes within a two-fold change.

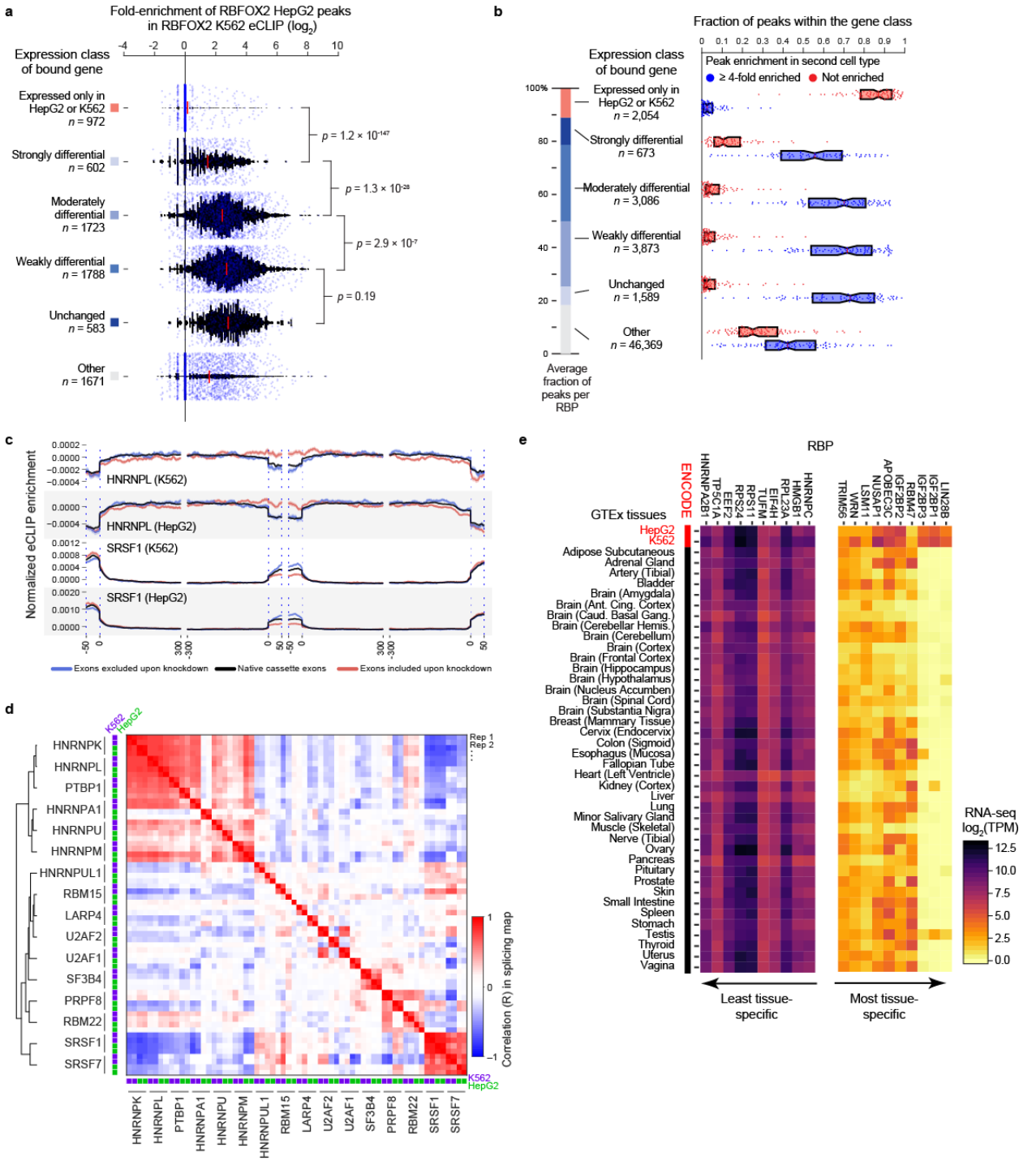

Figure 3-8: **Preservation of RBP binding and regulation across cell types**

Figure 3-8

(**a**) Each point indicates the fold-enrichment in K562 eCLIP of RBFOX2 for a reproducible RBFOX2 eCLIP peak in HepG2, with underlaid black histogram. Peaks are separated based on the relative expression difference of the bound gene between K562 and HepG2: unchanged (fold-difference ≤ 1.2), weakly ($1.2 <$ fold-difference ≤ 2), moderately ($2 <$ fold-difference ≤ 5) or strongly (fold-difference > 5) differential (each of which required expression TPM ≥ 1 in both K562 and HepG2), or cell type-specific genes (TPM < 0.1 in one cell type and TPM ≥ 1 in the other). Mean is indicated by red lines, with significance determined by Kolmogorov-Smirnov test.

(**b**) For each RBP profiled in both K562 and HepG2, points indicate the fraction of peaks in the first cell type associated with a given gene class that are (blue) at least four-fold enriched, or (red) not enriched (fold-enrichment ≤ 1) in the second cell type. Boxes indicate quartiles, with median indicated by red lines. (left) Stacked bar indicates the average fraction of peaks per RBP (see **Fig.** 3-S10**a** for per-RBP distributions).

(**c**) Cassette exon splicing maps for HNRNPL and SRSF1 in both K562 and HepG2 cells.

(**d**) Heatmap indicates correlation (Pearson R) between splicing maps for all RBPs profiled in both K562 and HepG2, hierarchically clustered at the RBP level.

(**e**) Expression of the 10 RBPs with the highest and lowest tissue specificity across the two ENCODE cell lines and 40 human tissues.

Expanding this analysis to all RBPs profiled in both cell types, we observed similar results: an average of 84.3% of peaks in genes with cell type-specific expression were not enriched in the other cell type, whereas 68.0%, 65.5%, and 67.6% of peaks in unchanging, weakly, or moderately differentially expressed genes were enriched by at least 4-fold in the second cell type, respectively (**Fig.** 3-8**b**). Requiring the strict IDR thresholds for peak identification, an average of 44.1% of peaks in genes with similar expression were preserved across cell types (**Fig.** 3-S10**c**). 48.9% of all peaks that showed no enrichment in the second cell type occurred in genes with cell type-specific expression (a 4.4-fold enrichment), whereas only 21.0% occurred in unchanging, weakly, or moderately differentially expressed genes, respectively (a 3-fold depletion) (**Fig.** 3-S10**c**). Thus, these results suggest that most RBP binding is preserved across cell types for similarly expressed genes, and that much of differential eCLIP signal between HepG2 and K562 likely reflects underlying gene expression differences more than differential binding.

Next, we asked whether an RBP's positional pattern of splicing regulation tended to be conserved across cell types. Considering splicing maps of cassette exons, we observed that binding of many RBPs had highly similar correlations with either inclusion or exclusion upon knockdown, including alternative splicing regulators HNRNPL and SRSF1 (**Fig.** 3-8**c-d**). We observed that the splicing maps for the same RBP across cell types had significantly higher correlation than random pairings of RBPs, when comparing across all 16 RBPs with both eCLIP and RNA-seq datasets (with sufficient splicing changes) in both K562 and HepG2 (**Fig.** 3-8**d**, **Fig.** 3-S10**d**).

Cell type-specific regulation of RNAs may also be achieved through differential modulation of RBP levels. To assess which RBPs confer regulation in such a manner, we calculated the expression of each RBP in HepG2 and K562 cells in addition to 40 diverse human tissues from the GTEx Project (**Fig.** 3-S11). Many RBPs had high expression in ENCODE cell lines and across a broad range of human tissues, including ribosomal proteins (RPL23A, RPS11, RPS24), translation factors (EIF4H, EEF2), and ubiquitously expressed splicing factors (HNRNPC, HNRNPA2B1) among the 10 least tissue-specific RBPs (**Fig.** 3-8**e**). However, several other RBPs had highly tissue-specific expression exhibiting either a pattern of high expression in one or a small number of human tissues (e.g., LIN28B, IGF2BP1/3) or were

differentially expressed by orders of magnitude across several human tissues (e.g., IGF2BP2 and APOBEC3C), indicating that the RNA targets and regulatory activity of these RBPs are likely modulated through cell type-specific gene expression programs. Of course, even RBPs with similar mRNA levels across tissues may have different protein levels or activity because of post-translational modifications, which are widespread among RBPs.

3.4 Discussion

Our study represents the largest effort to date to systematically study the roles of human RBPs by integrative approaches. The resulting catalog of functional RNA elements substantially expands the repertoire of regulatory components encoded in the human genome. While the impact of DNA binding proteins mostly culminates in effects on gene expression levels, RBP function encompasses a broader range of activities. RBP functions extend outside the nucleus and into the cytoplasm and organelles, and consist of multiple paths by which RNA substrates are altered (splicing, RNA editing/modification, RNA stability, localization, translation), expanding transcriptome and proteome complexity. We demonstrate the effectiveness of combining *in vivo* maps of RNA binding sites using eCLIP with orthogonal approaches, such as *in vitro* evaluation of RNA affinity for the same RBPs, chromatin association by ChIP-seq, and functional assessment of RNA variation by depletion experiments and RNA-seq. At the molecular level, we confirm that *in vivo* and *in vitro* preferences are highly correlated for RBPs interrogated here, and show that CLIP peaks containing motifs reflective of intrinsic RNA affinity are more predictive of regulation. We confirm using unbiased genome-wide analyses that SR and hnRNP proteins have broadly antagonistic effects on alternative splicing. Moreover, we extend genome-wide previous findings that alternative 3' splice site choice results from an "AG" scanning process initiating with branch point recognition, and implicate the upstream 5' splice sites of cassette exons in splicing regulation. We also implicate RNA structures bound by an RBP in processing of mitochondrial transcripts, and elucidate new RNA splicing maps for many RBPs. Furthermore, our data provide the first systematic investigation of chromatin-associated gene regulation and RNA processing at the level of RBP-nucleic acid interactions. At the cellular level, immunohistochemical analysis with our extensive repository of RBP-specific antibodies place these molecular interactions within subcellular contexts. We confirm localization of many RBPs to nuclear speckles, mitochondria, and other compartments, and identify many new proteins resident to these sites, emphasizing the necessity of localization data for interpreting RBP-RNA regulatory networks.

Here, we have surveyed the *in vivo* binding patterns of 126 RBPs, comprising roughly

10% of the human genes predicted to encode proteins that interact directly with RNA. Our observation that additional mapping of new RBPs leads to a greater increase in the functional RNA element coverage than mapping of the same RBPs in additional cell lines argues that expansion of these approaches to additional RBPs will be particularly informative. Additionally, the binding modes *in vivo* are highly preserved across genes expressed similarly in our two cell lines assayed (K562 and HepG2). Nevertheless, mapping of previously characterized RBPs in drastically different cell types (embryonic stem cells, post-mitotic cells such as neurons and muscle cells) and human tissues, with highly distinct transcriptomes, will undoubtedly yield new discoveries. Additionally, RNA processing is highly dynamically regulated during acute or chronic environmental influences such as stress, as new binding sites may arise from both environmental changes in RBP or RNA concentrations as well as from changes in post-translational modifications, binding partners, or subcellular distribution of RBPs. Thus, studying RBP subcellular localization and RBP-RNA substrate regulation under these conditions has potential to reveal new biology.

We expect that the data generated here will provide a useful framework upon which to build analyses of other aspects of RNA regulation, such as microRNA levels, RNA editing and modifications such as pseudouridylation and m6A methylation, translation efficiency, and mRNA half-life measurements. We have yet to integrate *in vivo* RNA structure probing data to evaluate how RBP-mediated RNA processing are influenced by local (Singh et al. [2007]) and long-range RNA structures (Lovci et al. [2013]). As we continue to embark on comprehensively characterizing all functional RNA elements, genome-scale CRISPR/Cas9-inspired genome-editing (Doudna and Charpentier [2014]) and RNA modulation (Nelles et al. [2016]) technologies will ultimately provide opportunities to study the impact on cellular and organismal phenotypes resulting from disruption of these RNA elements.

3.5 Supplementary Figures

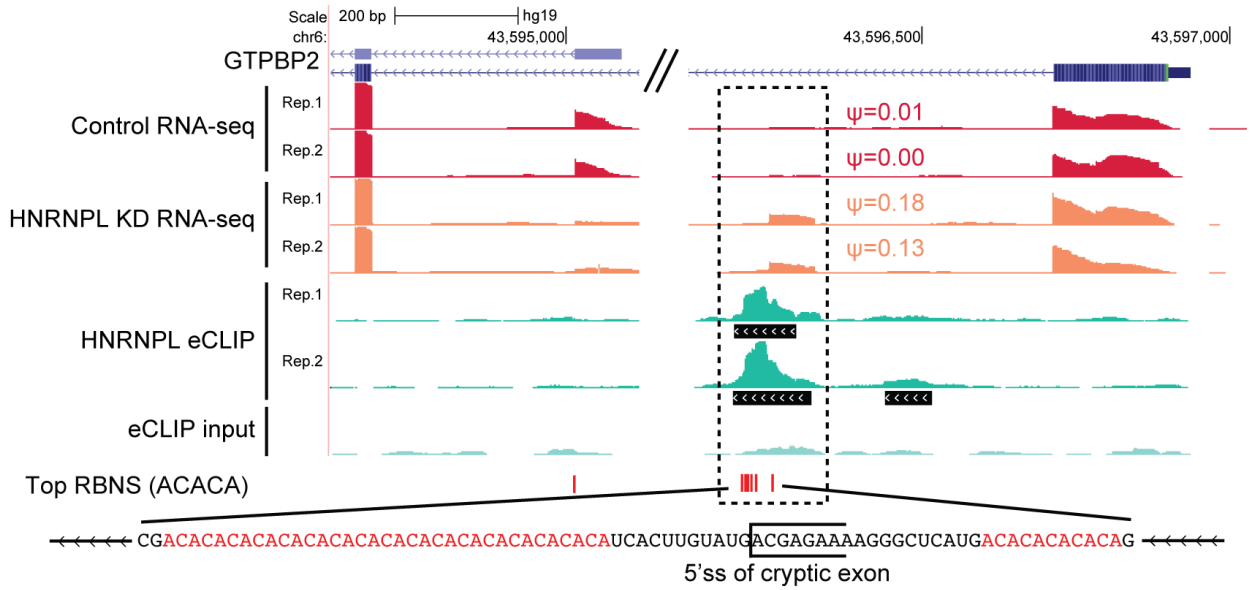

Figure 3-S1: **Integrative analysis of RBP data types in cryptic exon suppression** Shown is RNA-seq read density (reads per million), eCLIP read density (reads per million), and RBNS motif presence proximal to a 73 nt cryptic exon expected to induce nonsense-mediated decay (NMD) of GTP Binding Protein 2, a ribosome rescue factor whose loss induces neurodegeneration in certain genetic backgrounds (Ishimura et al. [2014]). eCLIP indicates that HNRNPL binds over the cryptic exon 5' splice site in a sequence-specific manner to a region rich in the top RBNS 5mer, ACACA.

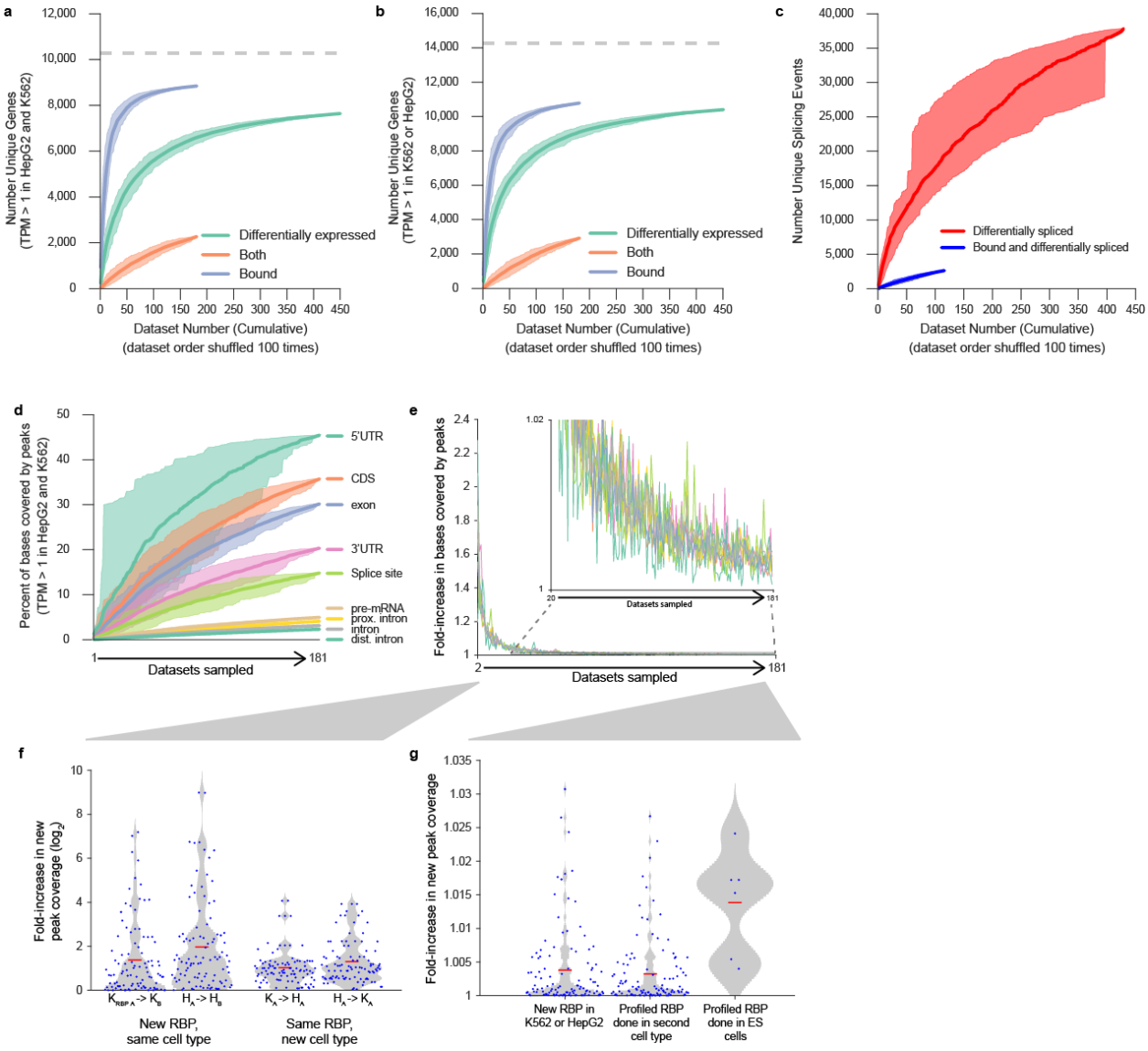

Figure 3-S2: **Saturation of RBP binding and regulation in the transcriptome**

Figure 3-S2

(**a**-**b**) Lines indicate the mean of 100 random orderings of each data type for the number of genes that are (green) differentially expressed upon RBP knockdown and RNA-seq (requiring FDR < 0.05, p-value < 0.05, and |fold-change| > 1), (blue) bound in eCLIP (overlapped by a IDR-reproducible peak with $p < 10^{-3}$ and fold-enrichment ≥ 8 in IP versus input), or (orange) both bound and differentially expressed in the same cell type. Grey dotted line indicates the total number of genes considered, either (**a**) with TPM > 1 in both HepG2 and K562, or (**b**) TPM > 1 in either K562 or HepG2. Shaded regions indicate tenth to ninetieth percentile.

(**c**) Lines indicate the mean of 100 random orderings of datasets for the number of (red) differential splicing events upon RBP knockdown (including cassette exons, alternative 5' and 3' splice sites, retained introns, and mutually exclusive exons; requiring FDR < 0.05, p-value < 0.05, and $|\Delta\Psi| > 0.05$), and (blue) exons both bound by an RBP and differentially spliced upon RBP knockdown in the same cell type (with binding defined as a peak located anywhere between the upstream intron 5' splice site and downstream intron 3' splice site). Shaded regions indicate tenth to ninetieth percentile.

(**d**) Lines indicate the mean cumulative fraction of bases covered by peaks for 100 random orderings of the 181 eCLIP datasets, separated by transcript regions as indicated, with shaded region indicating tenth and ninetieth percentiles, only considering genes expressed (with TPM > 1) in both K562 and HepG2.

(**e**) Data and colors as in (**d**), represented as fold-increase in mean bases covered by peaks from n to $n + 1$ eCLIP datasets.

(**f**). Points indicate the fold-increase in bases covered by peaks between sampling one or two datasets, separated by whether the second is the same RBP in a new cell type ($\mathrm{K}_A \rightarrow \mathrm{H}_A$ or $\mathrm{H}_A \rightarrow \mathrm{K}_A$ for RBP A profiled in K562 and then HepG2 or HepG2 and then K562, respectively) or a different RBP in the same cell type ($\mathrm{K}_A \rightarrow \mathrm{K}_B$ or $\mathrm{H}_A \rightarrow \mathrm{H}_B$ for RBP A followed by RBP B in either K562 or HepG2, respectively), with kernel smoothed density indicated by the shaded area. Red line indicates mean.

(**g**) Points indicate the fold-increase in bases covered by peaks between sampling all versus leaving one dataset out, separated by whether the RBP is (left) a newly profiled RBP or (center) a previously profiled RBP profiled in a second cell type (of either K562 or HepG2). (right) The fold-increase observed if an independent eCLIP experiment performed in H1 or H9 human embryonic stem cells is added (including RBFOX2, IGF2BP3, and two replicates each for IGF2BP1 and IGF2BP2). Red line indicates mean.

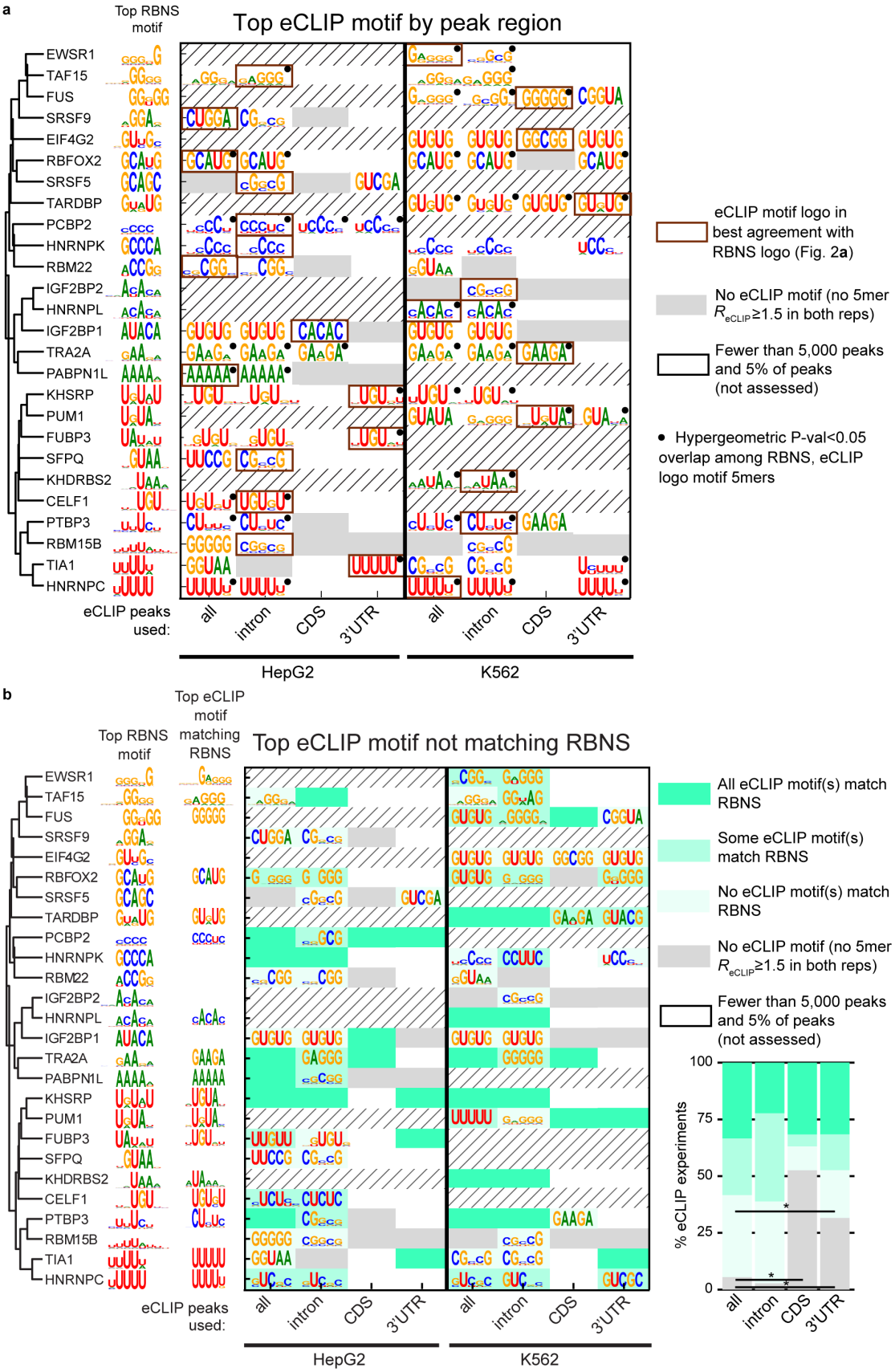

Figure 3-S3: **Comparison of *in vitro* RBNS-derived motifs with *in vivo* eCLIP-derived motifs**

Figure 3-S3

(**a**) Top motif derived from all eCLIP peaks as well as eCLIP peaks within intronic, CDS, and 3' UTR regions. Motifs were only derived if there were at least 5,000 peaks or 5% of total peaks in that region, averaged over the two eCLIP replicates. Dashed lines indicate eCLIP was not performed in that cell line. Filled circles indicate significant overlap ($p < 0.05$ by hypergeometric test) between RBNS and eCLIP motifs.
(**b**) The top eCLIP motif that does not match RBNS for the corresponding RBP (if any). The eCLIP motif was considered as matching RBNS if any of its constituent 5mers were among the RBNS Z≥3 5mers (always using at least 10 RBNS 5mers if there were fewer with Z≥3). Dashed lines indicate eCLIP was not performed in that cell line. (right) The percentage of eCLIP experiments aggregated over all RBP/cell types in each category of agreement with RBNS. Horizontal line indicates a significant difference in the proportion of a particular eCLIP/RBNS agreement category between eCLIP analysis of all peaks versus eCLIP analysis of intron, CDS, or 3' UTR peaks ($p < 0.05$ by Fisher's Exact Test).

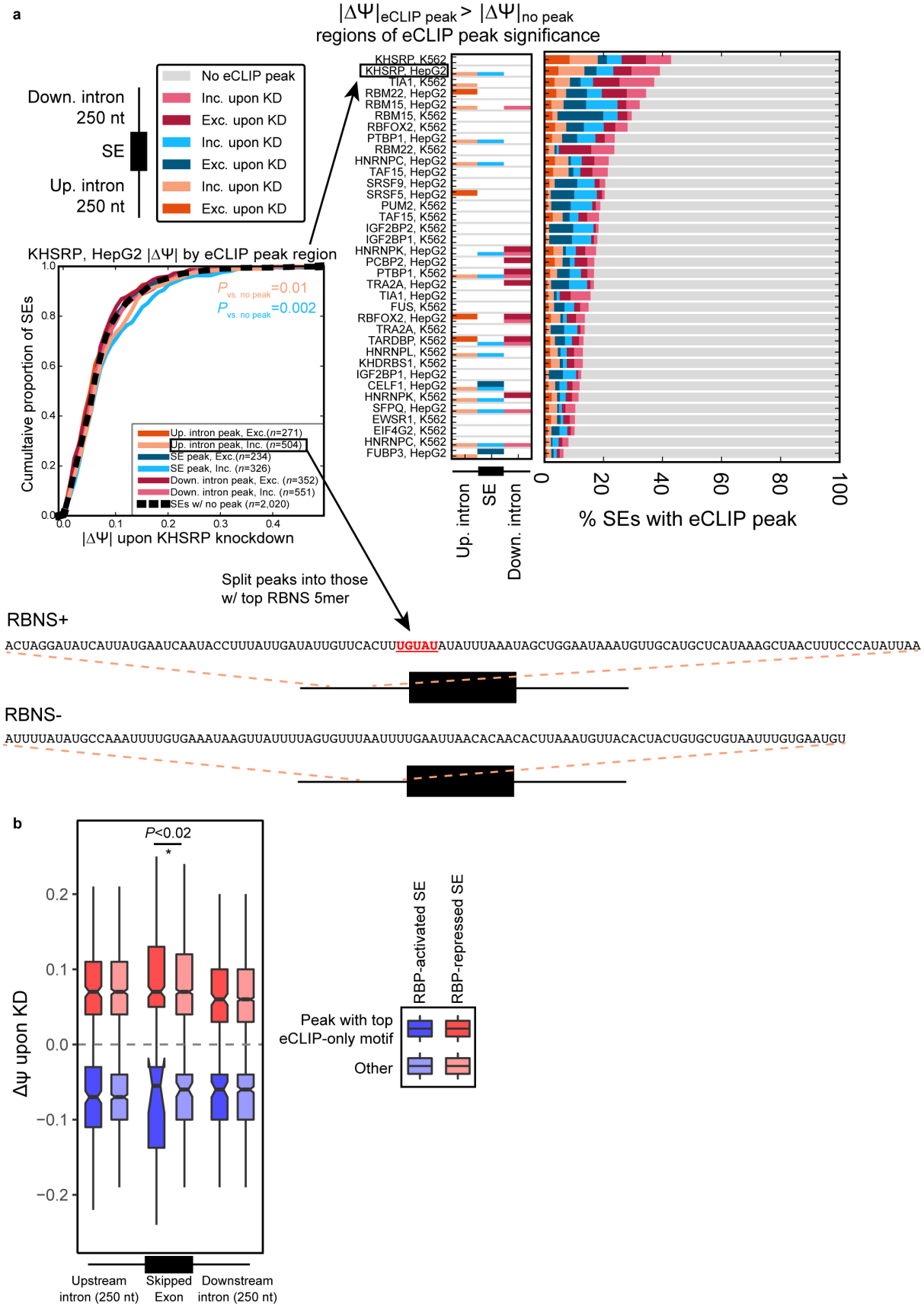

Figure 3-S4: **Splicing regulatory activity of RBNS+ and RBNS- eCLIP peaks**

Figure 3-S4

(**a**) Left: The distribution of $|\Delta\Psi|$ changes upon KD in each of the 6 eCLIP+ peak region/SE splicing change types compared to that of eCLIP- SEs for KHSRP in HepG2 (significant if $p < 0.05$ by Wilcoxon rank-sum test). Center: Regions of significance for eCLIP+ vs. eCLIP- SEs for each eCLIP experiment. Right: Proportion of SEs in each of the six eCLIP+ types for each eCLIP experiment. Bottom: Classification of eCLIP+ peaks into RBNS+ and RBNS- based on the presence of the top RBNS 5mer, shown here for two of the KHSRP peaks in HepG2.

(**b**) Same set of RBPs and corresponding eCLIP+ peak region/SE splicing change types as used in **Fig.** 3-3**c**, but separating eCLIP peaks on whether they contain the top 'eCLIP-only' 5mer (based on the motifs from **Fig.** 3-S3**b**) instead of the top RBNS 5mer.

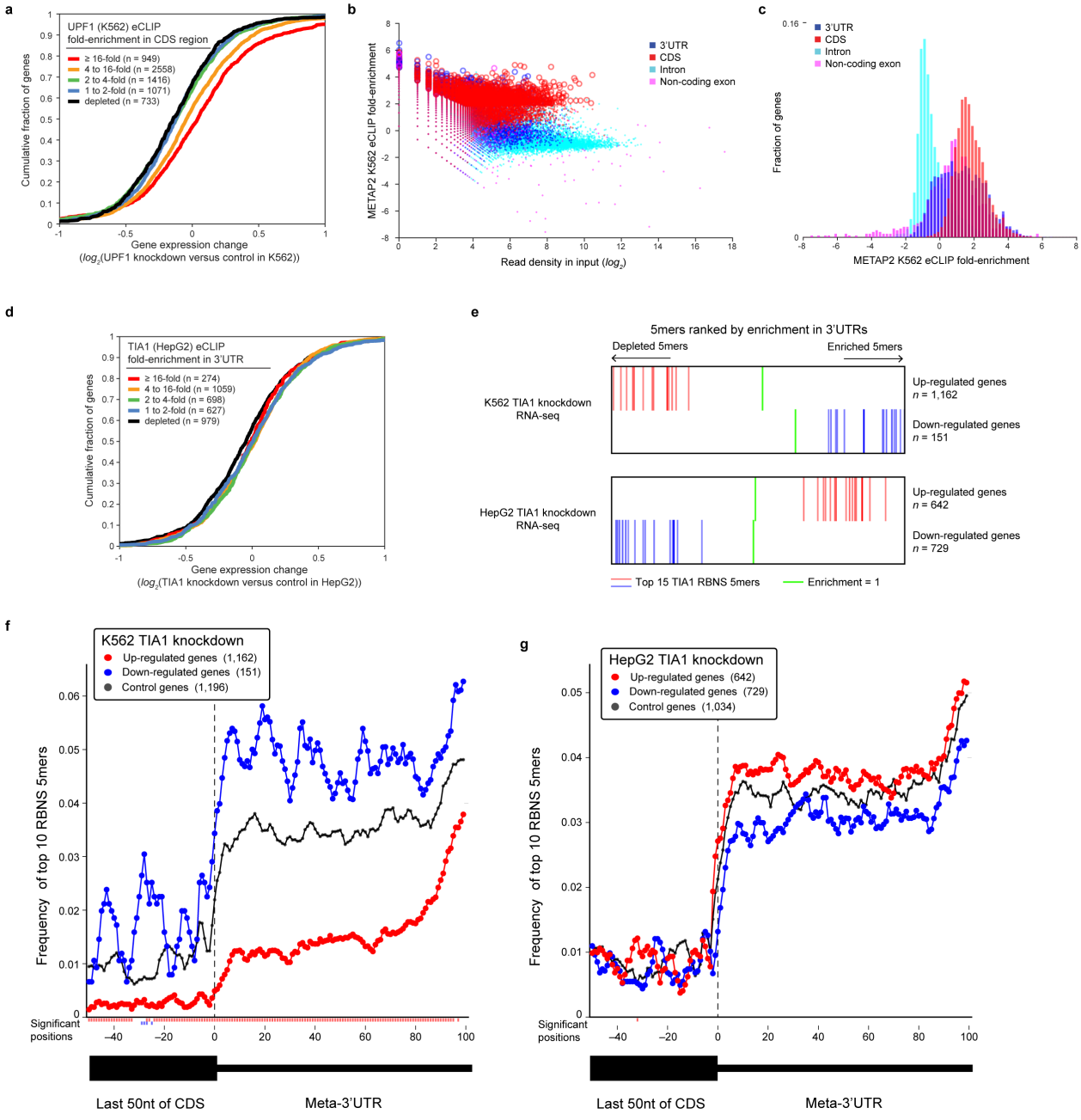

Figure 3-S5: **Association between RBP binding and RNA expression upon knockdown**

Figure 3-S5

(**a**) Lines indicate cumulative distribution plots of gene expression fold-change (UPF1 knockdown versus control) for indicated categories of UPF1 eCLIP binding in K562 cells.
(**b-c**) METAP2 K562 eCLIP region-level binding at 3' UTR, CDS, intronic, and non-coding exonic regions. (**b**) Points indicate read density in input (x-axis) versus fold-enrichment in METAP2 eCLIP (y-axis) for indicated transcript regions of all GENCODE v19 genes. Significantly enriched regions ($p < 10^{-5}$ and fold-enrichment > 4) are indicated by open circles.
(**c**) Histogram of METAP2 eCLIP fold-enrichment for the indicated transcript regions.
(**d**) Cumulative distribution plots of gene expression fold-change (TIA1 HepG2 knockdown versus control) for indicated categories of 3' UTR TIA eCLIP binding.
(**e**) Enrichment or depletion of the top 15 TIA1 RBNS 5mers in 3' UTRs of genes that are up- and down-regulated upon TIA1 knockdown in K562 and HepG2, relative to their frequency in control gene 3' UTRs (green lines indicate an enrichment of 1 (equal frequency in regulated gene 3' UTRs and control gene 3' UTRs)). All 1,024 5mers are ordered from lowest to highest enrichment from left to right in each row.
(**f-g**) Position-specific frequency of the top 10 TIA1 RBNS 5mers in the last 50 positions of the CDS and in a meta-3' UTR of (red) up-regulated, (blue) down-regulated, and (black) control genes upon TIA1 knockdown in (**f**) K562 and (**g**) HepG2 cells. Positions of motif density significantly different in up- or down-regulated genes relative to control genes are indicated below the x-axis (calculated using a binomial test comparing the number of regulated genes that do versus do not have one of the top 10 RBNS 5mers at that position versus the frequency observed in control genes).

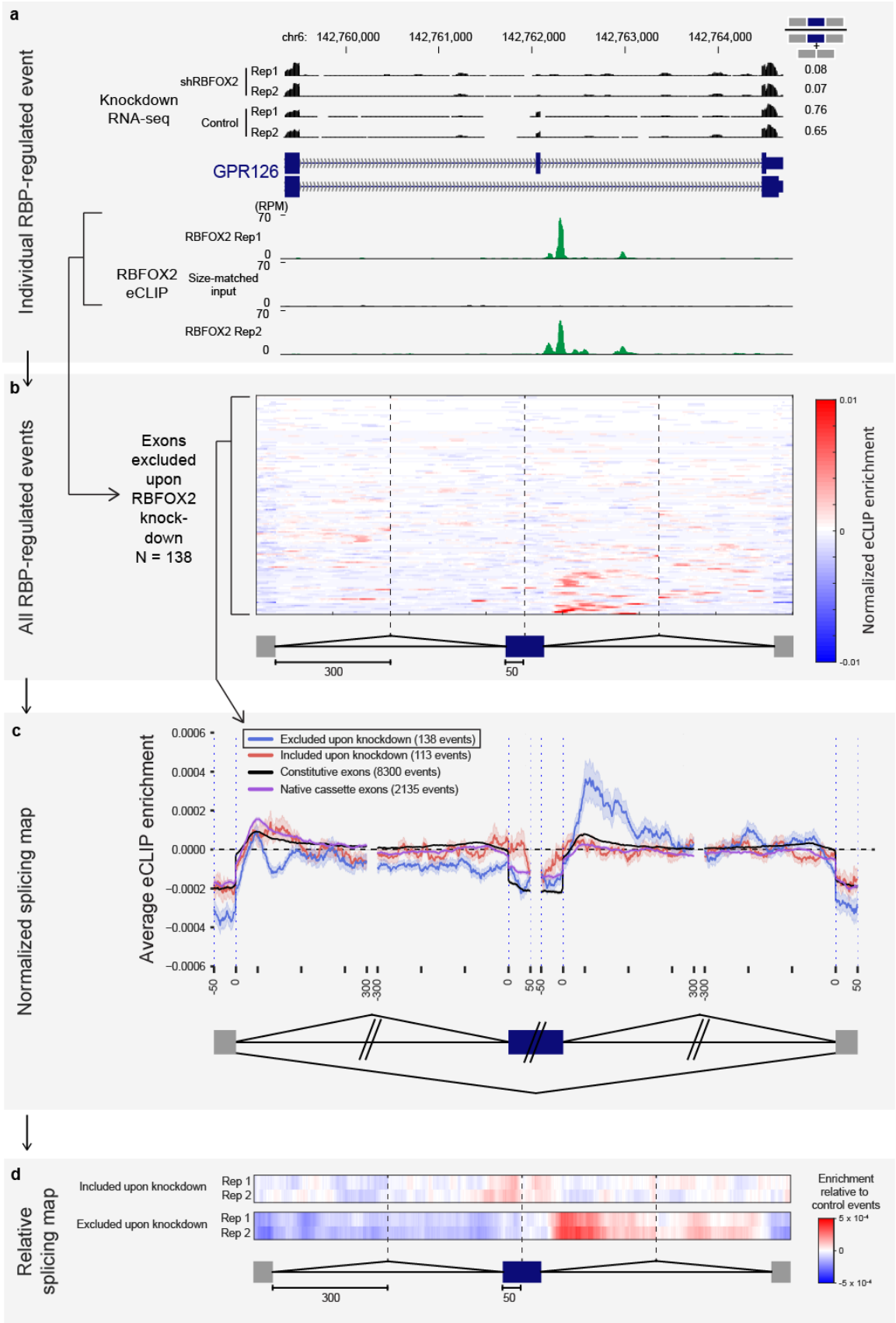

Figure 3-S6: **Generation of splicing maps for RBFOX2**

Figure 3-S6

(**a**) First, individual RBP-regulated splicing events are identified from significant changes in knockdown RNA-seq. Genome browser tracks indicate RNA-seq read density (as reads per million (RPM)) and eCLIP read density (RPM) of RBFOX2 in the same cell type, as well as its paired size-matched input.
(**b**) Next, each exon is normalized between IP versus input to obtain 'Normalized eCLIP enrichment'. The heatmap indicates normalized eCLIP enrichment for all exons significantly excluded upon RBFOX2 knockdown.
(**c**) Next, a 'splicing map' is created by calculating the mean and standard error of the mean of normalized eCLIP enrichment for each position across the region, removing the top and bottom 5% outlier values at each position. Lines in splicing map indicate 'Average eCLIP enrichment', defined as the mean normalized eCLIP enrichment for exons (red) included or (blue) excluded upon RBFOX2 knockdown. Also plotted are (purple) a control set of cassette exons (referred to as 'native' cassette exons) in wildtype HepG2 cells and (black) constitutive exons.
(**d**) A final simplified splicing map vector was calculated by subtracting the normalized eCLIP enrichment of control native cassette exons from that of either included or excluded exons at each position to calculate 'Enrichment relative to control events'.

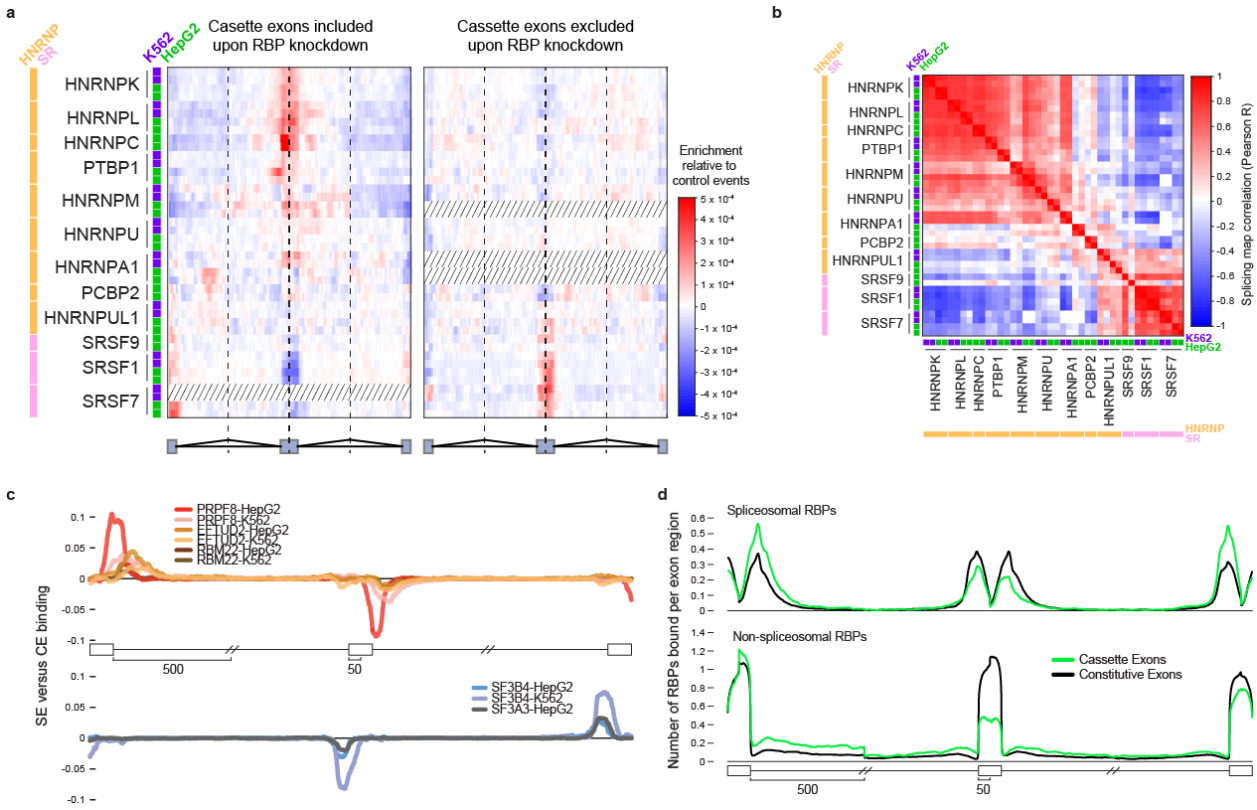

Figure 3-S7: **Splicing regulatory patterns of SR, HNRNP, and spliceosomal proteins**

(**a**) Relative splicing maps for cassette exons included (left) and excluded (right) upon knockdown (as described in **Fig.** 3-5**b**) are shown for all profiled SR and HNRNP proteins. Datasets were hierarchically clustered at the RBP level, and datasets with fewer than 100 events are indicated by slashed lines.

(**b**) Heatmap indicates the Pearson correlation (R) between splicing maps shown in (**a**), calculated across both included and excluded exon maps. Datasets are ordered identically to (**a**).

(**c**) Average binding patterns for indicated spliceosome-associated RBPs in 50 nt exonic and 500 nt intronic regions flanking splice sites. Lines indicate the difference in average binding for native cassette exons minus the average at constitutive exons.

(**d**) Lines indicate the average number of RBPs bound (out of 181 total datasets) in 50 nt exonic and 500 nt intronic regions flanking splice sites, separated by (top) core spliceosomal RBPs or (bottom) all other RBPs.

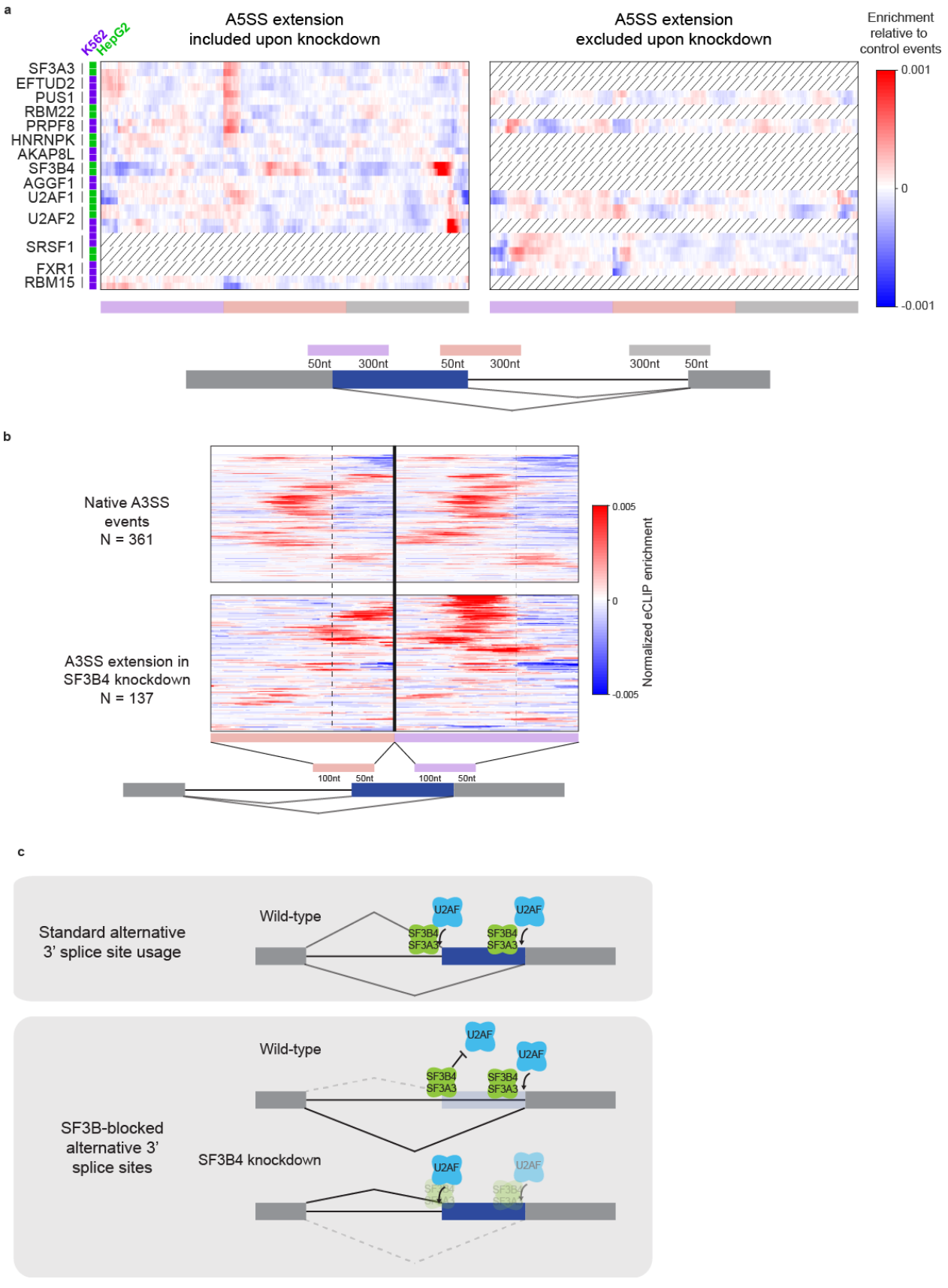

Figure 3-S8: **RNA maps for alternative 5' and 3' splice sites**

Figure 3-S8

(**a**) Heatmap indicates enrichment relative to control events at alternative 5' splice site events, for all RBPs with eCLIP and knockdown RNA-seq data (requiring a minimum of 50 significantly changing events upon knockdown). The region shown extends 50 nt into exons and 300 nt into introns.

(**b**) Heatmap indicates normalized eCLIP signal for SF3B4 in HepG2 at alternative 3' splice site events either (top) alternatively spliced in wildtype cells or (bottom) events with increased usage of the extended 3' splice site upon SF3B4 knockdown. The region shown extends 50 nt into exons and 100 nt into introns.

(**c**) Model for SF3B4 and SF3A3 blockage of 3' splice site recognition by U2AF.

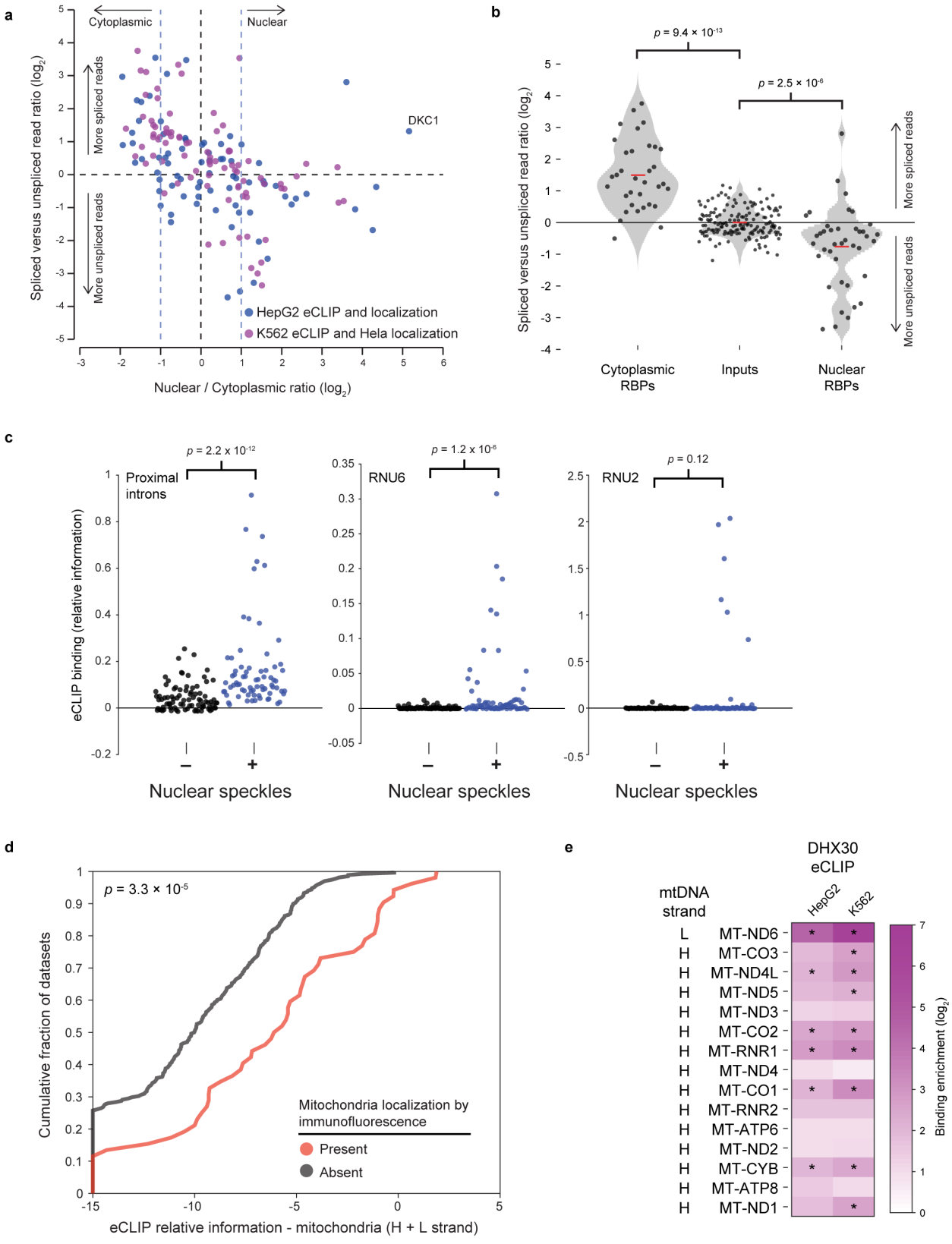

Figure 3-S9: **eCLIP binding patterns in subcellular space**

Figure 3-S9

(**a**) Points indicate nuclear versus cytoplasmic ratio from immunoflourescence (IF) imaging (x-axis) versus ratio of spliced versus unspliced exon junction reads, normalized to paired input (y-axis). RBPs profiled by eCLIP and IF in HepG2 are indicated in blue, and RBPs profiled by eCLIP in K562 (in purple) were paired with IF experiments performed in HeLa cells.

(**b**) Points indicate values as in (**a**), with RBPs separated into nuclear (nuclear / cytoplasmic ratio > 2) and cytoplasmic (nuclear / cytoplasmic ratio < 0.5). Significance was determined by Kolmogorov-Smirnov test, and red line indicates mean.

(**c**) Points indicate eCLIP relative information for all profiled RBPs for the indicated snRNA or pre-mRNA region, separated by the observation of co-localization at nuclear speckles. Significance was determined by Wilcoxon rank-sum test.

(**d**) Cumulative distribution curves indicate total relative information content for the mitochondrial genome for RBPs with mitochondrial localization by IF (red) and all other RBPs (black). Significance was determined by Kolmogorov-Smirnov test.

(**e**) Heatmap indicates DHX30 eCLIP binding across all exons for all mitochondrial protein-coding and rRNA transcripts. * indicates significant eCLIP binding (fold-enrichment > 4 and $p < 0.00001$ in IP versus input).

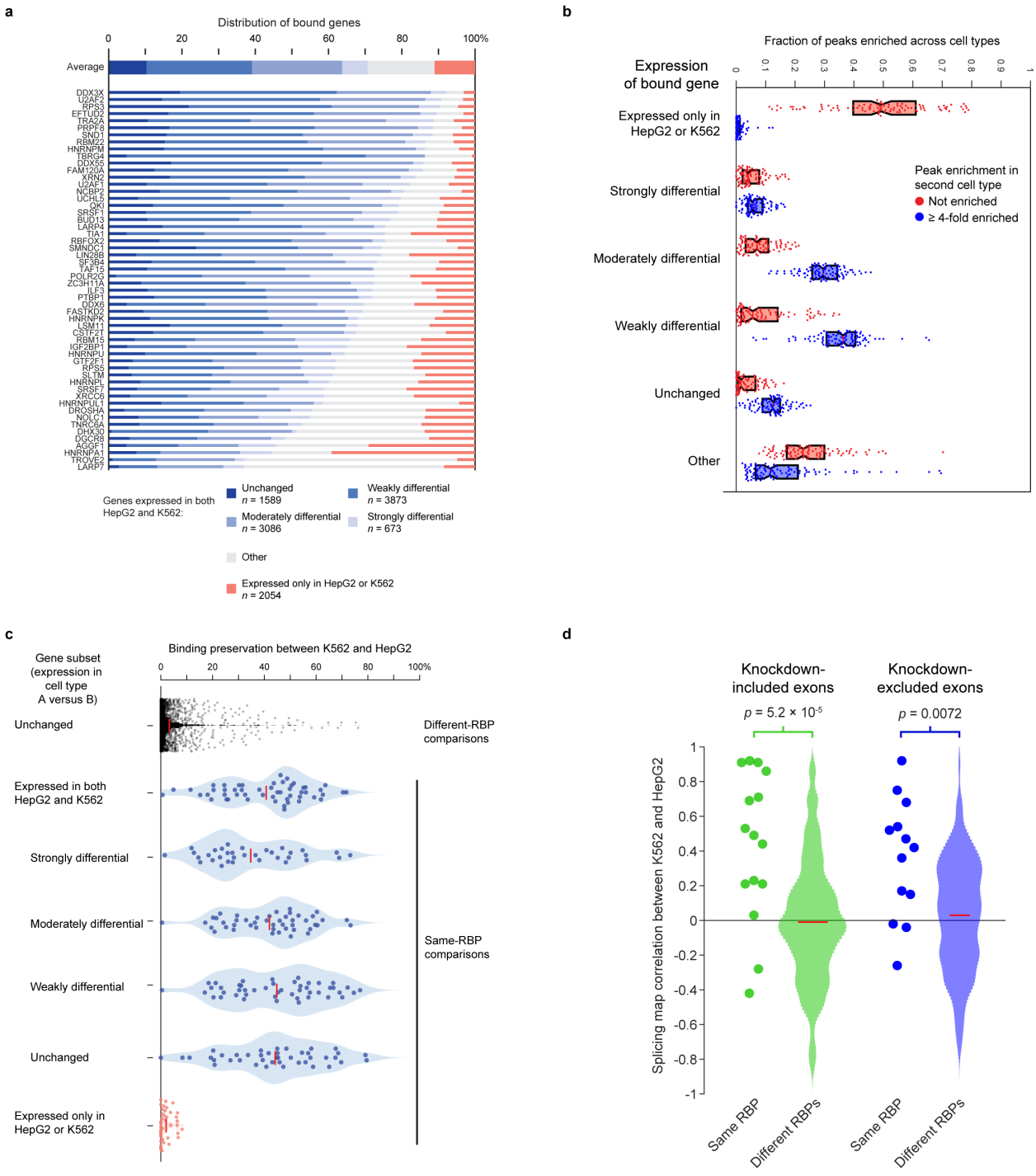

Figure 3-S10: **Preservation of binding across cell types**

Figure 3-S10

(**a**) Bars indicate the fraction of peaks observed for each RBP within sets of genes separated by their relative expression change between K562 and HepG2: unchanged (fold-difference $\leq$ 1.2), weakly (1.2 $<$ fold-difference $\leq$ 2), moderately (2 $<$ fold-difference $\leq$ 5) or strongly (fold-difference $>$ 5) differential, or cell type-specific genes (TPM $<$ 0.1 in one cell type and TPM $\geq$ 1 in the other). n indicates the number of genes meeting each criteria. For each RBP, the results shown are for the cell type with fewer total peaks.

(**b**) Each point represents one eCLIP dataset compared with the same RBP profiled in the second cell type. For the set of peaks from the first cell type that are not enriched (fold-enrichment $<$ 1) in the second cell type, red points indicate the fraction occurring in genes with the indicated expression difference between HepG2 and K562. Blue points similarly indicate the gene distribution of peaks four-fold enriched in the opposite cell type. Boxes indicate quartiles, with median indicated by the central red line.

(**c**) Points indicate the fraction of overlapping peaks between K562 and HepG2 for RBPs profiled in both cell types (blue or red), or between one RBP in K562 and a second in HepG2 (black), for sets of genes separated by their relative expression change between K562 and HepG2 as in (**a**).

(**d**) Plot represents the distribution of Pearson correlations between splicing maps as shown in **Fig.** 3-8**d**, separated by whether the comparison is between the same RBP or different RBPs profiled in two different cell types. Different RBPs are shown as smoothed histogram using a Normal kernel, and red line indicates mean. Significance was determined by Kolmogorov-Smirnov test.

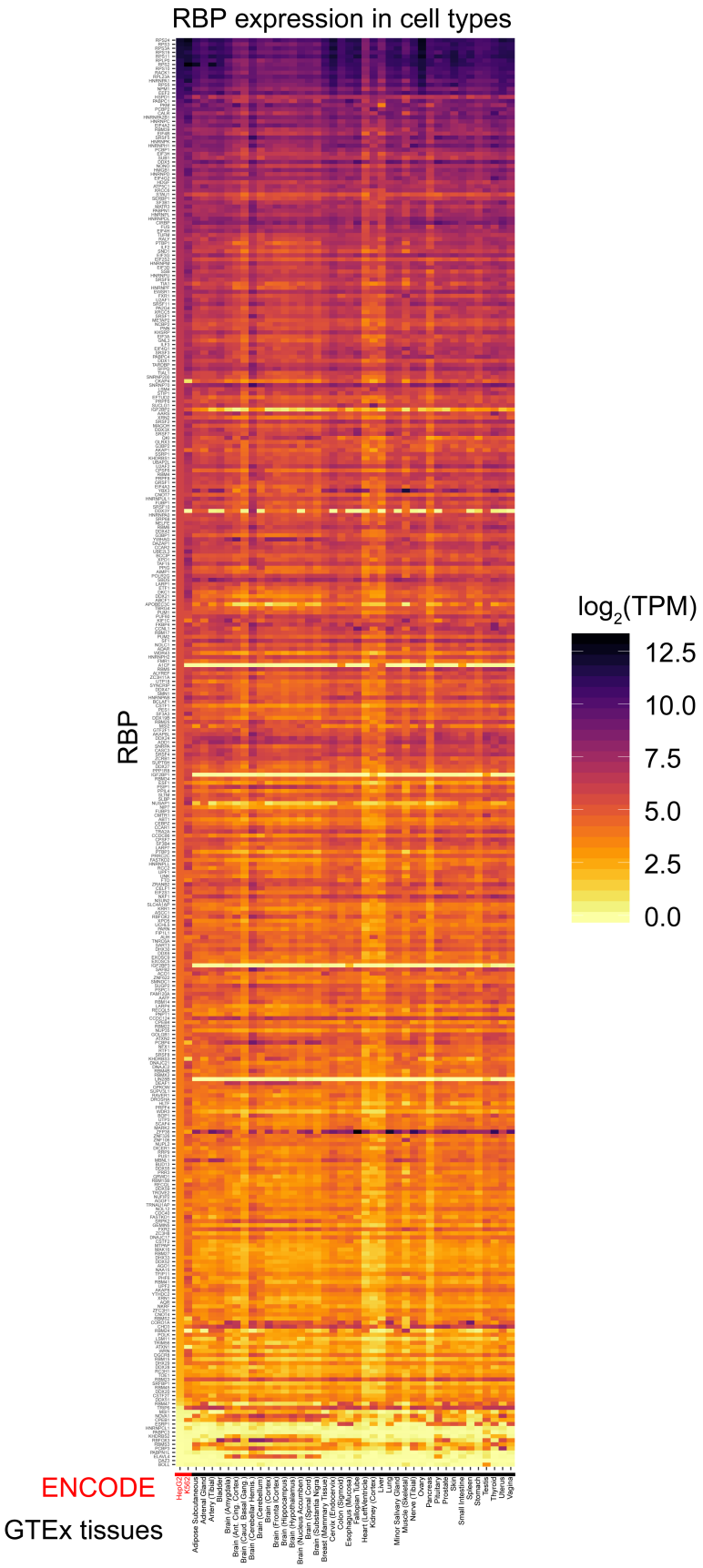

Figure 3-S11: **Expression of RBPs across tissues and cell types**

Figure 3-S11

Expression of the 352 RBPs (in Transcripts Per Million) investigated in this study in ENCODE cell lines HepG2 and K562 as well as 40 human tissues measured by the GTEx project. RBPs sorted by decreasing expression in HepG2.

3.6 Methods

3.6.1 RNA binding protein annotations and domains

RBPs were chosen from a previously described list of 1072 known RBPs, proteins containing RNA binding domains, and proteins characterized as associated with polyadenylated RNA, based on the availability of high quality antibodies (Sundararaman et al. [2016]). Annotation of RBP function was performed by integration of published literature, with manual inspection of references for less well-established annotations. Annotation of RNA binding domain presence was determined by UniProt Domain Descriptions, and a database of cell-essential genes was obtained from published high-throughput CRISPR screening efforts (Wang et al. [2015]).

3.6.2 eCLIP - experimental methods

Antibodies for eCLIP were pre-screened using a set of defined metrics (Sundararaman et al. [2016]). A 'biosample' of HepG2 or K562 cells was defined as a batch of cells starting from a single unfrozen stock, passaged for less than 30 days under standard ENCODE reference conditions, and validated for high viability and non-confluence at the time of crosslinking. All cells within a biosample were pooled and UV crosslinked on ice at 400 mJoules/cm^2 with 254 nm radiation. The biosample was then split into 20 million cell aliquots for eCLIP experiments.

eCLIP experiments were performed as previously described in a detailed Standard Operating Procedure (Van Nostrand et al. [2016]), which is provided as associated documentation with each eCLIP experiment on the ENCODE portal (https://www.encodeproject.org/documents/fa2a3246-6039-46ba-b960-17fe06e7876a/@@download/attachment/CLIP_SOP_v1.0.pdf). Briefly, 20 million crosslinked cells were lysed and sonicated, followed by treatment with RNase I (Thermo Fisher) to fragment RNA. Antibodies were pre-coupled to species-specific (anti-Rabbit IgG or anti-Mouse IgG) Dynabeads (Thermo Fisher), added to lysate, and incubated overnight at 4°C. Prior to immunoprecipitation (IP) washes, 2% of sample was removed to serve as the paired input sample.

For IP samples, high- and low-salt washes were performed, after which RNA was dephosphorylated with FastAP (Thermo Fisher) and T4 PNK (NEB) at low pH, and a 3' RNA adapter was ligated with T4 RNA Ligase (NEB). 10% of IP and input samples were run on an analytical PAGE Bis-Tris protein gel, transferred to PVDF membrane, blocked in 5% dry milk in TBST, incubated with the same primary antibody used for IP (typically at 1:4000 dilution), washed, incubated with secondary HRP-conjugated species-specific TrueBlot antibody (Rockland), and visualized with standard enhanced chemiluminescence imaging to validate successful IP. 90% of IP and input samples were run on an analytical PAGE Bis-Tris protein gel and transferred to nitrocellulose membranes, after which the region from the protein size to 75 kDa above protein size was excised from the membrane, treated with Proteinase K (NEB) to release RNA, and concentrated by column purification (Zymo). Input samples were then dephosphorylated with FastAP (Thermo Fisher) and T4 PNK (NEB) at low pH, and a 3' RNA adapter was ligated with T4 RNA Ligase (NEB) to synchronize with IP samples. Reverse transcription was then performed with AffinityScript (Agilent), followed by ExoSAP-IT (Affymetrix) treatment to remove unincorporated primer. RNA was then degraded by alkaline hydrolysis, and a 3' DNA adapter was ligated with T4 RNA Ligase (NEB). qPCR was then used to determine required amplification, followed by PCR with Q5 (NEB) and gel electrophoresis to size-select the final library. Libraries were sequenced on either the HiSeq 2000, 2500, or 4000 platform (Illumina). Each ENCODE eCLIP experiment consisted of IP from two independent biosamples, along with one paired size-matched input (sampled from one of the two IP lysates prior to IP washes).

3.6.3 eCLIP - data processing and peak identification

Data processing - including adapter trimming, repetitive element removal, unique genomic mapping, PCR duplicate removal, and peak calling versus paired size-matched input - were performed as previously described in a detailed Standard Operating Procedure (Van Nostrand et al. [2016]). Unless otherwise noted, reproducible and significant peaks were identified by merging peaks identified in each replicate, requiring that the peak meet an irreproducible discovery rate cutoff of 0.01 as well as p-value ≤ 0.001 and fold-enrichment ≥ 8 (using the geometric mean of $\log_2$(fold-enrichment) and -$\log_{10}$(p-value) between the two biological repli-

cates). For submission to the ENCODE portal, eCLIP datasets were required to pass several quality metrics, including minimum read depth, reproducibility, and peak saturation metrics. To quantify binding to snRNA and other multi-copy elements, a separate pipeline was developed to require unique mapping to element families instead of unique genomic positions. All analyses described in this manuscript used mapping to GRCh37 and GENCODE v19 annotations, but mapping to GRCh38 and GENCODE v24 annotations were also deposited at the ENCODE portal.

For analyses using binding considered at the level of regions (e.g. 3' UTR, CDS, or proximal intronic), read density was counted for the indicated region for both IP and paired input, and significance was determined by Fisher Exact test (or Yates' Chi-Square test if all observed and expected values were above 5). Only regions with at least 10 reads in one of IP or input, and where at least 10 reads would be expected in the comparison dataset given the total number of usable reads, were considered, and significant regions were defined as those with fold-enrichment ≥ 4 and p-value ≤ 0.00001.

To summarize relative enrichment between IP and input, information was defined as the Kullback-Leibler divergence (relative entropy): $p_i \times \log_2 (p_i/q_i)$, where p_i is the fraction of total reads in IP that map to a queried element i (peak, gene, or repetitive element), and q_i is the fraction of total reads in input for the same element.

3.6.4 Knockdown followed by RNA-seq (KD/RNAseq) - experimental methods

Individual RBPs were depleted from HepG2 or K562 cells by either RNA interference or CRISPR-mediated gene disruption. RNA interference was performed by transducing cells with lentiviruses expressing shRNAs (TRC collection) targeting an RBP of interest followed by puromycin selection for 5 days. CRISPR-mediated gene disruption was performed by transfecting cells with a plasmid expressing Cas9 and a gRNA targeting the RBP of interest, followed by puromycin selection for 5 days. In each case, depletions were performed in biological duplicate along with a pair of control depletions using a scrambled shRNA or gRNA. RNA was extracted from half of each sample and used to perform qRT-PCR to

confirm depletion of the targeted RBP transcript and to prepare RNA-seq libraries using the Illumina Tru-seq stranded mRNA library preparation kit. Protein was extracted from the other half of each sample and used to confirm depletion of the target RBP by Western blotting. Paired-end 100 bp reads were generated from the RNA-seq libraries to an average depth of 63 million reads per replicate.

3.6.5 KD/RNA-seq - data processing

Reads were aligned to both the hg19 assembly of the genome using the Gencode v19 annotation, respectively, using both TopHat2 (Kim et al. [2013]) and STAR (Dobin et al. [2013]). Gene expression levels were quantitated using RSEM (Li and Dewey [2011]) and Cufflinks (Trapnell et al. [2012]) and differential expression levels determined using DESeq2 (Anders and Huber [2010]) and Cuffdiff2 (Trapnell et al. [2013]). Alternative splicing was quantitated using rMATS (Shen et al. [2014]) and alternative poly(A) site use quantitated with MISO (Katz et al. [2010]). All analyses described in this manuscript used mapping to GRCh37 and GENCODE v19 annotations, but mapping to GRCh38 and GENCODE v24 annotations were also deposited at the ENCODE portal.

3.6.6 RNA Bind-N-Seq (RBNS) - experimental methods

RBNS experiments were performed as indicated in the protocol included on each experiment at the ENCODE portal. Briefly, randomized RNA oligonucleotides (20 or 40 nt) flanked by constant adapter sequences were synthesized and incubated with an SBP-tagged recombinant RBP (consisting minimally of all annotated RNA binding domains) at several concentrations (typically five, ranging from 5-1300 nM). RNA-protein complexes were isolated with streptavidin-conjugated affinity resin and eluted RNA was prepared for deep sequencing, resulting in 15-20 million reads per RBP pulldown concentration with a similar number of input reads sequenced per *in vitro* transcription reaction.

3.6.7 RBNS - data processing

RBNS $k$mer enrichments (R values) were calculated as the frequency of each $k$mer in the pulldown library reads divided by its frequency in the input library; enrichments from the pulldown library with the highest individual $k$mer R value were used for each RBP. Mean and SD of R values were calculated across all $k$mers for a given k to calculate the RBNS Z-score for each $k$mer. RBNS pipeline source code is available at: https://bitbucket.org/pfreese/rbns_pipeline.

RBNS motif logos were made using the following iterative procedure for k=5: the most enriched 5mer was given a weight equal to its excess enrichment over the input library (=R-1), and all occurrences of that 5mer were masked in both the pulldown and input libraries to eliminate subsequent counting of lower-affinity 'shadow' 5mers (e.g., GGGGA shifted by 1 from GGGGG). All enrichments were then recalculated on the masked read sets to obtain the resulting most enriched 5mer and its corresponding weight, with this process continuing until the enrichment Z-score (calculated from the original R values) was less than 3. All 5mers determined from this procedure were aligned to minimize mismatches to the most enriched 5mer, with a new motif initiated if the number of mismatches + offsets exceeds 2. The frequencies of each nucleotide in the position weight matrix, as well as the overall percentage of each motif, were determined from the weights of the individual aligned 5mers that went into that motif; empty unaligned positions before or after each aligned 5mer were given pseudocounts of 25% each nucleotide, and outermost positions of the motif logo were trimmed if they had >75% unaligned positions. To improve the robustness of the motif logos, the pulldown and input read sets were each divided in half and the above procedure was performed independently on each half; only 5mers identified in corresponding motif logos from both halves were included in the alignments to make the final motif logo. In **Fig.** 3-3**a**, only the top RBNS motif logo is shown if there were multiple (all motifs displayed on the ENCODE portal within the "Documents" box of each experiment).

The RBPs' top RBNS logos were clustered using Jensen-Shannon divergence (JSD)-based similarities computed by summing the score of the j overlapping positions between the motifs

of RBP A and RBP B:

$$\sum_{\text{aligned pos. } i=1,\ldots,j} \text{info}_{\text{A,i}} \times \text{info}_{\text{B,i}} \times \left(1 - \sqrt{\text{JSD}[\overrightarrow{ACGU}_{\text{A,i}} || \overrightarrow{ACGU}_{\text{B,i}}]}\right)$$

where $\text{info}_{\text{A,i}}$ and $\text{info}_{\text{B,i}}$ are the information content of motifs A and B at aligned position i, and A_i and B_i are the vectors of base frequencies at aligned position i in motifs A and B, respectively (with each A_i and B_i vector summing to 1). This score weights similarity more heavily at positions with higher information content and greater numbers of aligned positions. This similarity score was computed for each possible overlap of the two logos (with at least four positions overlapping, i.e., $j \geq 4$), and the top score with its corresponding alignment offset was used. The matrix of these scores was normalized to the maximum score over all RBP pairs and clustered using the linkage function with centroid method in scipy.cluster.hierarchy to obtain the dendrogram shown in **Fig.** 3-3**a**.

3.6.8 Immuno-Fluorescence, Microscopy Imaging and Data Processing

HepG2 cells were seeded in Poly-L-Lysine coated 96-well clear bottom plates (Corning Inc; plate number 3882 half-area microplates), at a concentration of 2,000 cells per well in DMEM + 10% FBS. After 72 hr in standard growth conditions (i.e. 37°C and 5% CO_2), cells were fixed with 3.7% formaldehyde, permeabilized in PBS + 0.5% Triton X-100 and blocked in PBS + 0.2% Tween-20 + 2% BSA (PBTB), all conducted for 20 min at room temperature. Primary antibodies directed against specific RBPs (all rabbit antibodies) and marker proteins were subsequently applied to the cells at a final concentration of 2 μg/mL in PBTB and incubated overnight at 4°C. The cells were next washed 3 times for 10 min each in PBST and incubated with secondary antibodies (Alexa647 donkey anti-rabbit and Alexa488 donkey anti-mouse, both diluted 1:500 in PBTB) for 90 min at room temperature. After 3 PBTB washes, the cells were counter-stained with DAPI for 5 min, washed 3 times in PBS and stored in PBS at 4°C. Subcellular marker antibodies and dilutions used are as follows: rat anti-Alpha Tubulin, MCA78G, 1:200 (Serotec, Bio-Rad); mouse anti-CD63, ab8219, 1:200 (Abcam);

mouse anti-Coilin, GTX11822, 1:100 (GeneTex Inc); mouse anti-DCP1a, sc100706, 1:200 (Santa Cruz Biotechnology); mouse anti-Fibrillarin, ab4566, 1:200 dilution (Abcam); mouse anti-GM130, #610822, 1:200 (Becton Dickinson); mouse anti-KDEL, ENZSPA827D, 1:200 (Enzo Life Sciences); mouse anti-Phospho Tyrosine, #9411S, 1:200 (NEB); mouse anti-PML, sc-966, 1:50 (Santa Cruz Biotechnology); mouse anti-SC35, GTX11826, 1:200 (GeneTex Inc). For staining with Mitotracker (Molecular Probes, M22426), cells were incubated with 100 nM of dye in tissue culture media for 45 min at 37°C prior to fixation. For staining with Phalloidin (Sigma, P5282), cells were incubated with 50 μg/ml of Phalloidin for 20 min prior to DAPI staining.

Imaging was conducted on an ImageXpress Micro high content screening system (Molecular Devices Inc). For each RBP/marker combination, 10-20 high resolution images were acquired in the DAPI, FITC, and Cy5 channels, using a 40x objective. Automated laser based auto-focusing and auto-exposure functions were employed for sample imaging, with exposure times ranging from 250-3000 ms, 100-500 ms and 50-100 ms, for RBP, Marker, and DAPI channels, respectively. Raw unprocessed grayscale images from individual channels were acquired as high resolution TIF files of 726 kb each. An in-house Matlab script was developed to batch normalize image intensity values and add blue, green, or red colors to the respective channels, which were subsequently merged as colour JPEG files. The final images were uploaded on a server accessible through the RBP Image Database website. A MySQL relational database (version 5.1.73) was implemented, along with a MyISAM storage engine, to store the images, data annotations, and characteristics. A controlled vocabulary of descriptors was devised to document RBP subcellular localization features.

Image analysis to quantify nuclear/cytoplasmic staining ratios, or to assess the degree of RBP targeting to punctate subcellular structures (e.g. Cajal bodies, nuclear speckles, nucleoli, Golgi, P-bodies), was conducted using 'Granularity', 'Colocalization', and 'Multi Wavelength Cell Scoring' analysis modules from the MetaXpress v3.1 software (Molecular Devices Inc), according to manufacturer recommendations. For localization categories including microtubules, actin, cell cortex, ER, focal adhesions, mitochondria and mitotic apparatus, manual localization grading was conducted by ranking candidate RBPS as strongly or weakly co-localized with respective protein markers. The Circos plot of localization co-

occurrance (**Fig.** 3-7**b**) was generated by drawing one line between every pair of categories for each RBP that shared both localization annotations. Nuclear annotations are indicated in purple, cytoplasmic in red, and lines between nuclear and cytoplasmic annotations are indicated in orange.

3.6.9 ChIP-seq - experimental methods

Chromatin immunoprecipitation was implemented according to ChIP Protocol optimized for RNA binding proteins (https://www.encodeproject.org/documents/e8a2fef1-580b-45ad-b29c-fffc3d527202/@@download/attachment/ChIP-seq_Protocol_for_RNA-Binding_Proteins_ENCODE_Fu_lab_RuiXiao.pdf). In brief, prior to coupling with RBP antibodies, magnetic beads were equilibrated by washing with ChIP dilution buffer and blocked with glycogen, BSA, and tRNA in ChIP dilution buffer. 10-20 million HepG2 and K562 cells were crosslinked in 1% formaldehyde diluted in PBS for 20 minutes and then quenched by adding glycine. Cell nuclei were extracted by resuspending cell pellet with cell lysis buffer with occasional inversion. Nucleus pellets, resuspended in nuclear lysis buffer, were sonicated with Branson Sonifier cell disruptor. 95% of nuclear lysate diluted in the final concentration of 1% triton X-100, 0.1% sodium deoxycholate, and 1X proteinase inhibitor cocktail was subjected to immunoprecipitation with antibody-coupled beads and 5% of nuclear lysate was used as input chromatin. Stringent washes were performed before elution. Input and immunoprecipitated chromatin DNAs were recovered by de-crosslinking, RNase A digestion, proteinase K treatment, phenol/chloroform extraction, and precipitation with ethanol. Library construction was mainly followed the instruction of the Illumina Preparing Samples for ChIP sequencing. Each library was barcoded for pooled sequencing. DNA Libraries between 200-400 bp were gel purified, quantified with Qubit, and subjected to Illumina HiSeq 2000/2500 sequencing.

3.6.10 ChIP-seq - data processing

Data processing was performed in accordance with ENCODE uniform transcription factor ChIP-seq pipeline (https://www.encodeproject.org/chip-seq/transcription_factor) and us-

ing hg19 as the reference human genome. All datasets containing >10 million usable reads from each replicate, passing IDR, and generating >200 peaks were used for final analysis.

3.6.11 Integrated Analysis

Saturation Analysis

Saturation analysis of eCLIP and KD/RNA-seq data was performed by randomly shuffling the order of datasets 100 times, subsampling 1 through all datasets, and calculating the desired metrics. Gene level saturation analysis of RBP binding was calculated first by taking all unique genes that were bound by an IDR-filtered peak in an eCLIP experiment. Then, each eCLIP experiment was iteratively added to the previous experiment, counting only unique genes in any experiment. Saturation analysis of differentially expression genes from KD/RNA-seq was similarly performed, based on differentially expressed genes identified with DESeq2. Genes were identified as differentially expressed if they had a $|\log_2(\text{fold-change})| > 1$ and an adjusted p-value < 0.05 between knockdown and control. Alternative versions of this analysis used: 3-2**b**) all genes (**Fig.**; only genes with TPM > 1 in HepG2 and K562 (**Fig.** 3-S2**a**); or only genes with TPM > 1 in either HepG2 or K562 (**Fig.** 3-S2**b**), using average gene-level expression from two rRNA-depleted RNA-seq experiments in HepG2 (ENCODE accession ENCFF533XPJ, ENCFF321JIT) and K562 (ENCFF286GLL, ENCFF986DBN). The set of differentially expressed and bound genes was determined by taking all genes differentially expressed upon RBP KD that contained at least one IDR-filtered peak in the corresponding eCLIP experiment in the same cell type.

Differentially spliced events were defined as those meeting p-value < 0.05, FDR < 0.1, and $|\Delta\Psi| > 0.05$ from rMATS analysis (described above). The number of unique events was defined as the number of non-overlapping events upon combining all experiments for a given sampling. A differentially spliced event was considered bound if for any RBP in which the event was differentially included upon KD, there was an eCLIP peak for the same RBP in the same cell type between the start of the upstream flanking exon and the end of the downstream flanking exon for cassette exons and mutually exclusive exons, start of the upstream flanking exon and end of the common exon region for A3SS, start of the common

exon and end of the common exon region for A5SS, and start of the upstream and stop of the downstream exons for retained introns.

To perform saturation of transcript regions, the highest expressed transcript for each gene was first identified using transcript-level quantifications from the same rRNA-depleted RNA-seq experiments described above (K562, accession numbers ENCFF424CXV, ENCFF073NHK; HepG2, accession numbers ENCFF205WUQ, ENCFF915JUZ). The following regions were then identified: the entire unspliced transcript (pre-mRNA), all exons (exon), 5' untranslated regions (5' UTR), coding sequence (CDS), 3' untranslated regions (3' UTR), all introns (intron), 100 nt intronic regions flanking the 5' and 3' splice sites (splice site), proximal intronic regions extending from 100 nt to 500 nt from the 5' and 3' splice site (prox. intron), and distal intronic regions extending from 500 nt and beyond from the 5' and 3' splice sites. Saturation calculations were then performed as described above for all genes (**Fig.** 3-2**c-d**, **Fig.** 3-S2**e-g**) or only genes with TPM $>$ 1 in both K562 and HepG2 (**Fig.** 3-S2**d**), and plotted as either the total number of bases covered (**Fig.** 3-2**c**), or the fraction of covered bases divided by the total number of bases in that annotation across all genes (**Fig.** 3-2**d**, **Fig.** 3-S2**d**). The ratio of bases covered was calculated by dividing the number of bases covered in subsampling of N+1 datasets divided by the number covered in subsampling N datasets.

Analysis of the fold-increase between one and two datasets (**Fig.** 3-S2**f**) was determined by first taking all 55 RBPs profiled in both HepG2 and K562, and calculating the fold-increase in covered bases by considering 110 comparisons including HepG2 followed by K562 and K562 followed by HepG2. Then, for each of the 110 datasets, 10 random other datasets were chosen from the same cell type, and for each of the 10 the fold-increase in covered bases from adding that dataset to the first was calculated. To compare the fold-increase between profiling new RBPs in additional cell lines, eCLIP datasets profiling RBFOX2, IGF2BP1, IGF2BP2, and IGF2BP3 in H9 human embryonic stem cells were obtained from the Gene Expression Omnibus (GSE78509, Conway et al. [2016]) and added as the 181st dataset. These were compared against profiling a new RBP in K562 or HepG2 (calculated by adding each of the 126 profiled RBPs as the 179 (if it was profiled in both cell types) or 180 (if it was profiled in only one cell type) datasets for other RBPs), or a profiled RBP done in

second cell type (calculated by sampling 180 datasets, and adding the 181st if it was an RBP already profiled in the other cell type).

Motif comparisons between RBNS and eCLIP

eCLIP 6mer Z-scores in **Fig.** 3-3**b** were calculated as previously described (Kapeli et al. [2016]). Briefly, peaks and a shuffled background set of peaks that preserves the region of binding (3' UTR, 5' UTR, CDS, exon, proximal and distal intron) were generated. EMBOSS compseq (http://structure.usc.edu/emboss/compseq.html) was used on these two peak sets and the Z-scores of the difference between real and background 6mer frequencies was calculated.

To produce eCLIP logos in a similar manner for comparison with RBNS logos, an analogous procedure was carried out on the eCLIP peak sequences (only eCLIP peaks at least 2-fold enriched were used): the two halves of the RBNS pulldown read set were replaced with the two eCLIP replicate peak sequence sets (each peak was extended 50 nt upstream of its 5' end as some RBPs have motif enrichments symmetrically around or only upstream of the peak starts), and the input RBNS sequences were replaced by random regions within the same gene as each peak that preserved peak length and transcript region (5' and 3' UTR peaks were chosen randomly within that region; intronic and CDS peaks were shuffled to a position within the same gene that preserved the peak start's distance to the closest intron/exon boundary to match sequence biases resulting from CDS and splicing constraints). The enrichment Z-score threshold for 5mers included in eCLIP logos was 2.8, as this threshold produced eCLIP logos containing the most similar number of 5mers to that of the $Z \geq 3$ 5mer RBNS logos. Each eCLIP motif logo was filtered to include only 5mers that occurred in both of the corresponding eCLIP replicate logos. eCLIP motif logos were made separately for all eCLIP peaks, only 3' UTR peaks, only CDS peaks, and only intronic peaks, with the eCLIP logo of those 4 (or 8 if CLIP was performed in both cell types) with highest similarity score to the RBNS logo shown in **Fig.** 3-3**a**, where the similarity score was the same as previously described to cluster RBNS logos (eCLIP logos for all transcript regions shown in **Fig.** 3-S3**a**). To determine significance of overlap between RBNS and eCLIP, a hypergeometric test was performed with 5mers in all RBNS logos, eCLIP logo 5mers (for peaks in

the region with highest similarity score to the RBNS logo), and 5mers in their intersection, relative to the background of all 1,024 5mers; overlap was deemed significant if $p < 0.05$. The top 'eCLIP-only' logo in each region was the highest eCLIP logo, if any, comprised of 5mers that had no overlap with any RBNS Z≥3 5mers (always using at least the top 10 RBNS 5mers if there were fewer than 10 with Z≥3).

All eCLIP/RBNS comparisons were for the same RBP with the following exceptions in which the eCLIP RBP was compared to a closely related RBNS protein: KHDRBS2 eCLIP versus KHDRBS1 RBNS; PABPN1 eCLIP versus PABPN1L RBNS; PTBP1 eCLIP versus PTBP3 RBNS; PUM2 eCLIP versus PUM1 RBNS; and RBM15 eCLIP versus RBM15B RBNS.

Splicing regulatory effects of RBNS+ and RBNS- eCLIP peaks

To assess the splicing regulatory effects of RBNS+ and RBNS- eCLIP peaks for **Fig. 3-3c**, only rMATS SEs with a Ψ between 0.05 and 0.95 in at least one of the control or KD were considered for each RBP. Each eCLIP peak (extended 50 nt 5' of the peak start) was first checked if it overlapped the SE, and if not then if it overlapped the upstream or downstream flanking 250 nt. To compare the magnitude of splicing changes upon KD for eCLIP+ vs eCLIP- SEs while minimizing the confounding factors of different wildtype host gene expression level and SE Ψ values among these two sets of SEs, a matched set of eCLIP- SEs was created by selecting for each eCLIP+ SE an SE in the same decile of wildtype gene expression and wildtype Ψ for each corresponding SE with an eCLIP peak. A CDF of the $|\Delta\Psi|$ changes upon KD was compared for the eCLIP+ versus eCLIP- SEs in each of the 6 SE direction/eCLIP region combinations ([included, excluded SE] × [peak over SE, upstream intron, downstream intron]), with significance $p < 0.05$ for a one-sided Wilcoxon rank-sum test that $|\Delta\Psi|_{\text{SE,peak}} > |\Delta\Psi|_{\text{SE,no peak}}$. If the eCLIP+ vs eCLIP- comparison was significant, the eCLIP peaks were divided into those that did and did not contain the top RBNS 5mer. The $|\Delta\Psi|$ values for all RBPs in each of the 6 SE direction/eCLIP regions were combined for comparison in **Fig. 3-3c**; see **Fig. 3-S4a** for RBPs that were significant in each region (12 included/4 excluded upon KD, upstream intron eCLIP peak; 11 included/2 excluded upon KD, SE eCLIP peak; 7 included/7 excluded upon KD, downstream intron

eCLIP peak). To assess eCLIP peaks with or without the top 'eCLIP-only' 5mer, the top 5mer from the aforementioned 'eCLIP-only' logo was used from the first region with an eCLIP-only logo among: all peaks; CDS peaks; intron peaks; and 3' UTR peaks (the more highly enriched 5mer if eCLIP was performed in both cell types). The resulting 'eCLIP-only' 5mers for Fig. 3-S4b were: CELF1 (CUCUC); EIF4G2 (GUGUG); EWSR1 (CGCGG); FUBP3 (UUGUU); FUS (GUGUG); HNRNPC (GUCGC); HNRNPK (UCCCC); HNRNPL (none); IGF2BP1 (GUGUG); IGF2BP2 (CGCCG); KHDRBS2: (none); KHSRP (none); PABPN1L (CGCGG); PCBP2 (CGGCG); PTBP3 (GAAGA); PUM2 (UUUUU); RBFOX2 (GGGGG); RBM22 (GGUAA); SFPQ (UCCGG); SRSF5 (CGGCG); SRSF9 (CUGGA); TAF15 (AGGGA); TARDBP (GAAGA); TIA1 (CGCCG); TRA2A (GAGGG).

Overlaps between RBP binding and gene expression perturbation upon KD/RNA-seq

To increase sensitivity for gene expression analysis, significant binding was determined at the level of transcript regions (including 5' UTR, CDS, 3' UTR, and introns) instead of using peaks. To identify significant enrichment between binding and expression changes, genes with significantly enriched binding to regions ($p \leq 0.00001$ and $\log_2$(fold-enrichment)≥ 4, as described above) were overlapped with the set of genes with significantly altered expression in KD/RNA-seq ($|\log_2(\text{fold-change})| > 1$ and an adjusted p-value < 0.05 between knockdown and control from DEseq2 analysis). Enrichment was calculated separately for knockdown-increased and knockdown-decreased genes, with significance determined by Fisher Exact test (or Yates' Chi-Square test if all observed and expected values were above 5).

For cumulative distribution plots, genes were separated based on their eCLIP fold-enrichment in IP versus input for the indicated transcript region. To filter out non-expressed genes, genes were included only if the region had at least 10 reads in one of IP or input, and where at least 10 reads would be expected in the comparison dataset given the total number of usable reads.

To perform TIA1 motif enrichment analysis, first the fold-enrichment of each 5mer was calculated by comparing the frequency in 3' UTRs of genes increased or decreased upon TIA1 knockdown in K562 or HepG2 with the frequency in a set of control genes upon

knockdown (changed genes upon KD: DEseq2 adjusted p-val < 0.05 and |fold-change| $>$ 1.5; control genes: DEseq2 p-val > 0.5 and |fold-change| < 1.1, subsetted to match the starting expression of changing genes upon KD). The top 15 5mers in TIA1 RBNS were then highlighted among in the ranked ordering of all 1,024 5mers. For positional analysis, a meta-3' UTR was created by normalizing all 3' UTRs to a 100 nt window. For each normalized position, the frequency of the top 10 TIA1 RBNS 5mers was calculated for each of the up-regulated, down-regulated, and control gene sets. Significance at each position was determined by $p < 0.05$ in a binomial test comparing the number of up- or down-regulated genes that have one of the top 10 RBNS 5mers at that position under the null frequency that it is equal to the corresponding frequency observed in control genes.

RBP binding correlation with knockdown-perturbed splicing (splicing maps)

RBP binding/splicing maps were generated using eCLIP normalized (reads per million) read densities overlapped with alternatively spliced (AS) regions from rMATS JunctionCountsOnly files from the same cell type. First, the set of differentially alternatively spliced events of the desired type (cassette/skipped exons (SE), alternative 5' splice site (A5SS), or alternative 3' splice site (A3SS) events were identified (**Fig.** 3-S6**a**), requiring rMATS p-value < 0.05, FDR < 0.1, and $|\Delta\Psi| > 0.05$ in knockdown versus control RNA-seq. To eliminate potential double counting of CLIP densities, overlapping AS events were additionally filtered to choose only the events containing the highest average inclusion junction count (IJC) among all replicates (using the bedtools v2.26 command merge (-o collapse -c 4) and pybedtools 0.7.9).

Next, for each splicing event, read densities were normalized across all regions separately for IP and input, in order to equally weigh each event. Per-position input probability densities were then subtracted from IP probability densities to attain position-level enrichment or depletion, for regions extending 50 nt into each exon and 300 nt into each intron composing the event, referred to as 'Normalized eCLIP enrichment' (**Fig.** 3-S6**b**). For shorter exons (<100 nt) and introns (<600 nt), densities were only counted until the boundary of the neighboring feature. Skipped exon (SE) maps were plotted using eCLIP densities overlapping the following 4 regions around AS events: 3' end of the upstream exon, 5' end of the

cassette, 3' end of the cassette, and 5' end of the downstream exon. Alternative 3' splice site (A3SS) maps were defined with three regions: 3' end of the upstream exon, 5' end of the longer transcript, and the 5' end of the shorter transcript. Alternative 5' splice site (A5SS) maps were defined with three regions: 3' end of the shorter transcript, 3' end of the longer transcript, and the 5' end of the downstream exon.

Plots of eCLIP signal enrichment (referred to as 'splicing maps') were then created by calculating the mean and standard error of the mean over all events after removing the highest (2.5%) and lowest (2.5%) outlying signal at each position, referred to as 'Average eCLIP enrichment' (**Fig.** 3-S6**c**). Splicing maps were only considered for RBPs with 100 or more altered cassette exon events, or 50 or more alternative 5' or 3' splice site events. As a background reference for cassette exon comparisons, sets of 1,832 (HepG2) and 2,244 (K562) 'native' cassette exons were identified which had $0.05 < \Psi < 0.95$ in at least half of control shRNA RNA-seq datasets for that cell type. Similar sets of 208 (K562) and 164 (HepG2) native alternative 5' splice site and 393 (K562) and 361 (HepG2) native alternative 3' splice site events were identified that had $0.05 < \Psi < 0.95$ in at least half of control shRNA RNA-seq datasets for that cell type. RBP-responsive event eCLIP enrichment was then calculated as eCLIP signal enrichment at RBP-regulated events minus eCLIP signal enrichment at native control events, referred to as 'Enrichment relative to control events' (**Fig.** 3-S6**d**). Native versus constitutive enrichment (**Fig.** 3-S7**c**) was similarly calculated as the eCLIP signal enrichment at native exons minus the eCLIP signal enrichment at constitutive exons (defined as exons with zero reads supporting exclusion in any control RNA-seq in that cell type).

Correlation between splicing maps was defined as the Pearson correlation (R) between a vector containing both included-upon knockdown and excluded-upon knockdown RBP-responsive event eCLIP enrichment for each RBP. If an RBP had less than the minimum required number of events (100 for cassette exons or 50 for alternative 5' or 3' splice site events) for either knockdown-included or knockdown-excluded events, the correlation was only calculated using the other event type.

To calculate the number of RBPs bound per exon, the set of spliceosomal RBPs was taken from manual annotation of RBP functions (described above and listed in Supplementary

Table 1). The number of reproducible (IDR) peaks at each position relative to splice sites was summed across all RBPs and divided by the total number of cassette or constitutive exons, respectively.

Comparison of DNA and RNA binding properties of RBPs

For integrative analyses, DNaseI HS data (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeOpenChromSynth), histone modifications by ChIP-seq from ENCODE/Broad Institute (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeBroadHistone), and eCLIP-seq data from ENCODE were downloaded and compared with RBP ChIP-seq data.

To explore the possibility that some RBP-chromatin association events might be coupled with their direct RNA binding activities in cells, RNA binding peaks were compared with DNA binding signals as assayed by ChIP-seq to quantify enrichment. Only eCLIP peaks in gene body regions (excluding promoter and terminator regions, defined as the 1 kb surrounding regions of TSS and TTS) were considered. The ChIP-seq signals were calculated for each eCLIP peak, together with surrounding regions that are 10 times the length of eCLIP peak on each side. Wilcoxon rank-sum tests were then performed to see whether ChIP-seq signal was enriched at the middle third regions.

To see whether those differentially-expressed genes after RBP knockdown were enriched in RBP binding at the chromatin level, an equal number of genes with similar expression level either with or without binding were randomly sampled, and the number of differentially-expressed genes after knockdown of the RBP were counted (fold change > 1.5 or $< 2/3$, adjusted p-value < 0.05 by DESeq2), and one-tailed Fisher's exact tests were then performed to test the dependence of RBP binding and differential expression. The above procedure was performed 100 times to give the distribution of the odds ratio. A significant dependence was defined when the null hypothesis was rejected at level of 0.05 for at least 95 times. The correlations between RBP association and genes with regulated alternative splicing events (A3SS, A5SS, RI, MXE and SE events) were investigated similarly.

Analysis of RBP regulatory features in subcellular space

Localization annotations and calculation of nuclear versus cytoplasmic ratio were generated from immunofluorescence imaging as described above. 'Nuclear RBPs' were defined as those with nuclear/cytoplasmic ratio ≥ 2, and 'Cytoplasmic RBPs' were defined as those with nuclear/cytoplasmic ratio ≤ 0.5. Spliced reads were defined as reads mapping across an annotated GENCODE v19 splice junction (extending at least 10 bases into each exon) and unspliced reads were defined as reads that overlapped an exon-intron junction (extending at least 10 bases into both the exon and intron regions). Significance between groups was determined by Wilcoxon rank-sum test.

Prediction of RNA secondary structure was performed using the RNAfold webserver (http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAfold.cgi, Gruber et al. [2008]) with default parameters. Shown is the MFE secondary structure prediction.

Preservation of RBP regulation across cell types

To consider binding across cell types, first the highest expressed transcript for each gene was identified using transcript-level quantifications from the same rRNA-depleted RNA-seq experiments described above (K562, accession numbers ENCFF424CXV, ENCFF073NHK; HepG2, accession numbers ENCFF205WUQ, ENCFF915JUZ) and used as representative for that gene. Next, genes were categorized based on their absolute fold-difference (FD) between K562 and HepG2: unchanged (FD ≤ 1.2), weakly ($1.2 <$ FD ≤ 2), moderately ($2 <$ FD ≤ 5) or strongly (FD > 5) differential (for each, requiring TPM ≥ 1 in both K562 and HepG2), or cell type-specific genes (TPM < 0.1 in one cell type and TPM ≥ 1 in the other). Peaks were then categorized based upon the expression change of their associated gene. Significance between groups was determined by Kolmogorov-Smirnov test.

Analysis of preservation of binding across cell types was considered in three ways. First, for each peak identified in one cell type, the fold-enrichment for that region in the other cell type was calculated and considered for each gene type (as in **Fig.** 3-8**a**). Two groups of peaks were then identified: those that were $\geq$4-fold enriched in the other cell type, and those that were not enriched in the other cell type. The fraction of peaks associated with a

gene class that were either ≥4-fold or not enriched were then considered for each gene class separately (**Fig.** 3-8**b**). Second, the set of peaks ≥4-fold enriched (and the set not enriched) was compiled across all genes, and the fractions associated with each gene class were then reported (**Fig.** 3-S10**b**). Finally, binding preservation was calculated by determining the fraction of IDR peaks identified in one cell type that overlap (requiring at least 1 nt overlap) IDR peaks identified in the second cell type (**Fig.** 3-S10**c**).

Correlation between splicing maps was performed as described above, considering all RBPs profiled by eCLIP and RNA-seq in both K562 and HepG2 that had at least 100 differentially included or excluded cassette exon events in both cell types.

RBP expression in tissues

Tissue specificity was measured as the entropy deviation from a uniform distribution among all tissues as in Gerstberger et al. [2014]. For each RBP, the log_2(TPM+1) was calculated for each of the 42 samples (HepG2, K562, and 40 tissues profiled by the GTEx consortium, GTEx Consortium [2015]), and the tissue specificity was computed as the difference between the logarithm of the total number of samples ($N = 42$) and the Shannon entropy of the expression values for an RBP:

$$S = H_{\max} - H_{\text{obs}} = \log_2(N) - \sum_{i=1,\ldots,N} \left[p_i \times \log_2(p_i)\right],$$

Where $p_i = x_i \;/\; \left(\sum_{i=1,\ldots,N} x_i\right)$ for $x_i = \log_2(\text{TPM}_i + 1)$ in sample i.

The data used for the analyses were obtained from dbGaP accession number phs000424.v2.p1 in Jan. 2015. TPMs were measured using kallisto (Bray et al. [2016]) on the following samples: Adipose-Subcutaneous: SRR1081567; AdrenalGland: SRR1120913; Artery-Tibial: SRR817094; Bladder: SRR1086236; Brain-Amygdala: SRR1085015; Brain-AnteriorCingulateCortex: SRR814989; Brain-CaudateBasalGanglia: SRR657731; Brain-CerebellarHemisphere: SRR1098519; Brain-Cerebellum: SRR627299; Brain-Cortex: SRR816770; Brain-FrontalCortex: SRR657777; Brain-Hippocampus: SRR614814; Brain-Hypothalamus: SRR661179; Brain-NucleusAccumben: SRR602808; Brain-SpinalCord: SRR613807; Brain-SubstantiaNigra: SRR662138; Breast-MammaryTissue: SRR1084674; Cervix: SRR1096057; Colon: SRR1091524; Esophagus:

SRR1085211; FallopianTube: SRR1082520; Heart-LeftVentricle: SRR815517; Kidney-Cortex: SRR809943; Liver: SRR1090556; Lung: SRR1081283; MinorSalivaryGland: SRR1081589; Muscle-Skeletal: SRR820907; Nerve-Tibial: SRR612911; Ovary: SRR1102005; Pancreas: SRR1081259; Pituitary: SRR1077968; Prostate: SRR1099402; Skin: SRR807775; Small-Intestine: SRR1093314; Spleen: SRR1085087; Stomach: SRR814268; Testis: SRR1081449; Thyroid: SRR808886; Uterus: SRR820026; Vagina: SRR1095599.

Chapter 4

Conclusion

Integration of the RBP characterizations included here with further cellular and functional data will pave the way for understanding, defining, and perhaps even predicting a code of RNA recognition and RBP functional activity in cells. Large-scale binding (e.g., eCLIP) and functional (KD/RNA-seq) data provides a catalog of information on the RNA side of RBP-RNA interactions, while biochemical and structural RBP binding data provide a molecular basis for defining cellular RBP-RNA interactions with greater confidence, creating a foundation for greater mechanistic understanding of post-transcriptional gene regulation. Interdisciplinary work, which this thesis endeavors to expand, will be needed to fully understand RBP-mediated RNA processing mechanisms and their aberrations in human disease.

4.1 Summary

Chapter 2 presented the affinity landscapes of 78 human RBPs using RNA Bind-n-Seq (RBNS), an unbiased assay that determines the sequence, structure, and context preferences of an RBP *in vitro* by deep sequencing of bound RNAs selected from a random pool of oligos. "RNA maps" of RBP activity in regulating alternative splicing events and gene expression levels were made using just the RBNS motifs without incorporation of data from crosslinking-based assays. An unexpectedly low diversity of RNA motifs was bound by the assayed RBPs, implying frequent convergence of binding specificity toward a relatively small set of RNA motifs, many with low compositional complexity composed primarily of one

or two bases. However, extensive preferences for contextual features distinct from short linear RNA motifs were observed, including spaced 'bipartite' motifs, biased flanking nucleotide composition, and bias away from or occasionally towards RNA structure. These *cis* contextual features are likely to play an important role in RNA recognition, adding to RBP regulation conferred at the *trans* level (e.g., RBP expression, subcellular localization, or post-translational modifications), and they likely enable targeting of distinct subsets of transcripts by different RBPs that recognize the same linear motif.

Chapter 3 presented an integration of five assays, each focused on a distinct aspect of RBP activity, to expand the catalog of functional elements encoded in the human genome by addition of those that function at the RNA level through interaction with RBPs. At least one of the assays (transcriptome-wide RNA binding sites of RBPs through eCLIP; determination of RBP-responsive genes and alternative splicing events through RBP KD/RNA-seq; *in vitro* RBP binding motifs through RBNS; RBP subcellular localization through systematic imaging; and RBP association with chromatin through ChIP-seq) was applied to 352 human RBPs involved in diverse cellular functions containing a multitude of RNA binding domain types. In addition to describing the generation of over 1000 replicated data sets and methods for processing them, findings included that *in vivo* binding is largely determined by *in vitro* binding specificity; dozens of RBPs are associated with changes in gene expression levels or alternative splicing, enabling construction of RNA maps of RBP context-dependent regulation of cassette exon, A3SS, and A5SS splicing; many RBPs are associated with chromatin and this coupling can affect gene expression levels and splicing outcomes; RBPs display a broad diversity of subcellular localization patterns with most factors exhibiting targeting to multiple structures in the nucleus and cytoplasm; and RBP binding is typically preserved among events expressed in different cell types.

4.2 Future Directions

4.2.1 Impact of post-transcriptional RNA and post-translational protein modifications on RBP-RNA binding

While the presence of RNA modifications in select RNA subtypes such as rRNAs, snRNAs, and snoRNAs has been known for decades, recent work enabled by high-throughput sequencing-based mapping technologies has uncovered at least seven 'epitranscriptomic' RNA modifications in mRNA: N^6-methyladenosine (m^6A), 2'-O-dimethyl-adenosine (m^6Am), N^1-methyladenosine (m^1A), pseudouridine (ψ), inosine (I), 5-methylcytidine (m^5C), and 5-hydroxymethylcytidine (hm^5C) (Kleiner [2017]). While m^6A, the most abundant internal modification of mRNA, has been implicated in diverse molecular functions including mRNA export, cap-independent translation, translational efficiency, mRNA stability, and mRNA structure (Gilbert et al. [2016]), less is known about the molecular consequences or functional significance of the other modifications. m^6A not only recruits effector YTH domain-containing reader proteins, but it has also been shown to bind HNRNPA2B1 to affect alternative splicing and primary microRNA processing (Alarcón et al. [2015]). Another recent study used a systematic mass spectrometry-based proteomics screen to identify 21 m^6A reader proteins in HeLa cells, including five proteins investigated in the RBNS assay in Chapter 2 (KHSRP, FUBP3, TARDBP, HNRNPF, and HNRNPH) as well as the stress granule protein G3BP1, whose binding was repelled by m^6A to negate that RBP's normal activity of stabilizing mRNAs in cells (Edupuganti et al. [2017]). Through transcriptome-wide profiling in HeLa and multiple mouse tissues, 5-methylcytosine (m^5C) has been mapped to regions immediately downstream of translation initiation sites, with the RRM-containing ALYREF export factor functioning as a specific mRNA m^5C-binding protein and mRNA export being promoted by the m^5C modification (Yang et al. [2017]). As these epitranscriptomic modifications may function at least in part by regulating the binding of specific RBPs, identifying which RBP(s) are affected by each modification and the functional outcomes resulting from such post-transcriptional RNA changes will be an important area of investigation going forward.

In additional to post-transcriptional modifications to the RNA, post-translational modifications (PTMs) of the protein *trans*-factors can affect RBP binding and RNA processing. While PTMs have classically been shown to alter RBP subcellular localization (including sub-nuclear localization and nucleocytoplasmic shuffling), protein complex formation, and enzymatic activity (Lovci et al. [2016]), studies have also shown that particular PTMs affect RNA binding ability and mediate cellular outcomes. For example, phosphorylation of HuR (ELAVL1) RRM1 induces its association with TRA2β exon 2 to regulate alternative splicing programs under oxidative stress (Akaike et al. [2014]), arginine methylation of the RG repeats of Sam68 abrogates that region's poly(U) binding activity *in vitro* and *in vivo* (Rho et al. [2007]), phosphorylation of a highly conserved serine in HNRNPL RRM4 induces binding to an RNA element in the Slo1 potassium channel transcript to regulate membrane depolarization/calcium signaling (Liu et al. [2012]), and tyrosine phosphorylation of the QKI C-terminus modulates its RNA binding activity to regulate central nervous system myelination (Zhang et al. [2003]). The effects of PTMs on RBP function have thus far been unpredictable (Lovci et al. [2016]), necessitating experimental probing of the effect of particular PTMs on individual RBPs. As the bacterially expressed and purified RBPs studied in Chapter 2 did not contain any PTMs the RBPs might have in mammalian cells, strategies to profile the effects of PTMs on RNA binding could include incubating the purified RBP with the known or putative PTM catalytic enyzme or purifying the RBP from human cells to obtain PTM-modified proteins.

4.2.2 Role of alternative protein isoforms and low-complexity domains in RNA binding specificity and higher-order protein assemblies

Over 95% of human multi-exonic genes undergo alternative splicing to produce different protein isoforms (Wang et al. [2008]). These alternative splicing patterns are often regulated in a tissue- and/or developmental-specific manner, resulting in distinct protein repertoires in different cell types. Recent work has demonstrated that many splicing factors are themselves alternatively spliced with different regulatory capacities of different protein isoforms

representing an important and recurrent mechanism underlying splicing regulation. While the RBNS assay in Chapter 2 was only performed on one isoform of each human RBP (typically the Uniprot reference isoform), as alternative splicing of RBPs can affect which RBD binding domain(s) are included in the protein and may directly alter RNA binding specificity or affinity, performing RBNS on different isoforms may provide insight into this mode of RNA binding regulation. For example, inclusion of alternative exon 2 of HNRNPD inhibits its RNA binding ability by about 10-fold, reducing its ability to mediate mRNA decay via binding to AU-rich elements. As this exon immediate precedes RRM1, its inclusion may affect the conformation of the RRM or its access to RNA substrates (Zucconi et al. [2010]). Autoregulation of the Fox proteins via alternative splicing of a highly conserved 93 nt exon produces an isoform that lacks the second half of its sole RRM, producing a dominant negative version of this splicing factor (Damianov and Black [2010]).

In addition to directly altering RBD regions, alternative splicing of RBPs can produce different isoforms to modulate their subcellular localization, interaction with protein partners, or functional activity. For example, mammalian-specific alternatively spliced exons in nearly all members of the hnRNP A and D families lead to differential inclusion of low-complexity, GY (glycine/tyrosine)-rich intrinsically disordered regions that control the formation of higher-order assemblies of these proteins to affect alternative splicing outcomes (Gueroussov et al. [2017]). Cataloging the repertoire of RBP isoforms present in different cell types or disease states, predicting the variants' effects on protein structure and/or function, and experimentally interrogating their influence on RNA binding ability (through RBNS) or binding locations and functional activity (through isoform-specific adaptations of eCLIP and RBP perturbation/RNA-seq) will be an interesting area of study going forward.

4.2.3 Integrative analysis of RBP binding data sets to relate genetic variation to RBP regulation & RBNS assay variants to probe altered interactions

The breadth of the RBP data presented in Chapter 3 enables integrative analyses to relate genetic variation to RBP regulation. For example, among the 18 RBPs with eCLIP, RBNS,

and KD/RNA-seq data, we identified 26 variants from the Exome Aggregation Consortium (ExAC, Lek et al. [2016]) that overlapped an eCLIP peak, disrupted an RBNS motif, and produced a splicing change upon knockdown of the corresponding RBP (Van Nostrand et al. [2017]). One such variant was in intron 66 of *UTRN* (dystrophin-related protein 1), which harbors an RBFOX2 eCLIP peak downstream of an alternatively spliced exon (**Fig.** 4-1**A**). The G→C variant disrupts the RBFOX2 binding motif (GCAUG) at the first position. RBNS data reveal that this variant substantially changes the RBFOX2 binding site: this top 5mer (GCAUG) has an enrichment value of 13.58, while the variant 5mer (CCAUG) has a enrichment of 0.89 (**Fig.** 4-1**B**), suggesting that the mutation disrupts RBFOX2 binding *in vivo*. Indeed, KD/RNA-seq of RBFOX2 in HepG2 indicated that in wildtype cells, the upstream exon was included in 87% of messages, whereas the inclusion of the exon was decreased to 28% in the RBFOX2 knockdown cells (**Fig.** 4-1**A**). Taken together, these data argue that this G→C variant disrupts RBFOX2 binding, leading to decreased inclusion of the upstream exon in over half of UTRN messages and an altered composition of protein isoforms. Overall, the actual number of variants that influence RNA metabolism is likely to be much larger than the 26 identified as other variants may affect aspects of RNA biology other than splicing (e.g., mRNA expression levels or translation). Furthermore, only 18 RBPs had all three assays (RBNS, eCLIP, and KD/RNA-seq) performed at the time of this analysis, so it will be worthwhile to revisit such analyses upon subsequent studies in which many more RBPs have been profiled with complementary *in vitro*, *in vivo*, and functional assays.

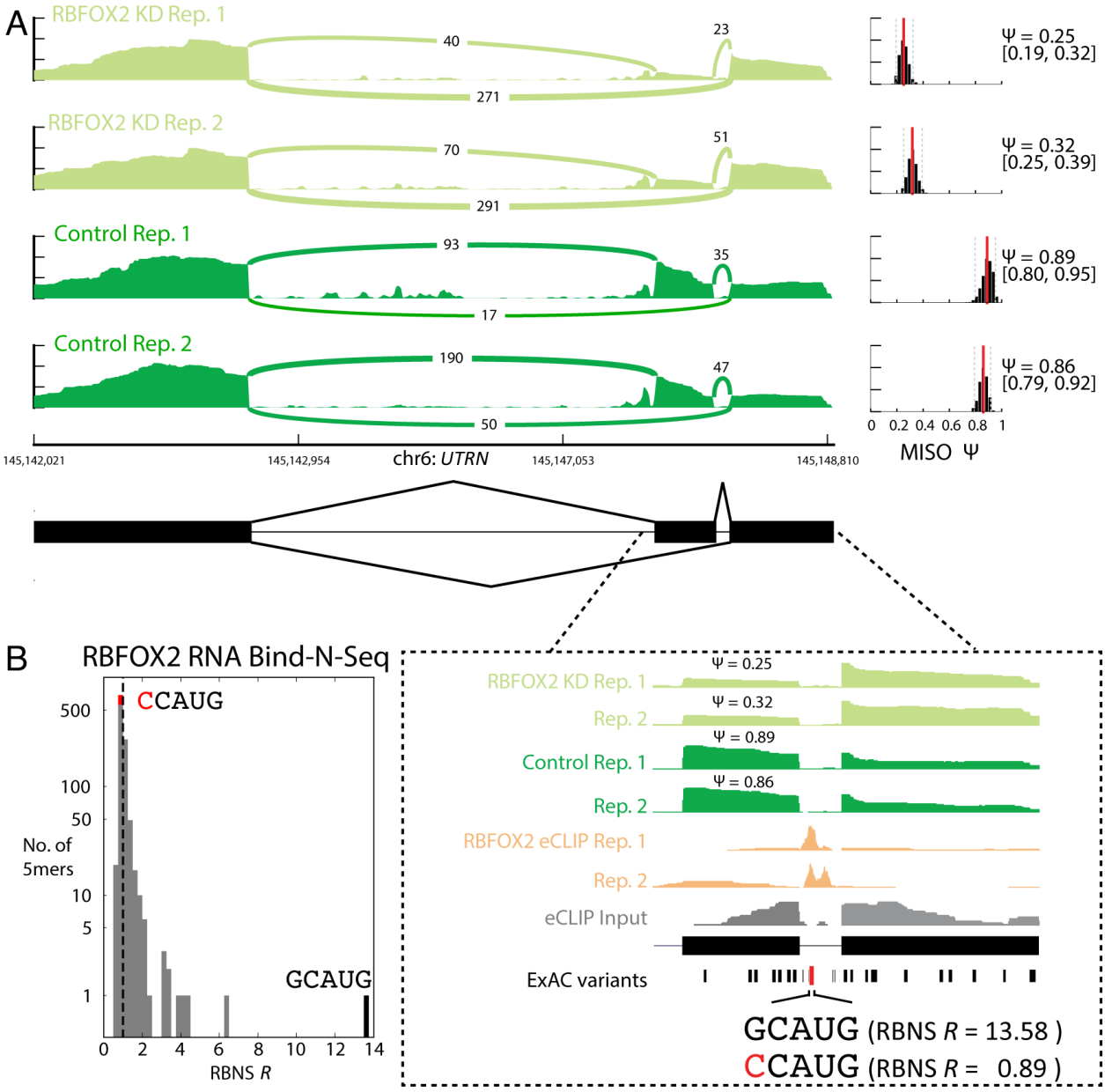

Figure 4-1: **Integrative analyses of RBP data can identify genetic variants that may impact RBP regulation**

To be included as panels of a figure in Moore et al.

(**A**) Control and RBFOX2 knockdown RNA-seq of exons 65-67 of the *UTRN* gene in HepG2 cells. Inclusion of the alternatively spliced exon 66 is reduced from 87% in control cells to 29% in RBFOX2 KD cells.

(**B**) (right) A strong RBFOX2 eCLIP binding peak in the downstream intron is consistent with this splicing factor enhancing inclusion of the upstream alternative exon. The minor allele of an ExAC SNP in the eCLIP peak in is expected to abrogate RBFOX2 binding as it abolishes the high affinity binding site determined from RBNS. (left) Effect of the ExAC variant on the RBFOX2 binding site as determined from RBNS data. The G→C SNP in the eCLIP peak changes the most enriched 5mer that likely mediates RBFOX2 binding (GCAUG $R = 13.78$) to a 5mer with no detectable *in vitro* binding (CCAUG $R = 0.89$).

An orthogonal approach that measures the affect of genetic variants on RBP binding could be achieved through a version of the RBNS assay that includes natural sequences rather than the random sequences used in Chapter 2. A natural sequence library of 109 nt long oligos derived from sequences flanking constitutive and alternative exons has successfully been used to measure secondary structure effects on an RBP's ability to bind these natural pre-mRNA regions *in vitro* (Taliaferro et al. [2016]), and a similar approach is currently being used in the Burge lab for a library of oligos derived from natural 3' UTR sequences. Adaptation of a natural sequence RBNS library to include oligos that contain not only the reference genome sequence but also disease SNPs in the transcriptome could define sets of RBP/genetic variant interactions that are abrogated or created in the context of specific human diseases. While this experiment is more limited in some respects than full integrative analysis of *in vitro*, *in vivo*, and functional assays, it does have the advantage that many RBPs can be profiled for their ability to bind the oligos in parallel, and RBPs and transcript sequences that exist only in restricted cell types can be assayed without performing eCLIP and/or KD/RNA-seq in new cell types. Going forward, utilizing high-throughput binding and functional technologies to understand RBP/RNA interactions in normal and disease states will be a ripe area of investigation.

Bibliography

OO Abudayyeh, JS Gootenberg, P Essletzbichler, S Han, J Joung, JJ Belanto, V Verdine, DBT Cox, MJ Kellner, A Regev, ES Lander, DF Voytas, AY Ting, and F Zhang. RNA targeting with CRISPR-Cas13. *Nature*, 550(7675):280–284, October 2017.

T Afroz, Z Cienikova, A Cléry, and FHT Allain. *One, Two, Three, Four! How Multiple RRMs Read the Genome Sequence*, volume 558 of *Methods in Enzymology*, pages 235–278. Elsevier, 2015.

V Agarwal, GW Bell, JW Nam, and DP Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4, August 2015.

AA Agrawal, E Salsi, R Chatrikhi, S Henderson, JL Jenkins, MR Green, DN Ermolenko, and CL Kielkopf. An extended U2AF(65)-RNA-binding domain recognizes the 3' splice site signal. *Nature Communications*, March 2016.

Y Akaike, K Masuda, Y Kuwano, K Nishida, K Kajita, K Kurokawa, Y Satake, K Shoda, I Imoto, and K Rokutan. HuR regulates alternative splicing of the TRA2β gene in human colon cancer cells under oxidative stress. *Molecular and Cellular Biology*, 34(15):2857–73, August 2014.

CR Alarcón, H Goodarzi, H Lee, X Liu, S Tavazoie, and SF Tavazoie. HNRNPA2B1 is a mediator of m6A-dependent nuclear RNA processing events. *Cell*, 162(6):1299–308, September 2015.

H Amrein, M Gorman, and R Nöthiger. The sex-determining gene *tra*-2 of *Drosophila* encodes a putative RNA binding protein. *Cell*, 55(6):1025–35, December 1988.

S Anders and W Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10), 2010.

H Antonicka and EA Shoubridge. Mitochondrial RNA granules are centers for posttranscriptional RNA processing and ribosome biogenesis. *Cell Reports*, pages 265–78, February 2015.

SD Auweter, FC Oberstrass, and FH Allain. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Research*, 34(17):4943–59, 2006.

D Baek, J Villén, C Shin, FD Camargo, SP Gygi, and DP Bartel. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, September 2008.

AG Baltz, M Munschauer, B Schwanhäusser, A Vasile, Y Murakawa, M Schueler, N Youngs, D Penfold-Brown, K Drew, M Milek, E Wyler, R Bonneau, M Selbach, C Dieterich, and M Landthaler. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Molecular Cell*, 46(5):674–90, June 2012.

FE Baralle and J Giudice. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, 18(7):437–451, July 2017.

J Baran-Gale, CL Kurtz, MR Erdos, C Sison, A Young, EE Fannin, PS Chines, and P Sethupathy. Addressing bias in small RNA library preparation for sequencing: A new protocol recovers microRNAs that evade capture by current methods. *Frontiers in Genetics*, 6 (352), December 2015.

Y Barash, JA Calarco, W Gao, Q Pan, X Wang, O Shai, BJ Blencowe, and BJ Frey. Deciphering the splicing code. *Nature*, 465(7294):53–9, May 2010.

C Barreau, L Paillard, and HB Osborne. AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Research*, 33(22):7138–50, January 2006.

C Beuck, S Qu, WS Fagg, M Ares Jr, and JR Williamson. Structural analysis of the quaking homodimerization interface. *Journal of Molecular Biology*, 423(5):766–781, November 2012.

V Botti, F McNicoll, MC Steiner, FM Richter, A Solovyeva, M Wegener, OD Schwich, I Poser, K Zarnack, I Wittig, KM Neugebauer, and M Müller-McNicoll. Cellular differentiation state modulates the mRNA export activity of SR proteins. *Journal of Cell Biology*, 216(7):1993–2009, July 2017.

RK Bradley, J Merkin, NJ Lambert, and CB Burge. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS Biology*, 10(1), January 2012.

KW Brannan, W Jin, SC Huelga, CA Banks, JM Gilmore, L Florens, MP Washburn, EL Van Nostrand, GA Pratt, MK Schwinn, DL Daniels, and GW Yeo. SONAR discovers RNA-binding proteins from analysis of large-scale protein-protein interactomes. *Molecular Cell*, 64(2):282–293, October 2016.

NL Bray, H Pimentel, P Melsted, and L Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–7, May 2016.

A Bregman, M Avraham-Kelbert, O Barkai, L Duek, A Guterman, and M Choder. Promoter elements regulate cytoplasmic mRNA decay. *Cell*, 147(7):1473–83, December 2011.

A Brünger. X-ray crystallography and NMR reveal complementary views of structure and dynamics. *Nature Structural Biology*, 4:862–5, October 1997.

JD Buenrostro, CL Araya, LM Chircus, CJ Layton, HY Chang, MP Snyder, and WJ Greenleaf. Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nature Biotechnology*, 32(6):562–8, June 2014.

E Buratti, AF Muro, M Giombi, D Gherbassi, A Iaconcig, and FE Baralle. RNA folding affects the recruitment of SR proteins by mouse and human polypurinic enhancer elements in the fibronectin EDA exon. *Molecular and Cellular Biology*, 24(3):1387–400, February 2004.

K Burger, B Mühl, M Kellner, M Rohrmoser, A Gruber-Eber, L Windhager, CC Friedel, L Dölken, and D Eick. 4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. *RNA Biology*, 10(10):1623–30, October 2013.

A Busch and KJ Hertel. Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdisciplinary Reviews: RNA*, 3(1):1–12, 2012.

ZT Campbell, D Bhimsaria, CT Valley, JA Rodriguez-Martinez, E Menichelli, JR Williamson, AZ Ansari, and Wickens M. Cooperativity in RNA-protein interactions: global analysis of RNA binding specificity. *Cell Reports*, 1(5):570–81, May 2012.

SM Carlson, CM Soulette, Z Yang, J Elias, AN Brooks, and O Gozani. RBM25 is a global splicing factor promoting inclusion of alternatively spliced exons and is itself regulated by lysine mono-methylation. *Journal of Biological Chemistry*, 292(32):13381–13390, August 2017.

A Castello, B Fischer, K Eichelbaum, R Horos, BM Beckmann, C Strein, NE Davey, DT Humphreys, T Preiss, LM Steinmetz, J Krijgsveld, and MW Hentze. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, 2149(6):1393–406, June 2012.

M Cereda, U Pozzoli, G Rot, P Juvan, A Schweitzer, T Clark, and J Ule. RNAmotifs: prediction of multivalent RNA motifs that control alternative splicing. *Genome Biology*, 15(1):R20, January 2014.

H Chang, J Lim, M Ha, and VN Kim. TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications. *Current Opinion in Structural Biology*, 53(6):1044–52, March 2014.

CD Chen, R Kobayashi, and DM Helfman. Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene. *Genes and Development*, 13(5):4562–71, August 1999.

ZG Chen, C Stauffacher, Y Li, T Schmidt, W Bomu, G Kamer, M Shanks, G Lomonossoff, and JE Johnson. Protein-RNA interactions in an icosahedral virus at 3.0 A resolution. *Science*, 245(4914):159–9, July 1989.

YD Choi, PJ Grabowski, PA Sharp, and G Dreyfuss. Heterogeneous nuclear ribonucleoproteins: role in RNA splicing. *Science*, 231(4745):1534–9, March 1986.

TB Chou, Z Zachar, and PM Bingham. Developmental expression of a regulatory gene is programmed at the level of splicing. *The EMBO Journal*, 6(13):4095–104, December 1987.

R Choudhury, SG Roy, YS Tsai, A Tripathy, LM Graves, and Z Wang. The splicing activator DAZAP1 integrates splicing control into MEK/Erk-regulated cell proliferation and migration. *Nature Communications*, 3078, 2014.

C Chu, QC Zhang, ST da Rocha, RA Flynn, M Bharadwaj, JM Calabrese, T Magnuson, E Heard, and HY Chang. Systematic discovery of Xist RNA binding proteins. *Cell*, 161 (2):404–16, April 2015.

Z Cieniková, FF Damberger, J Hall, FH Allain, and C Maris. Structural and mechanistic insights into poly(uridine) tract recognition by the hnRNP C RNA recognition motif. *Journal of the American Chemical Society*, 136(41):14536–44, October 2014.

A Cléry and FHT Allain. *From Structure to Function of RNA Binding Domains*, pages 137–58. RNA Binding Proteins. Landes Bioscience, January 2011.

AE Conway, EL Van Nostrand, GA Pratt, S Aigner, ML Wilbert, B Sundararaman, P Freese, NJ Lambert, S Sathe, TY Liang, A Essex, S Landais, CB Burge, DL Jones, and GW Yeo. Enhanced CLIP uncovers IMP protein-RNA targets in human pluripotent stem cells important for cell adhesion and survival. *Cell Reports*, 15(3):666–679, April 2016.

KB Cook, TR Hughes, and QD Morris. High-throughput characterization of protein-RNA interactions. *Briefings in Functional Genomics*, 14(1):74–89, January 2015.

KB Cook, S Vembu, KCH Ha, H Zheng, KU Laverty, TR Hughes, D Ray, and QD Morris. RNAcompete-S: Combined RNA sequence/structure preferences for RNA binding proteins derived from a single-step *in vitro* selection. *Methods*, 126:18–28, August 2017.

FH Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12: 138–63, 1958.

GE Crooks, G Hon, JM Chandonia, and SE Brenner. WebLogo: a sequence logo generator. *Genome Research*, 14(6):1188–90, June 2004.

DM Crothers, PE Cole, CW Hilbers, and RG Shulman. The molecular mechanism of thermal unfolding of *Escherichia coli* formylmethionine transfer RNA. *Journal of Molecular Biology*, 87(1):63–88, July 1974.

A Damianov and DL Black. Autoregulation of Fox protein expression to produce dominant negative splicing factors. *RNA*, 16(2):405–16, February 2010.

A Damianov, Y Ying, CH Lin, JA Lee, D Tran, AA Vashisht, E Bahrami-Samani, Y Xing, KC Martin, JA Wohlschlegel, and DL Black. Rbfox proteins regulate splicing as part of a large multiprotein complex LASR. *Cell*, 165(3):606–19, April 2016.

GM Daubner, A Cléry, and FH Allain. RRM-RNA recognition: NMR or crystallography...and new findings. *Current Opinion in Structural Biology*, 23(1):100–8, February 2013.

Y Ding, Y Tang, CK Kwok, Y Zhang, PC Bevilacqua, and SM Assmann. *In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, 505 (7485):696–700, January 2014.

A Dobin, CA Davis, F Schlesinger, J Drenkow, C Zaleski, S Jha, P Batut, M Chaisson, and TR Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013.

D Dominguez, P Freese, MS Alexis, A Su, M Hochman, T Palden, C Bazile, NJ Lambert, EL Van Nostrand, GA Pratt, GW Yeo, B Graveley, and CB Burge. Sequence, structure, and context preferences of human RNA binding proteins, Oct 2017. https://doi.org/10.1101/201996 bioRxiv; posted 10/12/2017.

KJ Doshi, JJ Cannone, CW Cobaugh, and RR Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, 5(5), August 2004.

JA Doudna and E Charpentier. The new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213), November 2014.

SR Eddy. How do RNA folding algorithms work? *Nature Biotechnology*, 22(11):1457–8, November 2004.

RR Edupuganti, S Geiger, RGH Lindeboom, H Shi, PJ Hsu, Z Lu, SY Wang, MPA Baltissen, PWTC Jansen, M Rossa, M Müller, HG Stunnenberg, C He, T Carell, and M Vermeulen. N6-methyladenosine (m^6A) recruits and repels proteins to regulate mRNA homeostasis. *Nature Structural & Molecular Biology*, 24(10):870–878, October 2017.

JM Engreitz, N Ollikainen, and M Guttman. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. *Nature Reviews Molecular Cellular Biology*, 17(12):756–770, December 2016.

S Erkelenz, WF Mueller, MS Evans, A Busch, K Schöneweis, KJ Hertel, and H Schaal. Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA*, 19(6):96–102, January 2013.

MR Fabian, F Frank, C Rouya, N Siddiqui, WS Lai, A Karetnikov, PJ Blackshear, B Nagar, and N Sonenberg. Structural basis for the recruitment of the human CCR4-NOT deadenylase complex by tristetraprolin. *Nature Structural & Molecular Biology*, 20(6):735–9, June 2013.

WG Fairbrother, RF Yeh, PA Sharp, and CB Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–13, August 2002.

J Font and JP MacKay. *Beyond DNA: Zinc Finger Domains as RNA-Binding Modules*, volume 558 of *Methods in Molecular Biology*, pages 479–491. Springer Protocols, 2010.

MB Friedersdorf and JD Keene. Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biology*, 15(1), January 2014.

RC Friedman, KK Farh, CB Burge, and DP Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, January 2009.

XD Fu and M Ares Jr. Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*, 15(10):689–701, October 2014.

XD Fu and T Maniatis. The 35-kDa mammalian splicing factor SC35 mediates specific interactions between U1 and U2 small nuclear ribonucleoprotein particles at the 3' splice site. *PNAS*, 89(5):1725–9, March 1992.

A Fukao and T Fujiwara. The coupled and uncoupled mechanisms by which *trans*-acting factors regulate mRNA stability and translation. *The Journal of Biochemistry*, 161(4): 309–314, April 2017.

A Fukao, Y Sasano, H Imataka, K Inoue, H Sakamoto, N Sonenberg, C Thoma, and T Fujiwara. The ELAV protein HuD stimulates cap-dependent translation in a Poly(A)- and eIF4A-dependent manner. *Molecular Cell*, 36(6):1007–17, December 2009.

T Fukunaga, H Ozaki, G Terai, K Asai, W Iwasaki, and H Kiryu. CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. *Genome Biology*, 15(1), January 2014.

G Gao, A Xie, SC Huang, A Zhou, J Zhang, AM Herman, S Ghassemzadeh, EM Jeong, S Kasturirangan, M Raicu, MA Sobieski, G Bhat, A Tatooles, EJ Benz Jr, TJ Kamp, and SC Dudley Jr. Role of RBM25/LUC7L3 in abnormal cardiac sodium channel splicing regulation in human heart failure. *Circulation*, 124(10):1124–31, September 2011.

K Gao, A Masuda, T Matsuura, and K Ohno. Human branch point consensus sequence is yUnAy. *Nucleic Acids Research*, 36(7):2257–67, April 2008.

H Ge and JL Manley. A protein factor, ASF, controls cell-specific alternative splicing of SV40 early pre-mRNA *in vitro*. *Cell*, 62(1):25–34, July 1990.

F Gebauer and MW Hentze. Molecular mechanisms of translational control. *Nature Reviews Molecular Cell Biology*, 5(10):827–35, October 2004.

S Gerstberger, M Hafner, and T Tuschl. A census of human RNA-binding proteins. *Nature Reviews Genetics*, 15(12):829–45, December 2014.

T Geuens, D Bouhy, and V Timmerman. The hnRNP family: insights into their role in health and disease. *Human Genetics*, 135(8):851–67, August 2016.

C Gilbert and JQ Svejstrup. RNA immunoprecipitation for determining RNA-protein associations *in vivo*. *Current Protocols in Molecular Biology*, August 2006.

WV Gilbert, TA Bell, and C Schaening. Messenger RNA modifications: Form, distribution, and function. *Science*, 352(6292):1408–12, June 2016.

G Giudice, F Sánchez-Cabo, C Torroja, and E Lara-Pezzi. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database*, April 2016.

T1 Glisovic, JL Bachorik, J Yong, and G Dreyfuss. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*, 582(14):1977–86, July 2008.

H Goodarzi, HS Najafabadi, P Oikonomou, TM Greco, L Fish, R Salavati, IM Cristea, and S Tavazoie. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature*, 485(7397):264–8, April 2012.

A Goren, O Ram, M Amit, H Keren, G Lev-Maor, I Vig, T Pupko, and G Ast. Comparative analysis identifies exonic splicing regulatory sequences-the complex definition of enhancers and silencers. *Molecular Cell*, 22(6):769–81, June 2006.

BR Graveley and T Maniatis. Arginine/serine-rich domains of SR proteins can function as activators of pre-mRNA splicing. *Molecular Cell*, 1(5):765–71, April 1998.

A Grimson, KK Farh, WK Johnston, P Garrett-Engele, LP Lim, and DP Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell*, 27 (1):91–105, July 2007.

NV Grishin. KH domain: one motif, two folds. *Nucleic Acids Research*, 29(3):638–43, February 2001.

AR Gruber, R Lorenz, SH Bernhart, R Neuböck, and IL Hofacker. The Vienna RNA websuite. *Nucleic Acids Research*, 36:W70–4, July 2008.

GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–60, May 2015.

S Gueroussov, RJ Weatheritt, D O'Hanlon, ZY Lin, A Narula, AC Gingras, and BJ Blencowe. Regulatory expansion in mammals of multivalent hnRNP assemblies that globally control alternative splicing. *Cell*, 170(2):324–339, July 2017.

RR Gutell, JC Lee, and JJ Cannone. The accuracy of ribosomal RNA comparative structure models. *Current Opinion in Structural Biology*, 12(3):301–10, June 2002.

M Hafner, M Landthaler, L Burger, M Khorshid, J Hausser, P Berninger, A Rothballer, M Ascano Jr, AC Jungkamp, M Munschauer, A Ulrich, GS Wardle, S Dewell, M Zavolan, and T Tuschl. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–41, April 2010.

M Hajiaghayi, A Condon, and HH Hoos. Analysis of energy-based algorithms for RNA secondary structure prediction. *BMC Bioinformatics*, 13(22), February 2012.

AS Halees, E Hitti, M Al-Saif, L Mahmoud, IA Vlasova-St Louis, DJ Beisang, PR Bohjanen, and K Khabar. Global assessment of GU-rich regulatory content and function in the human transcriptome. *RNA Biology*, 8(4):681–91, 2011.

M Halvorsen, JS Martin, S Broadaway, and A Laederach. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genetics*, 6(8), August 2010.

C Hauer, T Curk, S Anders, T Schwarzl, AM Alleaume, J Sieber, I Hollerer, M Bhuvanagiri, W Huber, MW Hentze, and AE Kulozik. Improved binding site assignment by high-resolution mapping of RNA-protein interactions using iCLIP. *Nature Communications*, 13(8), August 2015.

KJ Hertel and BR Graveley. RS domains contact the pre-mRNA throughout spliceosome assembly. *Trends in Biochemical Sciences*, 30(2):115–8, March 2005.

M Hiller, Z Zhang, R Backofen, and S Stamm. Pre-mRNA secondary structures influence exon recognition. *PLoS Genetics*, 3(11), November 2007.

DJ Hogan, DP Riordan, AP Gerber, D Herschlag, and PO Brown. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biology*, 6(10):e255, October 2008.

BP Hudson, MA Martinez-Yamout, HJ Dyson, and PE Wright. Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nature Structural & Molecular Biology*, 11(3):257–64, March 2004.

J Hui, G Reither, and A Bindereif. Novel functional role of CA repeats and hnRNP L in RNA stability. *RNA*, 9(8):931–6, August 2003.

NT Ingolia, S Ghaemmaghami, JR Newman, and JS Weissman. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924): 218–23, April 2009.

R Ishimura, G Nagy, I Dotu, H Zhou, XL Yang, P Schimmel, S Senju, Y Nishimura, JH Chuang, and SL Ackerman. Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. *Science*, 345(6195):6793–807, July 2014.

R Iwaoka, T Nagata, K Tsuda, T Imai, H Okano, N Kobayashi, and M Katahira. Structural insight into the recognition of r(UAG) by Musashi-1 RBD2, and construction of a model of Musashi-1 RBD1-2 bound to the minimum target RNA. *Molecules*, 22(7), July 2017.

CH Jan, RC Friedman, JG Ruby, and DP Bartel. Formation, regulation and evolution of *Caenorhabditis elegans* 3'UTRs. *Nature*, 469(7328):97–101, January 2011.

X Ji, Y Zhou, S Pandit, J Huang, H Li, CY Lin, R Xiao, CB Burge, and XD Fu. SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. *Cell*, 153(4):855–68, May 2013.

A Jolma, J Yan, T Whitington, J Toivonen, KR Nitta, P Rastas, E Morgunova, M Enge, M Taipale, G Wei, K Palin, JM Vaquerizas, R Vincentelli, NM Luscombe, TR Hughes, P Lemaire, E Ukkonen, T Kivioja, and J Taipale. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–39, January 2013.

DI Kao, GM Aldridge, IJ Weiler, and WT Greenough. Altered mRNA transport, docking, and protein translation in neurons lacking fragile X mental retardation protein. *PNAS*, 107(35):15601–6, August 2010.

K Kapeli, GA Pratt, AQ Vu, KR Hutt, FJ Martinez, B Sundararaman, R Batra, P Freese, NJ Lambert, SC Huelga, SJ Chun, TY Liang, J Chang, JP Donohue, L Shiue, J Zhang, H Zhu, F Cambi, E Kasarskis, S Hoon, M Ares Jr, CB Burge, J Ravits, F Rigo, and GW Yeo. Distinct and shared functions of ALS-associated proteins TDP-43, FUS and TAF15 revealed by multisystem analyses. *Nature Communications*, 12143, July 2016.

N Kastelic and M Landthaler. mRNA interactome capture in mammalian cells. *Methods*, 4126:38–43, August 2017.

PS Katsamba, S Park, and IA Laird-Offringa. Kinetic studies of RNA-protein interactions using surface plasmon resonance. *Methods*, 26(2):95–104, February 2002.

Y Katz, ET Wang, EM Airoldi, and CB Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–15, December 2010.

S Ke, S Shang, SM Kalachikov, I Morozova, L Yu, JJ Russo, J Ju, and LA Chasin. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Research*, 21(8): 1360–74, August 2011.

P Kerpedjiev, C Höner Zu Siederdissen, and IL Hofacker. Predicting RNA 3D structure using a coarse-grain helix-centered model. *RNA*, 21(6):1110–21, June 2015.

M Kertesz, Y Wan, E Mazor, JL Rinn, RC Nutter, HY Chang, and E Segal. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–7, September 2010.

D Kim, G Pertea, C Trapnell, H Pimentel, R Kelley, and SL Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4), April 2013.

OD King, AD Gitler, and J Shorter. The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease. *Brain Research*, 1462:61–80, June 2012.

RE Kleiner. Reading the RNA code. *Biochemistry*, September 2017.

J König, K Zarnack, G Rot, T Curk, M Kayikci, B Zupan, DJ Turner, NM Luscombe, and J Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology*, 17(7):909–915, July 2010.

AR Krainer, GC Conway, and D Kozak. Purification and characterization of pre-mRNA splicing factor SF2 from HeLa cells. *Genes & Development*, 4(7):1158–71, July 1990.

M Kullmann, U Göpfert, B Siewe, and L Hengst. ELAV/Hu proteins inhibit p27 translation via an IRES element in the p27 5'UTR. *Genes & Development*, 16(23):3087–99, December 2002.

SC Kwon, H Yi, K Eichelbaum, S Föhr, B Fischer, KT You, A Castello, J Krijgsveld, MW Hentze, and VN Kim. The RNA-binding protein repertoire of embryonic stem cells. *Nature Structural & Molecular Biology*, 20(9):1122–30, September 2013.

C Lagier-Tourenne, M Polymenidou, and DW Cleveland. TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. *Human Molecular Genetics*, 19(R1): R46–64, April 2010.

WS Lai, E Carballo, JR Strum, EA Kennington, RS Phillips, and PJ Blackshear. Evidence that tristetraprolin binds to AU-rich elements and promotes the deadenylation and destabilization of tumor necrosis factor alpha mRNA. *Molecular and Cellular Biology*, 19(6): 4311–23, June 1999.

N Lambert, A Robertson, M Jangi, S McGeary, PA Sharp, and CB Burge. RNA Bind-n-Seq: Quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Molecular Cell*, 54(5):887–900, June 2014.

CP Lapointe, MA Preston, D Wilinski, HAJ Saunders, ZT Campbell, and M Wickens. Architecture and dynamics of overlapped RNA regulatory networks. *RNA*, 23(11):1636–1647, November 2017.

MA Larkin, G Blackshields, NP Brown, R Chenna, PA McGettigan, H McWilliam, F Valentin, IM Wallace, A Wilm, R Lopez, JD Thompson, TJ Gibson, and DG Higgins. Clustal W and clustal X version 2.0. *Bioinformatics*, 23(21):2947–8, November 2007.

KA Leamy, SM Assmann, DH Mathews, and PC Bevilacqua. Bridging the gap between *in vitro* and *in vivo* RNA folding. *Quarterly Reviews of Biophysics*, 49, January 2016.

EK Lee, W Kim, K Tominaga, JL Martindale, X Yang, SS Subaran, OD Carlson, EM Mercken, RN Kulkarni, W Akamatsu, H Okano, N Perrone-Bizzozero, R de Cabo, JM Egan, and M Gorospe. RNA-binding protein HuD controls insulin translation. *Molecular Cell*, 45(6):826–35, March 2012.

FA Lefebvre, NAL Cody, LPB Bouvrette, J Bergalet, X Wang, and E Lécuyer. CeFra-seq: Systematic mapping of RNA subcellular distribution properties through cell fractionation coupled to deep-sequencing. *Methods*, 126:138–148, August 2017.

M Lek, KJ Karczewski, EV Minikel, KE Samocha, E Banks, T Fennell, AH O'Donnell-Luria, JS Ware, AJ Hill, BB Cummings, T Tukiainen, DP Birnbaum, JA Kosmicki, LE Duncan, K Estrada, F Zhao, J Zou, E Pierce-Hoffman, J Berghout, DN Cooper, N Deflaux, M DePristo, R Do, J Flannick, M Fromer, L Gauthier, J Goldstein, N Gupta, D Howrigan, A Kiezun, M Kurki, AL Moonshine, P Natarajan, L Orozco, GM Peloso, R Poplin, MA Rivas, V Ruano-Rubio, SA Rose, DM Ruderfer, K Shakir, PD Stenson, C Stevens, BP Thomas, G Tiao, MT Tusie-Luna, B Weisburd, HH Won, D Yu, DM Altshuler, D Ardissino, M Boehnke, J Danesh, S Donnelly, R Elosua, JC Florez, SB Gabriel, G Getz, SJ Glatt, CM Hultman, S Kathiresan, M Laakso, S McCarroll, M McCarthy, D McGovern, R McPherson, BM Neale, A Palotie, SM Purcell, D Saleheen, JM Scharf, P Sklar, PF Sullivan, J Tuomilehto, MT Tsuang, HC Watkins, JG Wilson, MJ Daly, DG MacArthur, and Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–91, August 2016.

B Li and CN Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(323), August 2011.

JB Li, EY Levanon, JK Yoon, J Aach, B Xie, E Leproust, K Zhang, Y Gao, and GM. Church. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science*, 324(5931):1210–3, May 2009.

X Li, G Quon, HD Lipshitz, and Q Morris. Predicting *in vivo* binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, 16(6):1096–107, June 2010.

DD Licatalosi, A Mele, JJ Fak, J Ule, M Kayikci, SW Chi, TA Clark, AC Schweitzer, JE Blume, X Wang, JC Darnell, and RB Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–9, November 2008.

LP Lim and CB Burge. A computational analysis of sequence features involved in recognition of short introns. *PNAS*, 98(20):11193–8, September 2001.

P Linder and E Jankowsky. From unwinding to clamping - the DEAD box RNA helicase family. *Nature Reviews Molecular & Cellular Biology*, 12(8):505–16, July 2011.

G Liu, A Razanau, Y Hai, J Yu, M Sohail, VG Lobo, J Chu, SK Kung, and J Xie. A conserved serine of heterogeneous nuclear ribonucleoprotein l (hnRNP l) mediates depolarization-regulated alternative splicing of potassium channels. *Journal of Biological Chemistry*, 287 (27):22709–16, June 2012.

R Lorenz, SH Bernhart, C Höner Zu Siederdissen, H Tafer, C Flamm, PF Stadler, and IL Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(26), November 2011.

MT Lovci, D Ghanem, H Marr, J Arnold, S Gee, M Parra, TY Liang, TJ Stark, LT Gehman, S Hoon, KB Massirer, GA Pratt, DL Black, JW Gray, JG Conboy, and GW Yeo. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nature Structural and Molecular Biology*, 20(12):1434–42, December 2013.

MT Lovci, MH Bengtson, and KB Massirer. *Post-Translational Modifications and RNA-Binding Proteins*, volume 907 of *Advances in Experimental Medicine and Biology*, pages 297–317. Springer Protocols, 2016.

MI Love, W Huber, and S Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 2014.

D Lu, MA Searles, and A Klug. Crystal structure of a zinc-finger-RNA complex reveals two modes of molecular recognition. *Nature*, 426(6962):96–100, November 2003.

JB Lucks, SA Mortimer, C Trapnell, S Luo, S Aviran, GP Schroth, L Pachter, JA Doudna, and AP Arkin. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *PNAS*, 108(27):11063–8, July 2011.

KE Lukong, KW Chang, EW Khandjian, and S Richard. RNA-binding proteins in human genetic disease. *Journal of Cell Biology*, 24(8):416–25, August 2008.

BM Lunde, C Moore, and G Varani. RNA-binding proteins: modular design for efficient function. *Nature Reviews Molecular and Cellular Biology*, 8(6):479–90, June 2007.

H Maatz, M Kolinski, N Hubner, and M Landthaler. Transcriptome-wide identification of RNA-binding protein binding sites using photoactivatable-ribonucleoside-enhanced crosslinking immunoprecipitation (PAR-CLIP). *Current Protocols in Molecular Biology*, 118(1):27.6.1–27.6.19, April 2017.

CD Mackereth, T Madl, S Bonnal, B Simon, K Zanier, A Gasch, V Rybin, J Valcárcel, and M Sattler. Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. *Nature*, 475(7356):408–11, July 2011.

AV Makeyev and SA Liebhaber. The poly(C)-binding proteins: a multiplicity of functions and a search for mechanisms. *RNA*, 8(3):265–78, March 2002.

L Martin, M Meier, SM Lyons, RV Sit, WF Marzluff, SR Quake, and HY Chang. Systematic reconstruction of RNA functional motifs with high-throughput microfluidics. *Nature Methods*, 9(12):1192–4, December 2012.

A Martini, R La Starza, H Janssen, C Bilhou-Nabera, A Corveleyn, R Somers, A Aventin, R Foà, A Hagemeijer, C Mecucci, and P Marynen. Recurrent rearrangement of the Ewing's sarcoma gene, EWSR1, or its homologue, TAF15, with the transcription factor CIZ/NMP4 in acute leukemia. *Cancer Research*, 62(19):5408–12, October 2002.

G Masliah, P Barraud, and FH Allain. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cellular and Molecular Life Sciences*, 70(11): 1875–95, June 2013.

DH Mathews, WN Moss, and DH Turner. Folding and finding RNA secondary structure. *Cold Spring Harbor Perspectives on Biology*, 2(12), December 2010.

E Matoulkova, E Michalova, B Vojtesek, and R Hrstka. The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA Biology*, 9(5):563–76, May 2012.

DM Mauger, C Lin, and MA Garcia-Blanco. hnRNP H and hnRNP F complex with Fox2 to silence fibroblast growth factor receptor 2 exon IIIc. *Molecular and Cellular Biology*, 28(17):5403–19, September 2008.

C Mayr. Regulation by 3'-untranslated regions. *Annual Review of Genetics*, 51:171–94, 2017.

K Mazan-Mamczarz, S Galbán, I López de Silanes, JL Martindale, U Atasoy, JD Keene, and M Gorospe. RNA-binding protein HuR enhances p53 translation in response to ultraviolet light irradiation. *PNAS*, 100(14):8354–9, July 2003.

AJ McCullough and SM Berget. G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Molecular and Cellular Biology*, 17(8):4562–71, August 1997.

CA McHugh, CK Chen, A Chow, CF Surka, C Tran, P McDonel, A Pandya-Jones, M Blanco, C Burghard, A Moradian, MJ Sweredoski, AA Shishkin, J Su, ES Lander, S Hess, K Plath, and M Guttman. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature*, 521(7551):232–6, May 2015.

NH Meyer, K Tripsianes, M Vincendeau, T Madl, F Kateb, R Brack-Werner, and M Sattler. Structural basis for homodimerization of the Src-associated during mitosis, 68-kDa protein (Sam68) Qua1 domain. *Journal of Biological Chemistry*, 285(37):28893–901, September 2010.

J Miller, AD McLachlan, and A Klug. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO*, 4(6):1609–14, June 1985.

SF Mitchell and R Parker. Principles and properties of eukaryotic mRNPs. *Molecular Cell*, 54(4):547–58, May 2014.

J Moore, MJ Purcaro, HE Pratt, CB Epstein, N Shoresh, J Adrian, T Kawli, CA Davis, A Dobin, R Kaul, J Halow, EL Van Nostrand, P Freese, DU Gorkin, Y He, M Mackiewicz, The ENCODE Consortium, JM Cherry, RM Myers, B Ren, BR Graveley, JA Stamatoyannopoulos, MB Gerstein, LA Pennacchio, T Gingeras, MP Snyder, BE Bernstein, B Wold, RC Hardison, and Z Weng. ENCODE Phase III: Building an encyclopaedia of candidate regulatory elements for human and mouse. Under review, Dec. 2017.

MJ Moore and NJ Proudfoot. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell*, 136(4):688–700, February 2009.

F Moretti, C Kaiser, A Zdanowicz-Specht, and MW Hentze. Regulation of sexual differentiation in *D.melanogaster* via alternative splicing of RNA from the transformer gene. *Cell*, 50(5):739–47, August 1987.

F Moretti, C Kaiser, A Zdanowicz-Specht, and MW Hentze. PABP and the poly(A) tail augment microRNA repression by facilitated miRISC binding. *Nature Structural & Molecular Biology*, 19(6):603–8, May 2012.

M Müller-McNicoll, V Botti, AM de Jesus Domingues, H Brandl, OD Schwich, MC Steiner, T Curk, I Poser, K Zarnack, and KM Neugebauer. SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. *Genes & Development*, 30(5):553–66, March 2016.

J Murn, M Teplova, K Zarnack, Y Shi, and DJ Patel. Recognition of distinct RNA motifs by the clustered CCCH zinc fingers of neuronal protein Unkempt. *Nature Structural & Molecular Biology*, 23(1):16–23, January 2016.

S Naftelberg, IE Schor, G Ast, and AR Kornblihtt. Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annual Review of Biochemistry*, 84: 165–98, 2015.

DA Nelles, MY Fang, MR O'Connell, JL Xu, SJ Markmiller, JA Doudna, and GW Yeo. Programmable RNA tracking in live cells with CRISPR/Cas9. *Cell*, 165(2):488–96, April 2016.

G Nicastro, MF García-Mayoral, D Hollingworth, G Kelly, SR Martin, P Briata, R Gherzi, and A Ramos. Noncanonical G recognition mediates KSRP regulation of let-7 biogenesis. *Nature Structural & Molecular Biology*, 19(12):1282–6, December 2012.

G Nicastro, IA Taylor, and A Ramos. KH-RNA interactions: back in the groove. *Current Opinion in Structural Biology*, 30:63–70, February 2015.

CB Nielsen, N Shomron, R Sandberg, E Hornstein, J Kitzman, and CB Burge. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*, 13(11):1894–910, November 2007.

JK Nussbacher, R Batra, C Lagier-Tourenne, and GW Yeo. RNA-binding proteins in neurodegeneration: Seq and you shall receive. *Trends in Neurosciences*, 38(4):226–36, April 2015.

FC Oberstrass, SD Auweter, M Erat, Y Hargous, A Henning, P Wenter, L Reymond, B Amir-Ahmady, S Pitsch, DL Black, and FH Allain. Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science*, 309(5743):2054–7, September 2005.

P Oikonomou, H Goodarzi, and S Tavazoie. Systematic identification of regulatory elements in conserved 3' UTRs of human transcripts. *Cell Reports*, 7(1):281–92, April 2014.

DH Ostareck, A Ostareck-Lederer, IN Shatsky, and MW Hentze. Lipoxygenase mRNA silencing in erythroid differentiation: The 3'UTR regulatory complex controls 60S ribosomal subunit joining. *Cell*, 104(2):281–90, January 2001.

BA Ozdilek, VF Thompson, NS Ahmed, CI White, RT Batey, and JC Schwartz. Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding. *Nucleic Acids Research*, 45(13):7984–7996, July 2017.

PS Page-McCaw, K Amonlirdviman, and PA Sharp. PUF60: a novel U2AF65-related splicing activity. *RNA*, 5(12):1548–60, December 1999.

MP Paronetto, T Achsel, A Massiello, CE Chalfant, and C Sette. The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x. *Journal of Cell Biology*, 176(7):929–39, March 2007.

S Piñol Roma, YD Choi, MJ Matunis, and G Dreyfuss. Immunopurification of heterogeneous nuclear ribonucleoprotein particles reveals an assortment of RNA-binding proteins. *Genes & Development*, 2(2):215–27, February 1988.

CC Query, RC Bentley, and JD Keene. A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. *Cell*, 57(1):89–101, April 1989.

M Rabani, M Kertesz, and E Segal. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *PNAS*, 105(39):14885–90, September 2008.

O Rackham, TR Mercer, and A Filipovska. The human mitochondrial transcriptome and the RNA-binding proteins that regulate its expression. *Wiley Interdisciplinary Reviews: RNA*, 3(5):675–95, 2002.

N Rasche, O Dybkov, J Schmitzová, B Akyildiz, P Fabrizio, and R Lührmann. Cwc2 and its human homologue RBM22 promote an active conformation of the spliceosome catalytic centre. *The EMBO Journal*, 31(6):1591–604, March 2012.

D Ray, H Kazan, ET Chan, L Peña Castillo, S Chaudhry, S Talukder, BJ Blencowe, Q Morris, and TR Hughes. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnology*, 27(7):667–70, July 2009.

D Ray, H Kazan, KB Cook, MT Weirauch, HS Najafabadi, X Li, S Gueroussov, M Albu, H Zheng, A Yang, H Na, M Irimia, LH Matzat, RK Dale, SA Smith, CA Yarosh, SM Kelly, B Nabet, D Mecenas, W Li, RS Laishram, M Qiao, HD Lipshitz, F Piano, AH Corbett, RP Carstens, BJ Frey, RA Anderson, KW Lynch, LO Penalva, EP Lei, AG Fraser, BJ Blencowe, QD Morris, and TR Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–7, July 2013.

D Ray, KCH Ha, K Nie, H Zheng, TR Hughes, and QD Morris. RNAcompete methodology and application to determine sequence preferences of unconventional RNA-binding proteins. *Methods*, pages 3–15, April 2017.

J Rho, S Choi, CR Jung, and DS Im. Arginine methylation of Sam68 and SLM proteins negatively regulates their poly(U) RNA binding activity. *Archives of Biochemistry and Biophysics*, 466(1):49–57, October 2007.

DC Rio. Filter-binding assay for analysis of RNA-protein interactions. *Cold Spring Harbor Protocols*, 2012(10):1078–81, October 2012.

OS Rissland. The organization and regulation of mRNA-protein complexes. *Wiley Interdisciplinary Reviews: RNA*, 8(1), January 2017.

X Roca, AR Krainer, and IC Eperon. Pick one, but be quick: 5' splice sites and the problems of too many choices. *Genes and Development*, 27(2):129–44, January 2013.

AB Rosenberg, RP Patwardhan, J Shendure, and G Seelig. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell*, 163(3):698–711, October 2015.

S Rouskin, M Zubradt, S Washietl, M Kellis, and JS Weissman. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*. *Nature*, 505(7485): 701–5, January 2014.

JC Schwartz, X Wang, ER Podell, and TR Cech. RNA seeds higher-order assembly of FUS protein. *Cell Reports*, 5(4):918–25, November 2013.

JC Schwartz, TR Cech, and RR Parker. Biochemical properties and biological functions of FET proteins. *Annual Review of Biochemistry*, 84:355–79, 2015.

MG Seetin and DH Mathews. *RNA structure prediction: an overview of methods*, volume 905 of *Methods in Molecular Biology*, pages 99–122. Springer Protocols, 2012.

M Selbach, B Schwanhäusser, N Thierfelder, Z Fang, R Khanin, and N Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, September 2008.

P Shannon, A Markiel, O Ozier, NS Baliga, JT Wang, D Ramage, N Amin, B Schwikowski, and T Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–504, November 2003.

LV Sharova, AA Sharov, T Nedorezov, Y Piao, N Shaik, and MS Ko. Database for mRNA half-life of 19,977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Research*, 16(1):45–58, February 2009.

H Shen and MR Green. A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. *Molecular Cell*, 16(3):363–73, November 2004.

H Shen, JL Kan, and MR Green. Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. *Molecular Cell*, 13(3):367–76, February 2004.

S Shen, JW Park, J Huang, KA Dittmar, ZX Lu, Q Zhou, RP Carstens, and Y Xing. MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Research*, 40(8):e61, April 2012.

S Shen, JW Park, ZX Lu, L Lin, MD Henry, YN Wu, Q Zhou, and Y Xing. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *PNAS*, 111(51):E5593–601, December 2014.

IM Silverman, F Li, A Alexander, L Goff, C Trapnell, JL Rinn, and BD Gregory. RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. *Genome Biology*, 15(1), January 2014.

G Singh, G Pratt, GW Yeo, and MJ Moore. The clothes make the mRNA: Past and present trends in mRNP fashion. *Annual Reviews of Biochemistry*, 84:325–54, 2015.

NN Singh, RN Singh, and EJ Androphy. Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. *Nucleic Acids Research*, 35(2):371–89, 2007.

R Singh, J Valcárcel, and MR Green. Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science*, 268(5214):1173–6, May 1995.

H Siomi, MJ Matunis, WM Michael, and G Dreyfuss. The pre-mRNA binding K protein contains a novel evolutionarily conserved motif. *Nucleic Acids Research*, 21(5):1193–8, March 1993.

L Skalska, M Beltran-Nebot, J Ule, and RG Jenner. Regulatory feedback from nascent RNA to chromatin and transcription. *Nature Reviews Molecular Cellular Biology*, 18(5): 331–337, May 2017.

CW Smith, TT Chu, and B. Nadal-Ginard. Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Molecular and Cellular Biology*, 13(8):4939–52, August 1993.

SA Smith, D Ray, KB Cook, MJ Mallory, TR Hughes, and KW Lynch. Paralogs hnRNP L and hnRNP LL exhibit overlapping but distinct RNA binding constraints. *PLoS One*, 8 (11):e80701, November 2013.

N Sonenberg, MA Morgan, D Testa, RJ Colonno, and AJ Shatkin. Interaction of a limited set of proteins with different mRNAs and protection of 5'-caps against pyrophosphatase digestion in initiation complexes. *Nucleic Acids Research*, 7(1):15–29, September 1979a.

N Sonenberg, KM Rupprecht, SM Hecht, and AJ Shatkin. Eukaryotic mRNA cap binding protein: purification by affinity chromatography on sepharose-coupled m7GDP. *PNAS*, 76(9):4345–9, September 1979b.

DL Spector and AI Lamond. Nuclear speckles. *Cold Spring Harbor Perspectives in Biology*, 3(2), February 2011.

N Spies, CB Burge, and DP Bartel. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Research*, 23(12):2078–90, December 2013.

RC Spitale, RA Flynn, QC Zhang, P Crisalli, B Lee, JW Jung, HY Kuchelmeister, PJ Batista, EA Torre, Kool ET, and HY Chang. Structural imprints *in vivo* decode RNA regulatory mechanisms. *Nature*, 519(7544):486–90, March 2015.

B Stebbins-Boaz, Q Cao, CH de Moor, R Mendez, and JD Richter. Maskin is a CPEB-associated factor that transiently interacts with elF-4E. *Molecular Cell*, 4(6):1017–27, December 1999.

AO Subtelny, SW Eichhorn, GR Chen, H Sive, and DP Bartel. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature*, 508(7494):66–71, April 2014.

Y Sugimoto, J König, S Hussain, B Zupan, T Curk, M Frye, and J Ule. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biology*, 13(8), August 2012.

Y Sugimoto, A Vigilante, E Darbo, A Zirra, C Militti, A D'Ambrogio, NM Luscombe, and J Ule. hiCLIP reveals the *in vivo* atlas of mRNA secondary structures recognized by Staufen 1. *Nature*, 519(7544):491–4, March 2015.

B Sundararaman, L Zhan, SM Blue, R Stanton, K Elkins, S Olson, X Wei, EL Van Nostrand, GA Pratt, SC Huelga, BM Smalec, X Wang, EL Hong, JM Davidson, E Lécuyer, BR Graveley, and GW Yeo. Resources for the comprehensive discovery of functional RNA elements. *Molecular Cell*, 61(6):903–13, March 2016.

JM Taliaferro, NJ Lambert, PH Sudmant, D Dominguez, JJ Merkin, MS Alexis, C Bazile, and CB Burge. RNA sequence context effects measured *in vitro* predict *in vivo* protein binding and regulation. *Molecular Cell*, 64(2):294–306, October 2016.

J Talkish, G May, Y Lin, JL Woolford Jr, and CJ McManus. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *Nature*, 20(5):713–20, May 2014.

H Tani and N Akimitsu. Genome-wide technology for determining RNA stability in mammalian cells: historical perspective and recent advantages based on modified nucleotide labeling. *RNA Biology*, 9(10):1233–8, October 2012.

X Tao and G Gao. Tristetraprolin recruits eukaryotic initiation factor 4E2 to repress translation of AU-rich element-containing mRNAs. *Molecular and Cellular Biology*, 35(22): 3921–3, November 2015.

SA Tenenbaum, CC Carson, PJ Lager, and JD Keene. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *PNAS*, 97(26):14085–90, December 2000.

M Teplova, J Song, HY Gaw, A Teplov, and DJ Patel. Structural insights into RNA recognition by the alternate-splicing regulator CUG-Binding Protein 1. *Structure*, 18(10):1364–77, October 2010.

M Teplova, L Malinina, JC Darnell, J Song, M Lu, R Abagyan, K Musunuru, A Teplov, SK Burley, RB Darnell, and DJ Patel. Protein-RNA and protein-protein recognition by dual KH1/2 domains of the neuronal splicing factor Nova-1. *Structure*, 19(7):930–44, July 2011.

P Thandapani, TR O'Connor, TL Bailey, and S Richard. Defining the RGG/RG motif. *Molecular Cell*, 50(5):613–23, June 2013.

T Thisted, DL Lyakhov, and SA Liebhaber. Optimized RNA targets of two closely related triple KH domain proteins, heterogeneous nuclear ribonucleoprotein K and αCP-2KL, suggest distinct modes of RNA recognition. *Journal of Biological Chemistry*, 276(20): 17484–96, May 2001.

JM Tome, A Ozer, JM Pagano, D Gheba, GP Schroth, and JT Lis. Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. *Nature Methods*, 11(6):683–8, June 2014.

C Trapnell, A Roberts, L Goff, G Pertea, D Kim, DR Kelley, H Pimentel, SL Salzberg, JL Rinn, and L Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–78, March 2012.

C Trapnell, DG Hendrickson, M Sauvageau, L Goff, JL Rinn, and L Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, 31(1):46–53, January 2013.

T Treiber, N Treiber, U Plessmann, S Harlander, JL Daiß, N Eichner, G Lehmann, K Schall, H Urlaub, and G Meister. A compendium of RNA-binding proteins that regulate microRNA biogenesis. *Molecular Cell*, 66(2):270–284, April 2017.

YS Tsai, SM Gomez, and Z Wang. Prevalent RNA recognition motif duplication in the human genome. *RNA*, 20(5):702–12, May 2014.

C Tuerk and L Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968):505–10, August 1990.

DH Turner, N Sugimoto, JA Jaeger, CE Longfellow, SM Freier, and R Kierzek. Improved parameters for prediction of RNA structure. *Cold Spring Harbor Symposium on Quantitative Biology*, 52:123–33, July 1987.

J Ule, KB Jensen, M Ruggiu, A Mele, A Ule, and RB Darnell. CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 302(5648):1212–5, November 2003.

J Ule, G Stefani, A Mele, M Ruggiu, X Wang, B Taneri, T Gaasterland, BJ Blencowe, and RB Darnell. An RNA map predicting Nova-dependent splicing regulation. *Nature*, 444 (7119):580–6, November 2006.

JG Underwood, AV Uzilov, S Katzman, CS Onodera, JE Mainzer, DH Mathews, TM Lowe, SR Salama, and D Haussler. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature Methods*, 7(12):995–1001, December 2010.

R Valverde, L Edwards, and L Regan. Structure and function of KH domains. *The FEBS Journal*, 275(11):2712–26, June 2008.

EL Van Nostrand, GA Pratt, AA Shishkin, C Gelboin-Burkhart, MY Fang, B Sundararaman, SM Blue, TB Nguyen, C Surka, K Elkins, R Stanton, F Rigo, M Guttman, and GW Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6):508–14, June 2016.

EL Van Nostrand, P Freese, GA Pratt, X Wang, X Wei, R Xiao, SM Blue, D Dominguez, NAL Cody, S Olson, B Sundararaman, L Zhan, C Bazile, LPB Bouvrette, J Chen, MO Duff, KE Garcia, C Gelboin-Burkhart, M Hochman, NJ Lambert, H Li, TB Nguyen, T Palden, I Rabano, S Sathe, R Stanton, J Bergalet, B Zhou, A Su, R Wang, BA Yee,

AL Louie, S Aigner, X Fu, E Lecuyer, CB Burge, BR Graveley, and GW Yeo. A large-scale binding and functional map of human RNA binding proteins, Aug 2017. http://doi.org/10.1101/179648 bioRxiv; posted 8/23/2017.

IA Vlasova, NM Tahoe, D Fan, O Larsson, B Rattenbacher, JR Sternjohn, J Vasdewani, G Karypis, CS Reilly, PB Bitterman, and PR Bohjanen. Conserved GU-rich elements mediate mRNA decay by binding to CUG-binding protein 1. *Molecular Cell*, 29(2):263–70, February 2008.

MC Wahl, CL Will, and R Lührmann. The Spliceosome: Design principles of a dynamic RNP machine. *Cell*, 136(4):701–18, February 2009.

DC Wai, M Shihab, JK Low, and JP MacKay. The zinc fingers of YY1 bind single-stranded RNA with low sequence specificity. *Nucleic Acids Research*, 44(19):9153–9165, November 2016.

R Walczak, E Westhof, P Carbon, and A Krol. A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. *RNA*, 2(4):376–79, April 1996.

Y Wan, K Qu, QC Zhang, RA Flynn, O Manor, Z Ouyang, J Zhang, RC Spitale, MP Snyder, E Segal, and HY Chang. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, 505(7485):706–9, January 2014.

ET Wang, R Sandberg, S Luo, I Khrebtukova, L Zhang, C Mayr, SF Kingsmore, GP Schroth, and CB Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–6, November 2008.

T Wang, K Birsoy, NW Hughes, KM Krupczak, Y Post, JJ Wei, ES Lander, and DM Sabatini. Identification and characterization of essential genes in the human genome. *Science*, 350(6264):1096–101, November 2015.

X Wang and TM Tanaka Hall. Structural basis for recognition of AU-rich element RNA by the HuD protein. *Nature Structural & Molecular Biology*, 8(2):141–5, February 2001.

Y Wang, M Ma, X Xiao, and Z Wang. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nature Structural & Molecular Biology*, 19(10): 1044–52, October 2012.

Y Wang, X Xiao, J Zhang, R Choudhury, A Robertson, K Li, M Ma, CB Burge, and Z Wang. A complex network of factors with overlapping affinities represses splicing through intronic elements. *Nature Structural & Molecular Biology*, 20(1):36–45, January 2013.

ZF Wang, ML Whitfield, TC Ingledue, Z Dominski, and WF Marzluff. The protein that binds the 3' end of histone mRNA: a novel RNA-binding protein required for histone pre-mRNA processing. *Genes & Development*, 10(23):3028–40, December 1996.

MB Warf and JA Berglund. Role of RNA structure in regulating pre-mRNA splicing. *Trends in Biochemical Sciences*, 35(3):169–78, March 2010.

C Wei, R Xiao, L Chen, H Cui, Y Zhou, Y Xue, J Hu, B Zhou, T Tsutsui, J Qiu, H Li, L Tang, and XD Fu. RBFox2 binds nascent RNA to globally regulate polycomb complex 2 targeting in mammalian genomes. *Molecular Cell*, 62(6):875–889, June 2016.

GH Wei, G Badis, MF Berger, T Kivioja, K Palin, M Enge, M Bonke, A Jolma, M Varjosalo, AR Gehrke, J Yan, S Talukder, M Turunen, M Taipale, HG Stunnenberg, E Ukkonen, TR Hughes, ML Bulyk, and J Taipale. Genome-wide analysis of ETS-family DNA-binding *in vitro* and *in vivo*. *The EMBO Journal*, 29(13):2147–60, July 2010.

EC Wheeler, EL Van Nostrand, and GW Yeo. Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. *Wiley Interdisciplinary Reviews: RNA*, August 2017.

CP Wigington, J Jung, EA Rye, SL Belauret, AM Philpot, Y Feng, PJ Santangelo, and AH Corbett. Post-transcriptional regulation of programmed cell death 4 (PDCD4) mRNA by the RNA-binding proteins human antigen R (HuR) and T-cell intracellular antigen 1 (TIA1). *Journal of Biological Chemistry*, 290(6):3468–87, February 2015.

EM Wissink, EA Fogarty, and A Grimson. High-throughput discovery of post-transcriptional *cis*-regulatory elements. *BMC Genomics*, 17(177), March 2016.

JT Witten and J Ule. Understanding splicing regulation through RNA splicing maps. *Trends in Genetics*, 27(3):89–97, March 2011.

X Wu and G Brewer. The regulation of mRNA stability in mammalian cells: 2.0. *Gene*, 500 (1):10–21, May 2012.

X Xie, J Lu, EJ Kulbokas, TR Golub, V Mootha, K Lindblad-Toh, ES Lander, and M Kellis. Systematic discovery of regulatory motifs in human promoters and 3'UTRs by comparison of several mammals. *Nature*, 434(7031):338–45, March 2005.

HY Xiong, B Alipanahi, LJ Lee, H Bretschneider, D Merico, RK Yuen, Y Hua, S Gueroussov, HS Najafabadi, TR Hughes, Q Morris, Y Barash, AR Krainer, N Jojic, SW Scherer, BJ Blencowe, and BJ Frey. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218):1254806, January 2015.

AV Yakhnin, H Yakhnin, and P Babitzke. *Gel mobility shift assays to detect protein-RNA interactions*, volume 905 of *Methods in Molecular Biology*, pages 201–11. Springer Protocols, 2012.

S Yamasaki, G Stoecklin, N Kedersha, M Simarro, and P Anderson. T-cell intracellular antigen-1 (TIA-1)-induced translational silencing promotes the decay of selected mRNAs. *Journal of Biological Chemistry*, 282(41):30070–7, October 2007.

X Yang, Y Yang, BF Sun, YS Chen, JW Xu, WY Lai, A Li, X Wang, DP Bhattarai, W Xiao, HY Sun, Q Zhu, HL Ma, S Adhikari, M Sun, YJ Hao, B Zhang, CM Huang, N Huang, GB Jiang, YL Zhao, HL Wang, YP Sun, and YG Yang. 5-methylcytosine promotes mRNA export - NSUN2 as the methyltransferase and ALYREF as an m5C reader. *Cell Research*, 27(5):606–625, May 2017.

GW Yeo, NG Coufal, TY Liang, GE Peng, XD Fu, and FH Gage. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature Structural & Molecular Biology*, 16(2):130–7, February 2009.

PD Zamore, JR Williamson, and R Lehmann. The pumilio protein binds RNA through a conserved domain that defines a new class of RNA-binding proteins. *RNA*, 3(12):1421–33, December 1997.

G Zhai, M Iskandar, K Barilla, and PJ Romaniuk. Characterization of RNA aptamer binding by the Wilms' tumor suppressor protein WT1. *Biochemistry*, 40(7):2032–40, February 2001.

X Zhang, C Yan, J Hang, LI Finci, J Lei, and Y Shi. An atomic structure of the human spliceosome. *Cell*, 169(5):918–929, May 2017.

Y Zhang, Z Lu, L Ku, Y Chen, H Wang, and Y Feng. Tyrosine phosphorylation of QKI mediates developmental signals to regulate mRNA metabolism. *The EMBO Journal*, 22 (8):1801–10, April 2003.

J Zhao, TK Ohsumi, JT Kung, Y Ogawa, DJ Grau, K Sarma, JJ Song, RE Kingston, M Borowsky, and JT Lee. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular Cell*, 40(6):939–53, December 2010.

A Zhou, AC Ou, A Cho, EJ Benz Jr, and SC Huang. Novel splicing factor RBM25 modulates Bcl-x pre-mRNA 5'splice site selection. *Molecular and Cellular Biology*, 28(19):5924–36, October 2008.

FY Zong, X Fu, WJ Wei, YG Luo, M Heiner, LJ Cao, Z Fang, R Fang, D Lu, H Ji, and J Hui. The RNA-binding protein QKI suppresses cancer-associated aberrant splicing. *PLoS Genetics*, 10(4):e1004289, April 2014.

BE Zucconi, JD Ballin, BY Brewer, CR Ross, J Huang, EA Toth, and GM Wilson. Alternatively expressed domains of AU-rich element RNA-binding protein 1 (AUF1) regulate RNA-binding affinity, RNA-induced protein oligomerization, and the local conformation of bound RNA ligands. *Journal of Biological Chemistry*, 285(50):39127–39, December 2010.